B.M.S. COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU, Belagavi)

DEPARTMENT OF MACHINE LEARNING

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

Course :Mini Project

Course Code: 24AM5PWMPW

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

Final Review

Date: 15-02-2025

Presented By,

Student Name & USN :

Vishesh Bishnoi(1BM22AI155)

Siddharth Sahay(1BM22AI128)

Varsh Gandhi(1BM23AI413)

Mathias Prajwal Dsouza(1BM23AI408)

Semester & Section: **5B & 5C**

Batch Number: 3

Faculty In-Charge:

Dr. Sandeep Varma N

Associate Professor

Department of Machine Learning

B.M.S. College of Engineering

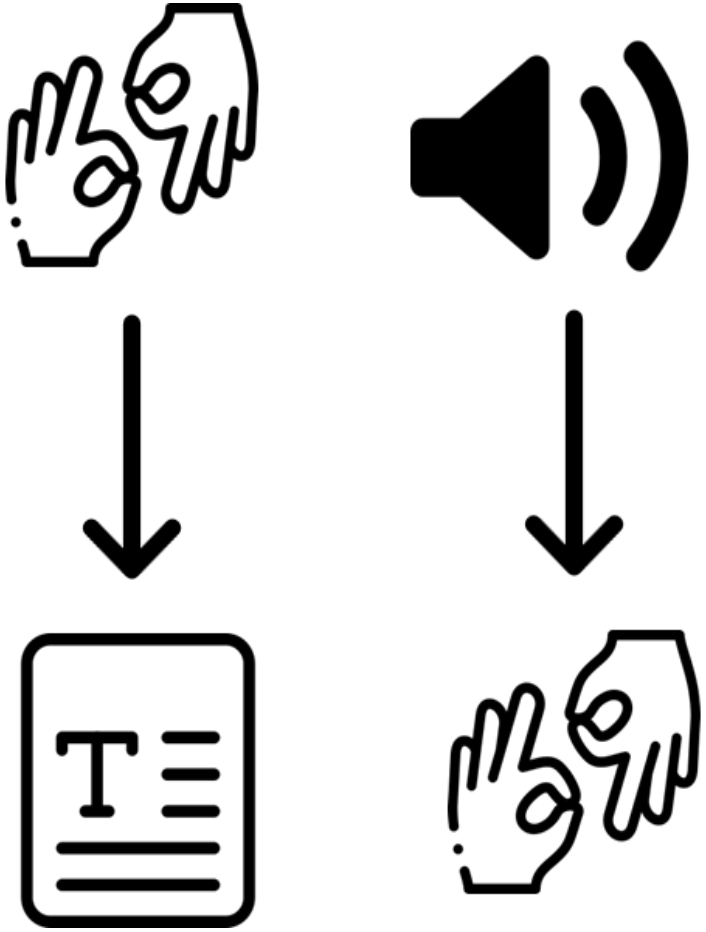
Agenda

- Introduction
- Literature Review
- Open Issues
- Problem Statement
- Proposed Architecture
- Functional & Non-Functional Requirements
- Low-Level Design
- Methodology
- Implementation
- Progress in Project so far
- Experiment results and analysis
- Testing and Validation
- Conclusion

Introduction

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 a multimodal architecture 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.



Literature review

AUTHOR / TITLE / YEAR	APPLIED METHODOLOGY / ALGORITHM USED	FINDINGS	RESULTS	LIMITATIONS
M. Kowsigan, Rahul Dhawan, Ankan Kundu/ "An Efficient Speech to Sign Language Conversion and Text Recognition through Live Gesture" /**2024**	Combined speech-to-sign language conversion and live gesture-to-text recognition using machine learning and custom datasets.	Demonstrates inclusivity by supporting ASL and ISL. Achieved significant improvement in gesture recognition using NLP and CNNs.	Achieved over 85% accuracy in bidirectional communication, enhancing interaction for impaired users.	Requires further work on regional sign language nuances, real-time improvements, and cultural adaptability.
Jeevanandham P, George Britt A, Hariharan A "Real-Time Hand Sign Language Translation: Text and Speech Conversion“/ **2024**	Used Mediapipe for landmark detection and Random Forest for sign classification, combined with CNN for gesture recognition.	Successfully converted sign gestures to text and speech in real time using webcam input.	Achieved fast and reliable gesture recognition with an average processing time of 0.3 seconds per gesture.	Limited by complexity of dynamic gestures and handling diverse regional sign languages.
Dr. M. Kavitha, Aditi Chatterjee, Shivam Shrivastava, and Gourav Sarkar, “Formation of Text from Indian Sign Language using Convolutional Neural Networks,”/ **2022**	ISL dataset was used with 19167 images of different numbers and letters. The paper used CNN integrated with a GUI made using tkinter library in python	Using CNN models in ISL translation is better than other models like SVM and ANN	Achieved 98.82% accuracy and training loss of 0.0578 using the proposed CNN model	There is limited vocabulary in the dataset chosen for the training. The models has been made for single handed gestures

Literature review

AUTHOR / TITLE / YEAR	APPLIED METHODOLOGY / ALGORITHM USED	FINDINGS	RESULTS	LIMITATIONS
Dr. K. Anitha Sheela, Chevella Anil Kumar, Jella Sandhya, Nadia Begum Shaik, and Gaddam Ravindra, Indian Sign Language Translator, **2022**	Mediapipe hands was used for hand gesture detection. Inception V3 and LSTM was used for extracting temporal features from gesture sequences.	The system focuses on static and dynamic gesture categorization for number, letters, greeting and medical terms.	ISL translation achieved 96.26% accuracy for ISL recognition system.	Dataset only includes 76 gestures which is insufficient for ISL converge. Model has difficulty in gestures with similar gestures.
Jashwanth Peguda,V Sai Sriharsha Santosh,Y Vijayalata,Ashlin Deepa R N and Vaddi Mounish /"Speech to Sign Language Translation for Indian Languages"/**2022**	The paper uses Wavelet-based MFCC with GMM for speech recognition, followed by an LSTM encoder-decoder model for text translation, and maps the translated text to Indian Sign Language gestures.	It shows that combining Wavelet-based MFCC with GMM & LSTM improves accuracy in speech-to-text and gesture recognition for ISL systems. This hybrid approach outperforms traditional methods by capturing both spectral and temporal features effectively.	The hybrid model achieved an accuracy of **93.5%** in speech-to-text conversion and **91.2%** in gesture recognition	The research is limited by the need for large datasets and high computational complexity, hindering real-time application in resource-constrained environments.
Pankaj Sonawane,Karan Shah,Parth Patel,Shikhar Shah and Jay Shah/"Speech To Indian Sign Language (ISL)Translation System"/**2021**	Kinect was used for motion capture of ISL gestures,mapped onto a 3D avatar in Unity. Google's Speech-to-Text API processed spoken input, which was parsed and synchronized with the corresponding ISL gestures for real-time communication.	Proposed system successfully translates speech to ISL in 91% of cases	The system achieved 91% accuracy with ISL.Misinterpretations due to speech-to-text limitations were addressed, enhancing the overall functionality.	Challenges with speech-to-text accuracy, leading to occasional misinterpretations of signs.the lack of a comprehensive ISL database & the variability in regional dialects complicate accurate translation

Open Issues

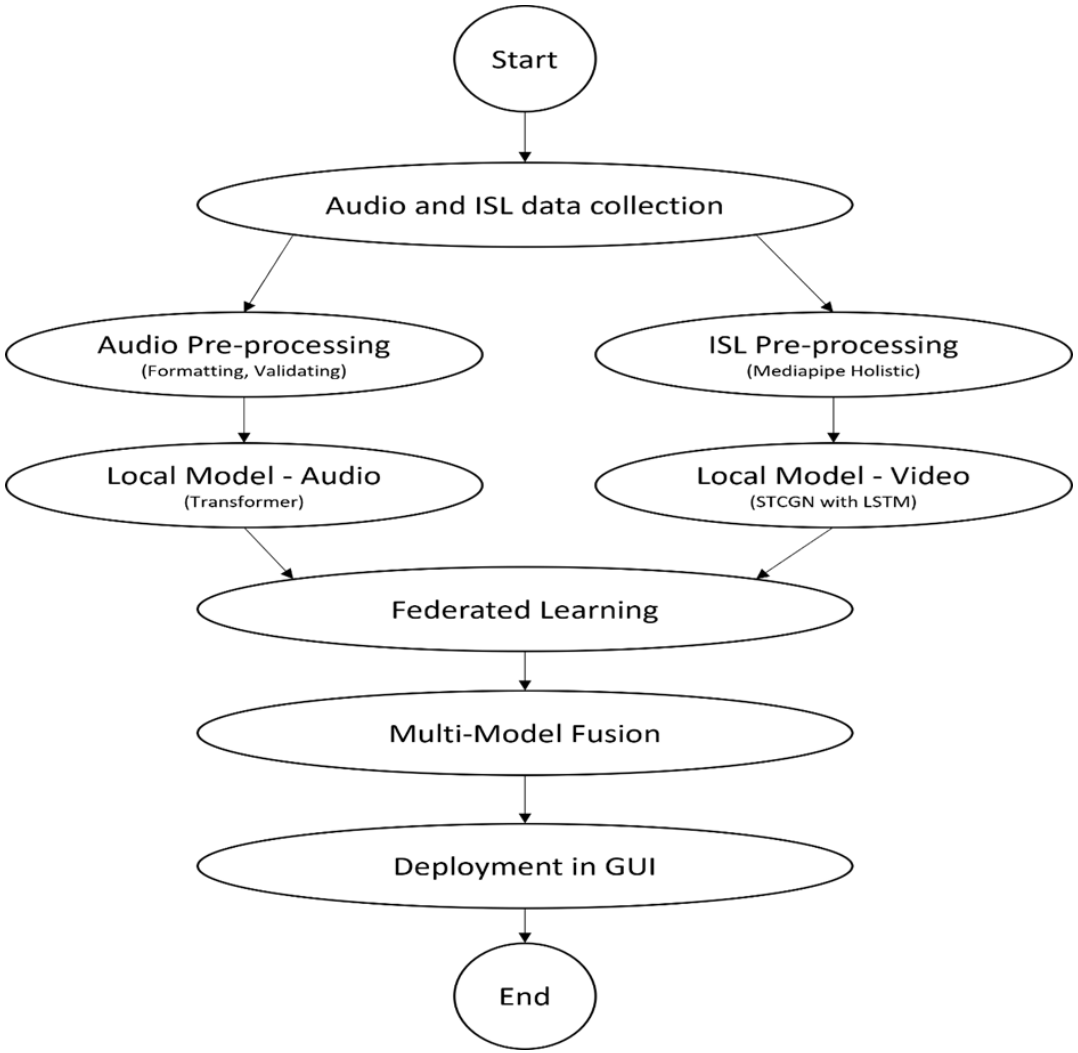
Here are some open issues and challenges with implementing ISL-to-text and audio-to-ISL conversion using a multimodal architecture and federated learning:

1. **Limited ISL Dataset Availability :**Current datasets for Indian Sign Language (ISL) are limited in scale, diversity, and regional variations, affecting the model's ability to generalize effectively.
2. **Gesture Similarity Challenges:**ISL gestures often involve subtle differences, making it difficult for models to distinguish between similar signs.
3. **Federated Learning Constraints**: Federated learning requires significant computational resources and reliable network connectivity for training across decentralized nodes, which can be a challenge in resource-constrained environments.
4. **Cultural and Regional Adaptation:**Indian Sign Language (ISL) varies regionally, influenced by India’s diverse languages and cultures. These adaptations enrich ISL but make standardization challenging.
5. **Dynamic Gesture Recognition Complexity:**Recognizing dynamic gestures involving sequential movements is more complex compared to static signs.

Problem Statement

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.

Proposed Architecture



Proposed Architecture

The system architecture is designed to process multi-modal inputs (audio and video) for generating sign language representations using federated learning. Here's a step-by-step breakdown of the architecture:

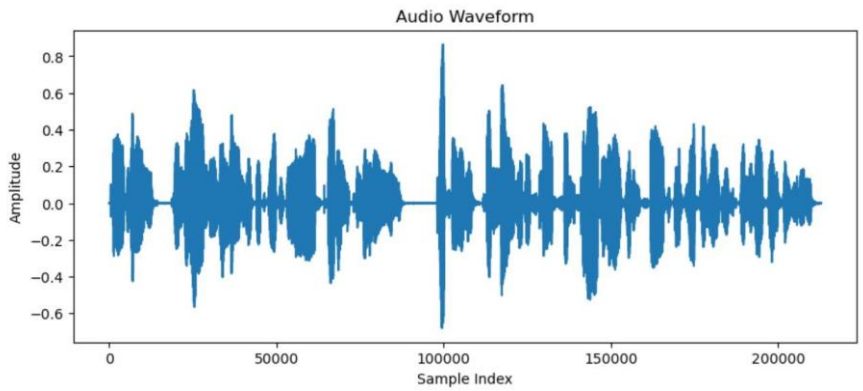
1. **Input Collection :**

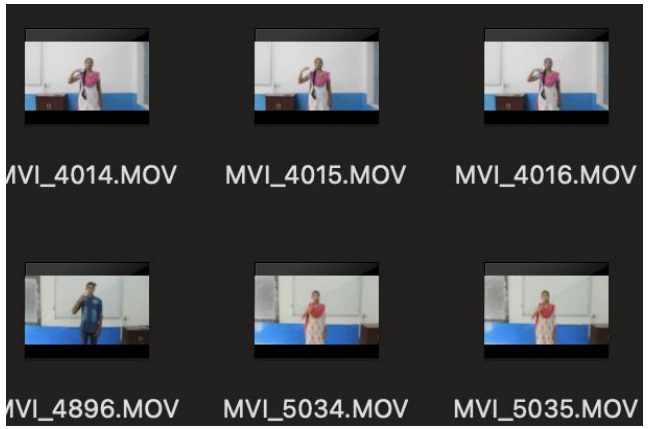
   a. **Audio Inputs (Audio Input 1 & 2):**

   - Audio streams are collected from two separate sources, such as microphones or audio-recording devices, ensuring diverse inputs for robust processing.

   a. **Video Inputs (Video Input 1 & 2):**

   - Video datasets featuring ISL gestures are sourced from various online platforms, focusing on body gestures and hand movements relevant to sign language interpretation.





Proposed Architecture

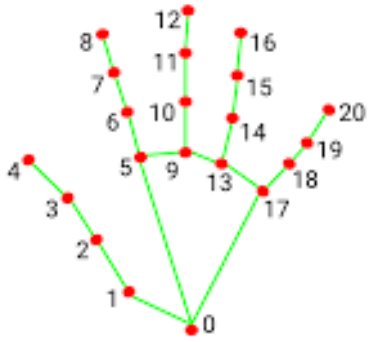
2. Audio and Video Preprocessing

a. Speech-to-Text Processing (Shared for Audio Inputs):

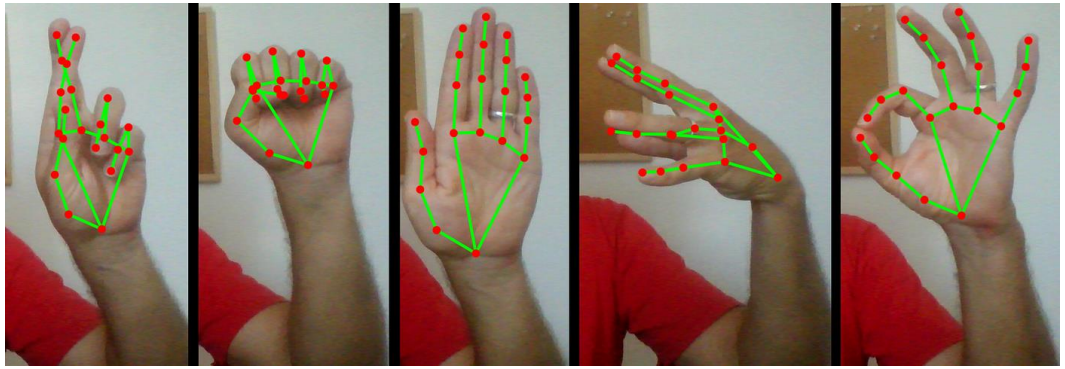
- STFT extraction and power normalization.
- Enhances accuracy by pre-processing the audio to remove noise and normalize the speech.
- Outputs text that represents the spoken content, preparing it for local model training.

a. Pose & Gesture Detection (Shared for Video Inputs):

- Body postures and key landmarks are detected using MediaPipe Holistic, enabling the interpretation of body movements in sign language.
- Hand gestures are analyzed with MediaPipe Holistic, providing essential gesture data for accurate sign language interpretation.



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP



Proposed Architecture

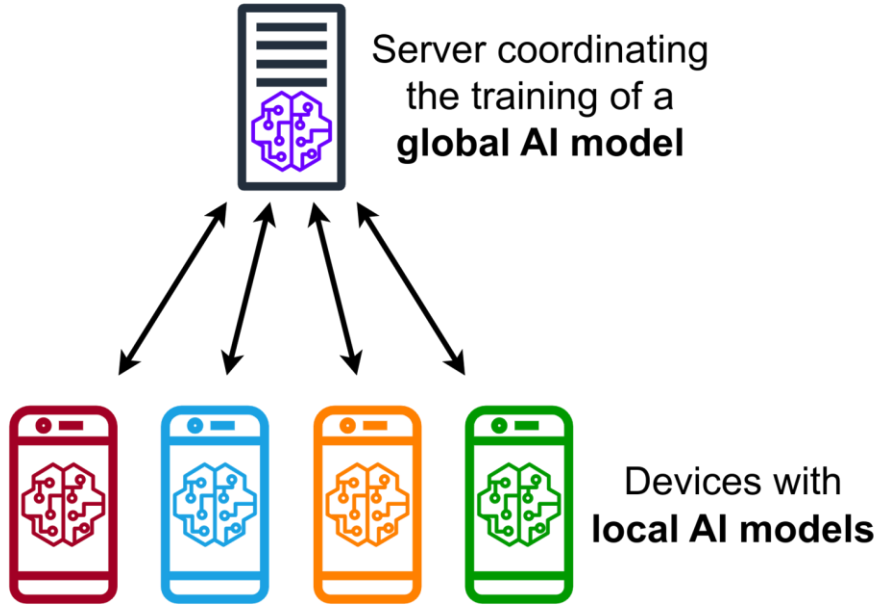
3. Local Model Training

a. Decentralized Training Nodes:

- Four local training models operate independently to
- Audio Models (2 Nodes): These models perform speech-to-text conversion,
- Video Models (2 Nodes): Utilize skeleton data consisting of 75 key landmarks to learn and classify hand gestures, facilitating accurate gesture recognition.

a. Benefits of Local Training :

- Local training allows models to be trained on different datasets across multiple nodes, enabling the processing of larger data volumes. This approach improves training efficiency, reduces processing time, and ensures better model performance through diverse data.

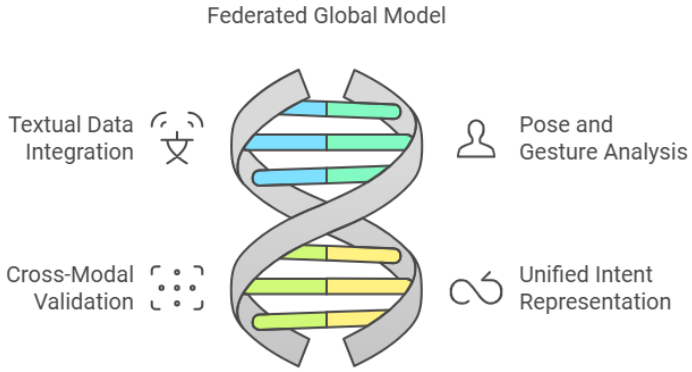


Proposed Architecture

4. Federated Learning Server

a. Global Model Aggregation:

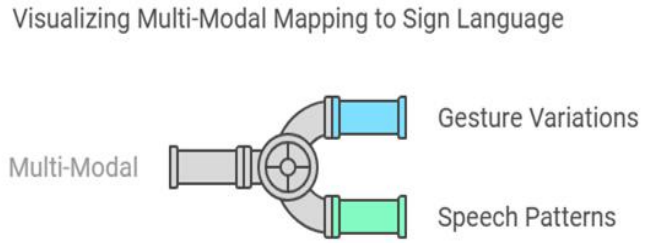
- Local model updates (parameters or gradients) from all nodes are sent to a centralized federated learning server.
- The server aggregates these updates using algorithms like Federated Averaging (FedAvg) to create a global model which is trained on large and diverse datasets.



5. Multi-Modal Fusion:

Integration of Audio and Video Insights:

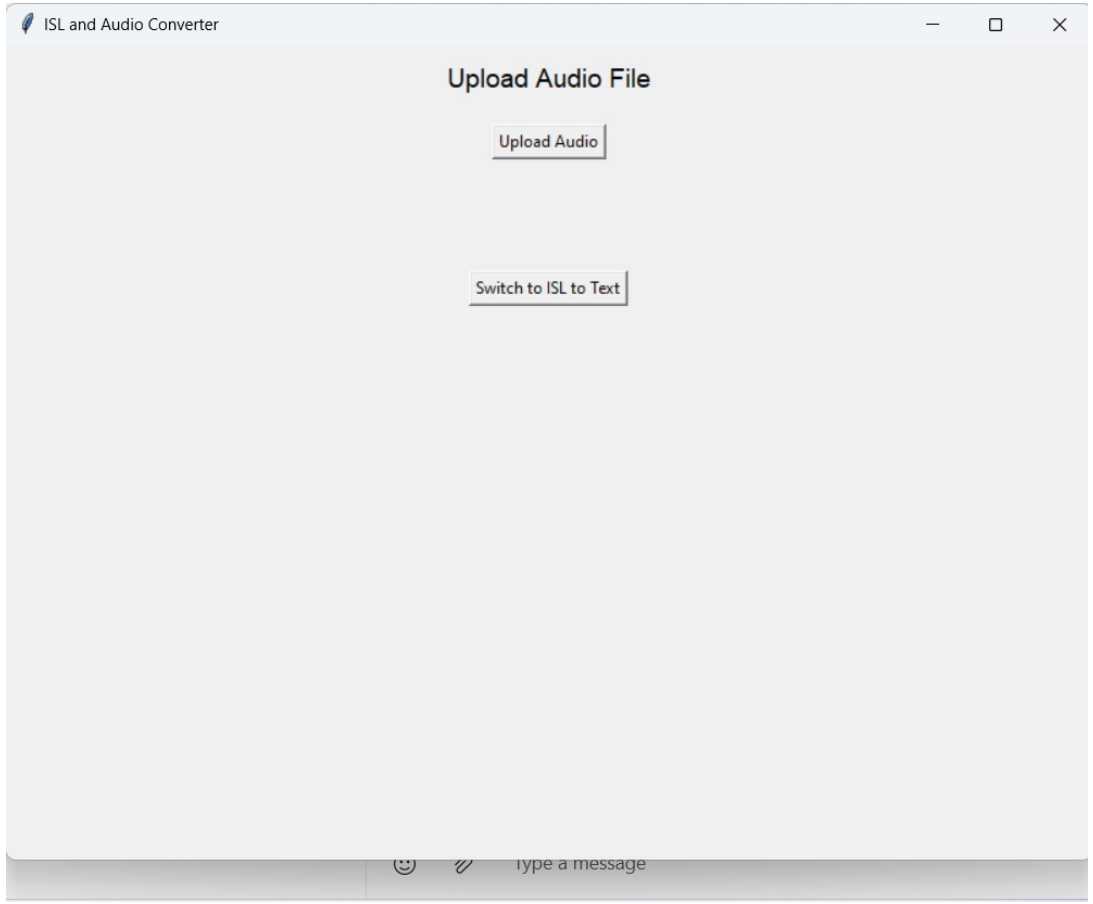
- The federated global model integrates the textual data from speech-to-text with pose and gesture key points from video inputs.
- Cross-validates information across modalities to resolve ambiguities and improve overall understanding.
- Produces a unified representation of the user's intent.

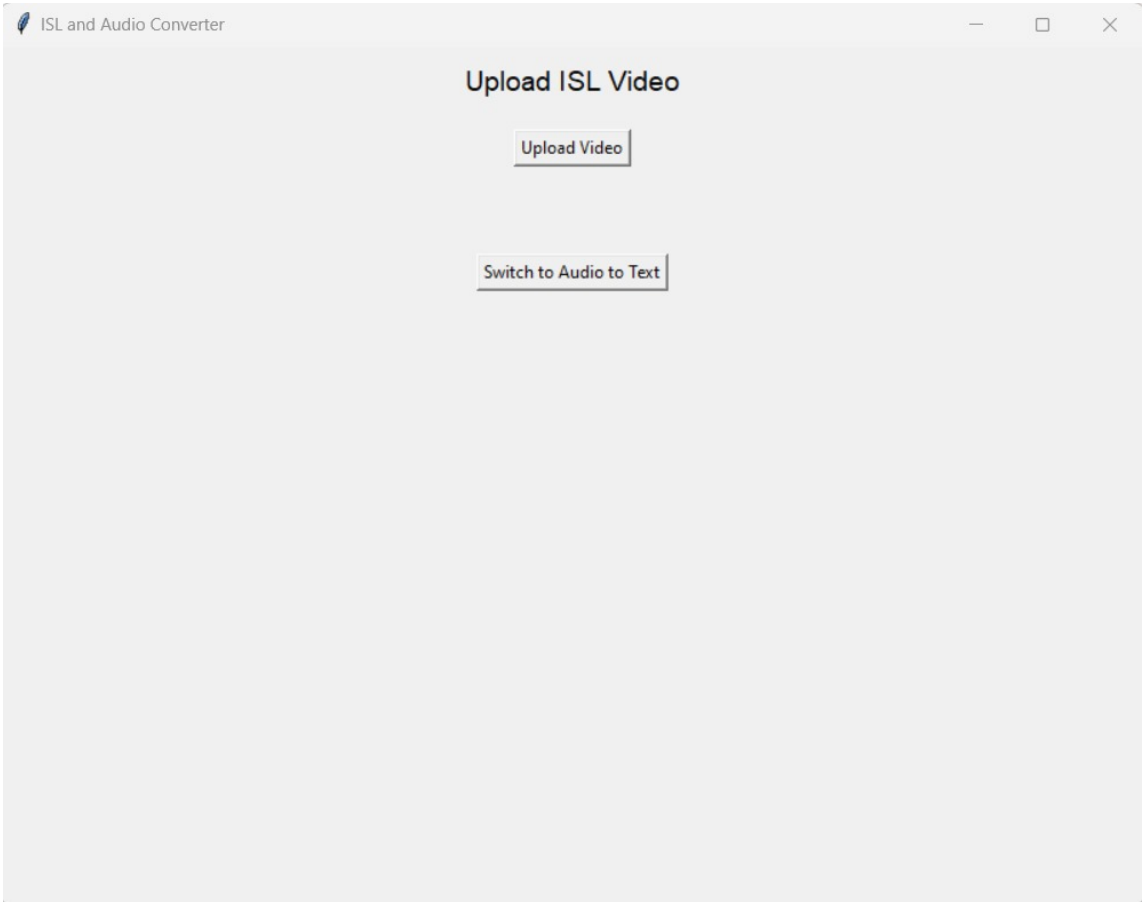


Proposed Architecture

6. Graphical User Interface

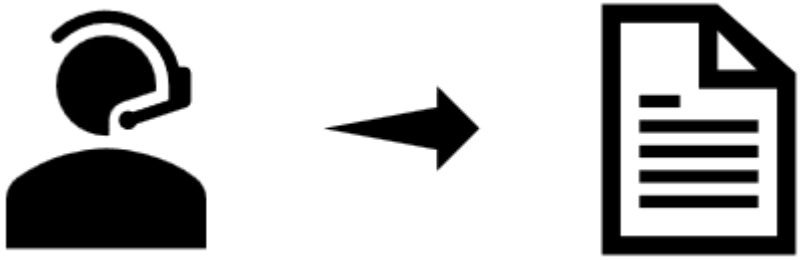
The GUI interface simplifies the translation process for the user, concealing the underlying complexity of the system.





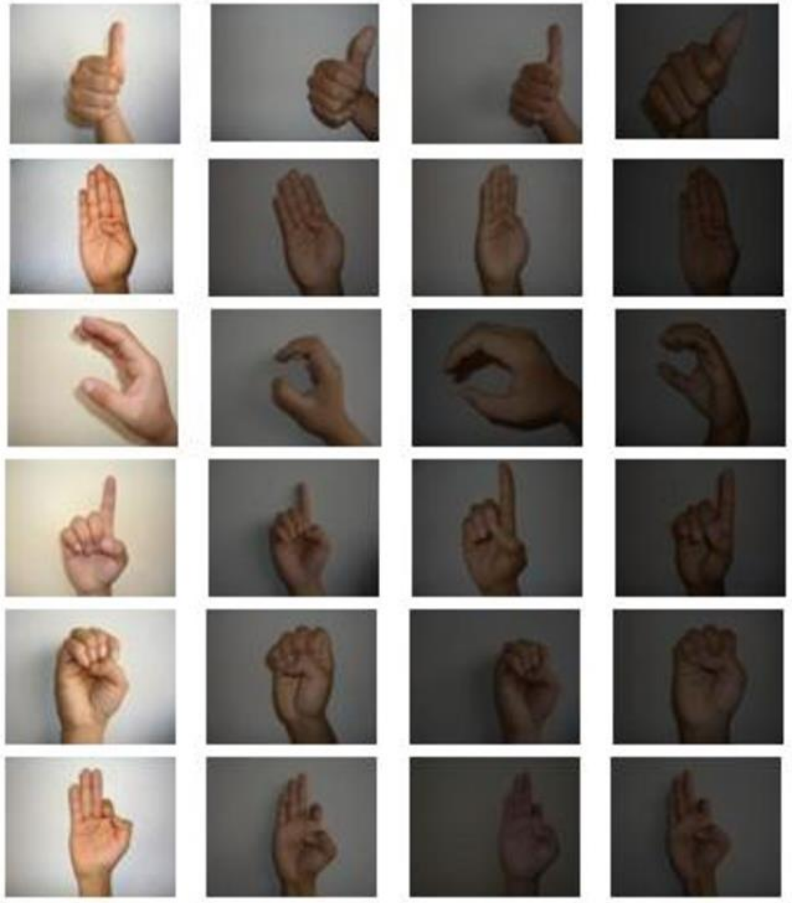
Functional requirements

1. **Diverse Data Distribution**: Enable training models on decentralized ISL & Audio datasets .
2. **Sign Recognition**: Accurately detect and interpret Indian Sign Language gestures.
3. **Translation Output**: Convert recognized signs into corresponding text and speech input into isl gesture.
4. **Multimodal Support**: ISL-to-text and speech-to-ISL gesture.

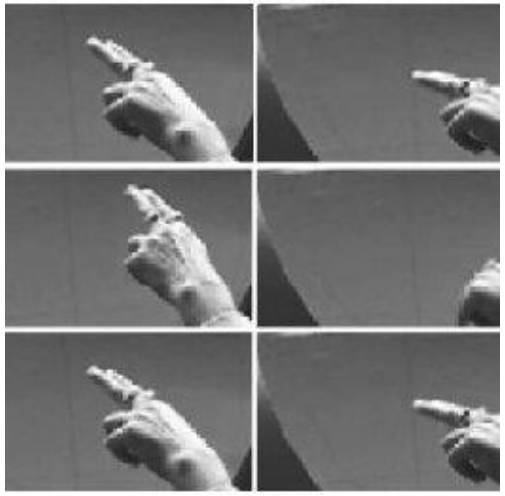


Non- Functional Requirements

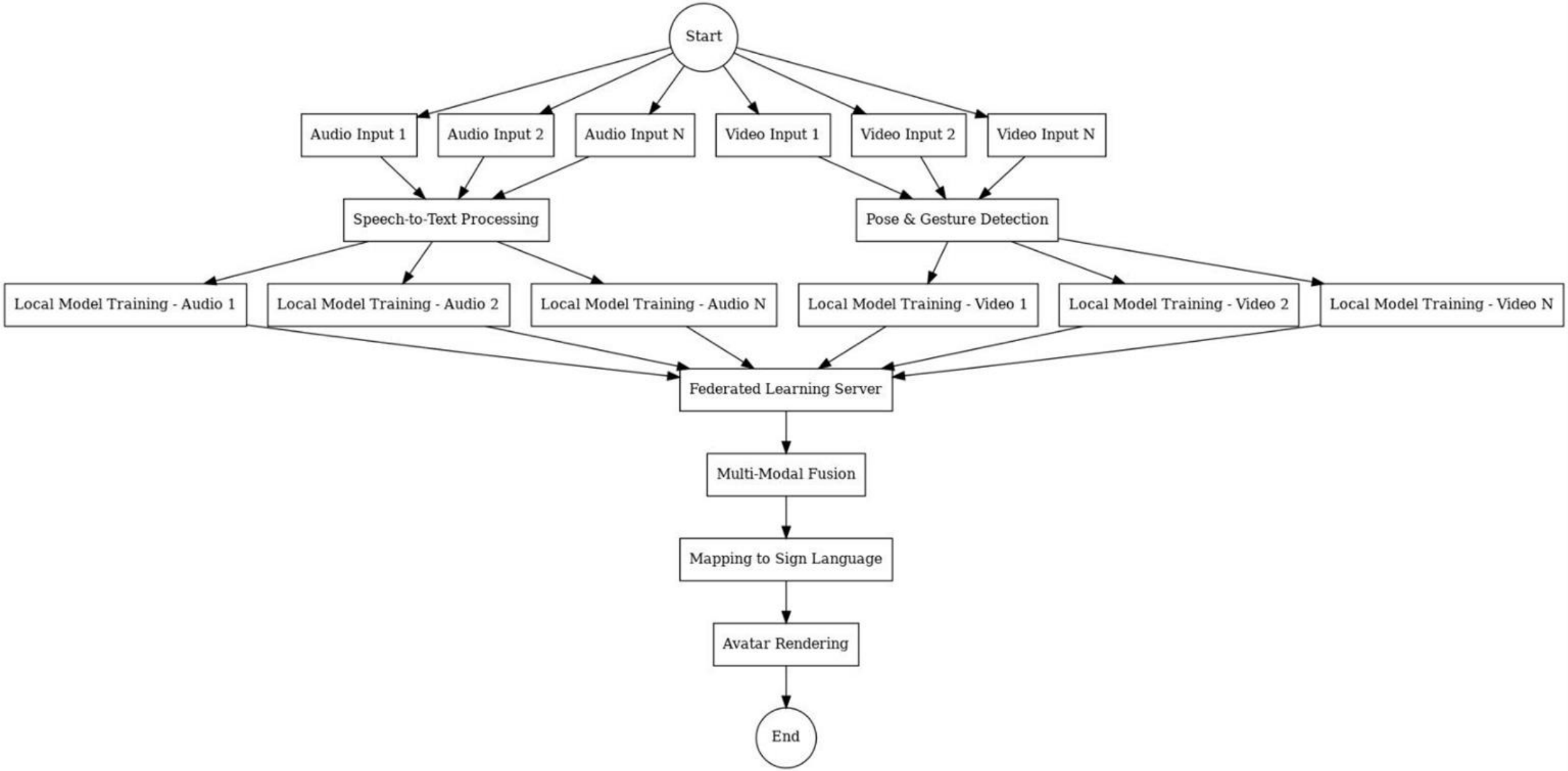
1. **Model Accuracy**: Achieve moderate translation accuracy (e.g., >65% for word level gestures).
2. **Robustness**: Handle noisy environments, partial gestures, and varied lighting conditions.
3. **Ease to use:** User interface is intuitive and requires no more than a few steps to complete a translation.



Graphical
User Interface

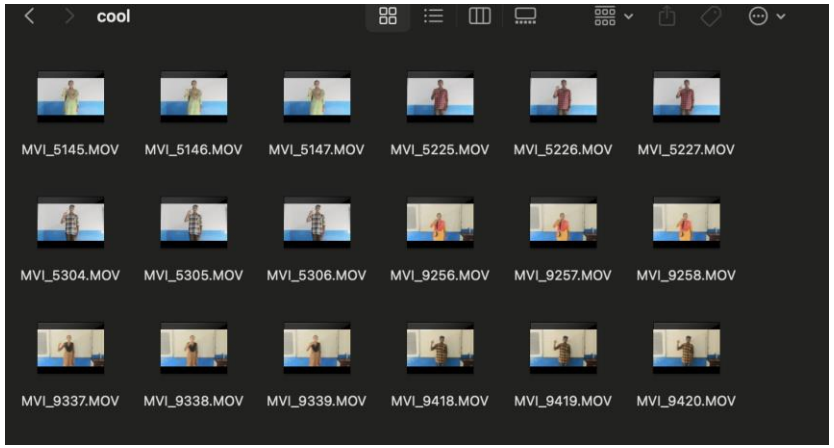


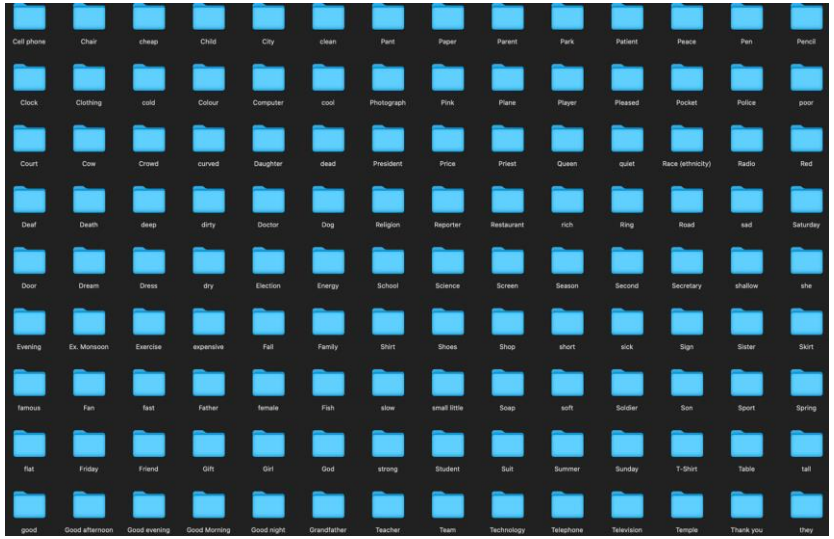
Low-Level Design



Low-Level Design

1. **Data Input:-**
   - **Video Input:-**
     - Moderate-sized ISL video dataset
     - Approximately 5,000 videos
     - Covers 262 classes (adjectives & common words)
     - Video length: 2 to 4 seconds
     - Recorded at 25 frames per second (fps) for clear gesture representation
   - **Audio Input:-**
     - Audio dataset sourced from "LJ Speech" on Kaggle
     - 13,100 short audio clips
     - Single speaker reading passages from seven non-fiction books
     - Clip length: 1 to 10 seconds, total duration ~24 hours
     - Texts published between 1884–1964, recordings from 2016–2017 (LibriVox project, public domain)

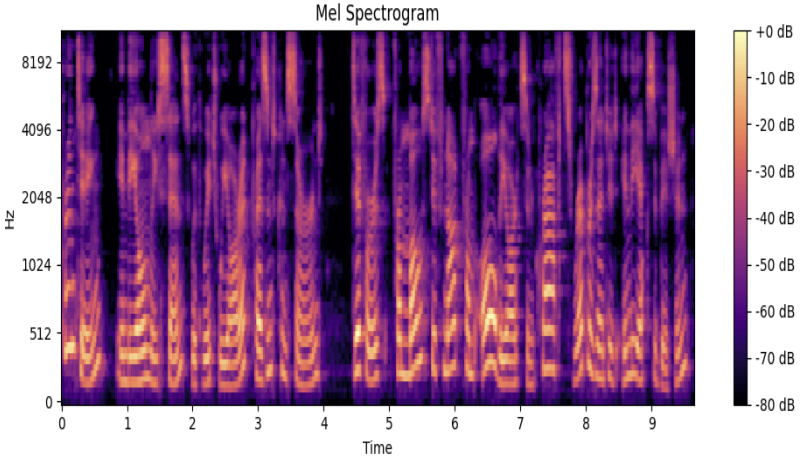


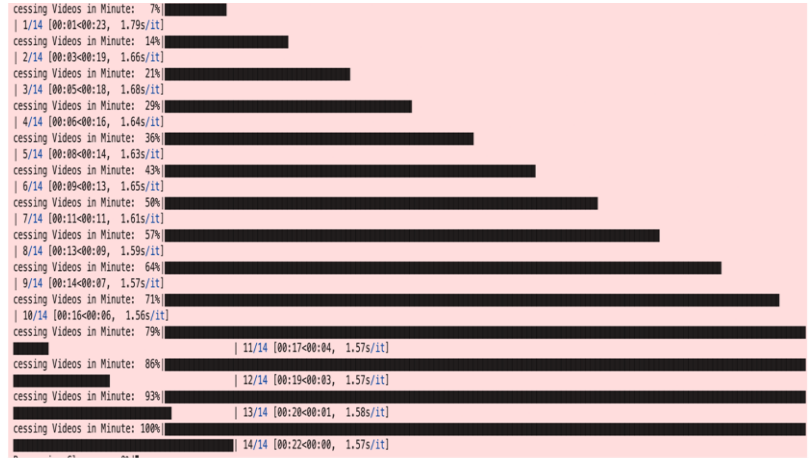


Low-Level Design

2. Data Preprocessing:-

- **Audio Preprocessing for:-**
  - Convert raw audio into normalized spectrograms
  - Apply Short-Time Fourier Transform (STFT) for frequency features
  - Power normalization for consistent feature scaling
  - Pad or truncate audio to 10 seconds (2754 frames)
  - Create TensorFlow dataset with batch size 4, pairing audio with text targets
- **Pose Detection:-**
  - Use MediaPipe Holistic for pose detection
  - Process video to maintain 50 frames (2 seconds at 25 FPS)
  - Extract 33 body key points and 21 hand key points
  - Process frames in RGB format for pose and hand landmark detection
  - Handle missing data by substituting zeros
  - Store extracted landmark data in JSON format for gesture classification

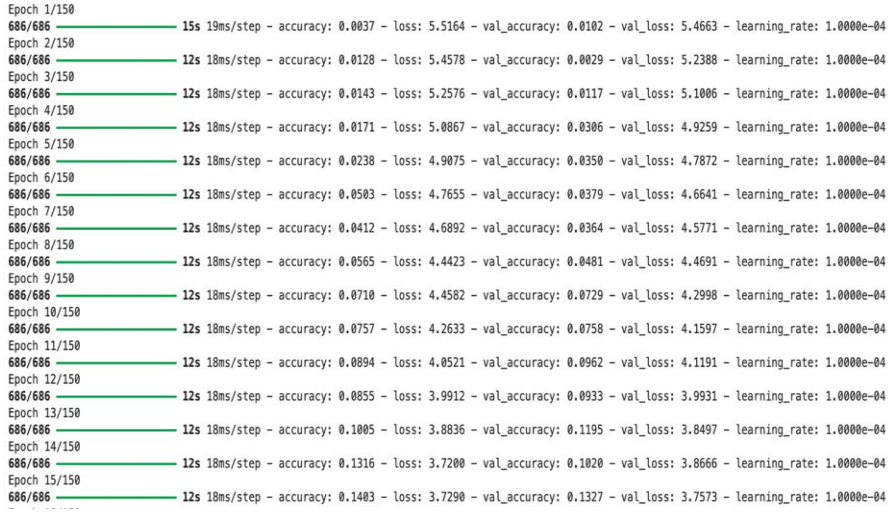




Low-Level Design

3. Local model:-

- **STCGN with LSTM Model for Low-Level Design:-**
  - Train STCGN model with LSTM for gesture classification
  - Apply temporal convolutional layers for spatial and temporal feature extraction
  - Use LSTM layers for sequence modeling and a sliding window technique for consistent input length
  - Normalize data and utilize class weights to handle imbalanced data
  - Incorporate dropout layers for regularization
  - Save the trained model and label encoder for future inference



- **Speech to text:-**
  - Speech-to-text conversion using transformer-based model
  - Trained on "LJ Speech" audio dataset with 13,100 transcribed clips
  - Transformer architecture processes sequential data efficiently
  - Captures contextual relationships for accurate transcription
  - Optimizes model to recognize speech patterns and map to text

```
target:     <its present manual filing system is obsolete#>
prediction: <its present manual fileng systems is obsely t.>
```

```
target:     <the secret service and the department of the treasury now recognize this critical need.>
prediction: <the secret service and the department of the treasury now recognize this critical med.>
```

Low-Level Design

4. Federated Learning:-

- Four client models deployed:
  - Two for gesture classification (STCGN-LSTM)
  - Two for speech-to-text conversion (transformer-based)
- Each client trains locally on its respective dataset to maintain data privacy
- Local models periodically share parameters with a centralized server
- Server aggregates parameters to create a robust global model
- Approach improves generalization, learning from diverse data, and enhances performance without centralized data storage

```
Received weights from client ISL.
Sent updated ISL weights.
Connected to ('10.127.7.183', 50958)
Received weights from client ISL.
Sent updated ISL weights.
Connected to ('10.127.7.183', 50959)
Client ISL has completed training.
Connected to ('10.127.7.129', 51747)
Received weights from client STT.
Sent updated audio_to_txt weights.
Connected to ('10.127.7.131', 53451)
Received weights from client STT.
Sent updated audio_to_txt weights.
```

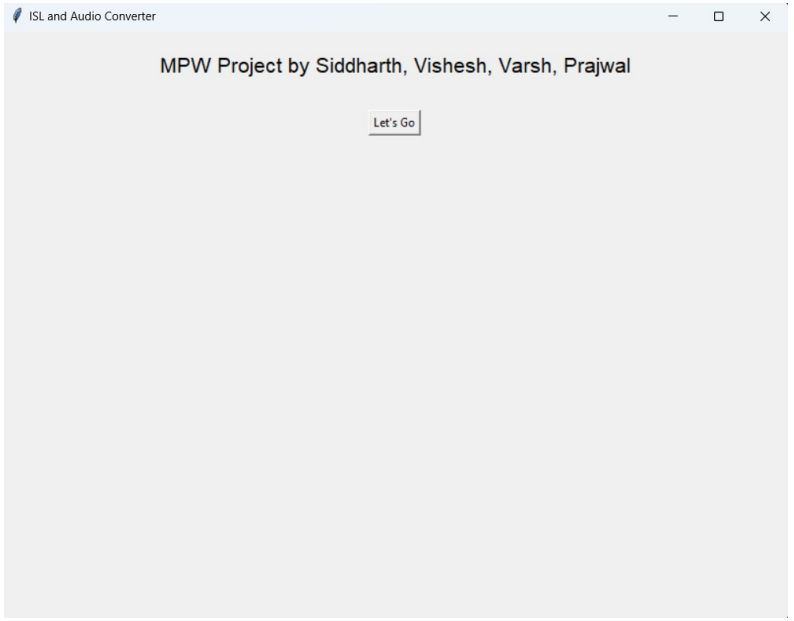
5. Mapping:-

- Visualizes gestures for a sentence using pose and hand skeleton data
- Links words/phrases to corresponding skeleton files for accurate gesture mapping
- Draws pose and hand landmarks on frames with distinct colors and connections
- Uses weighted blending over multiple frames for seamless gesture transitions
- Generates a video that reconstructs ISL gestures for the input sentence

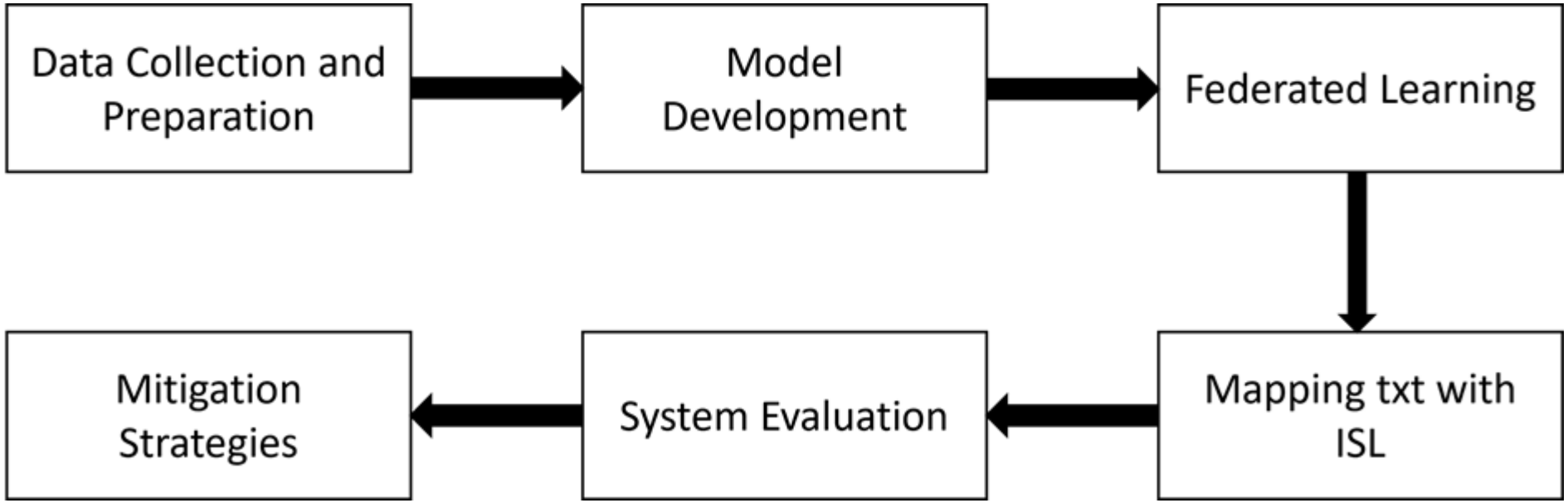
Low-Level Design

6. GUI:-

- User-friendly desktop application built with tkinter
- Supports ISL video and audio file uploads for text conversion
- Displays media and predicts translated text using backend functions
- Asynchronous video and audio processing with threading for smooth UI
- Real-time media playback and translation display
- Clear error handling and seamless file management for a smooth experience



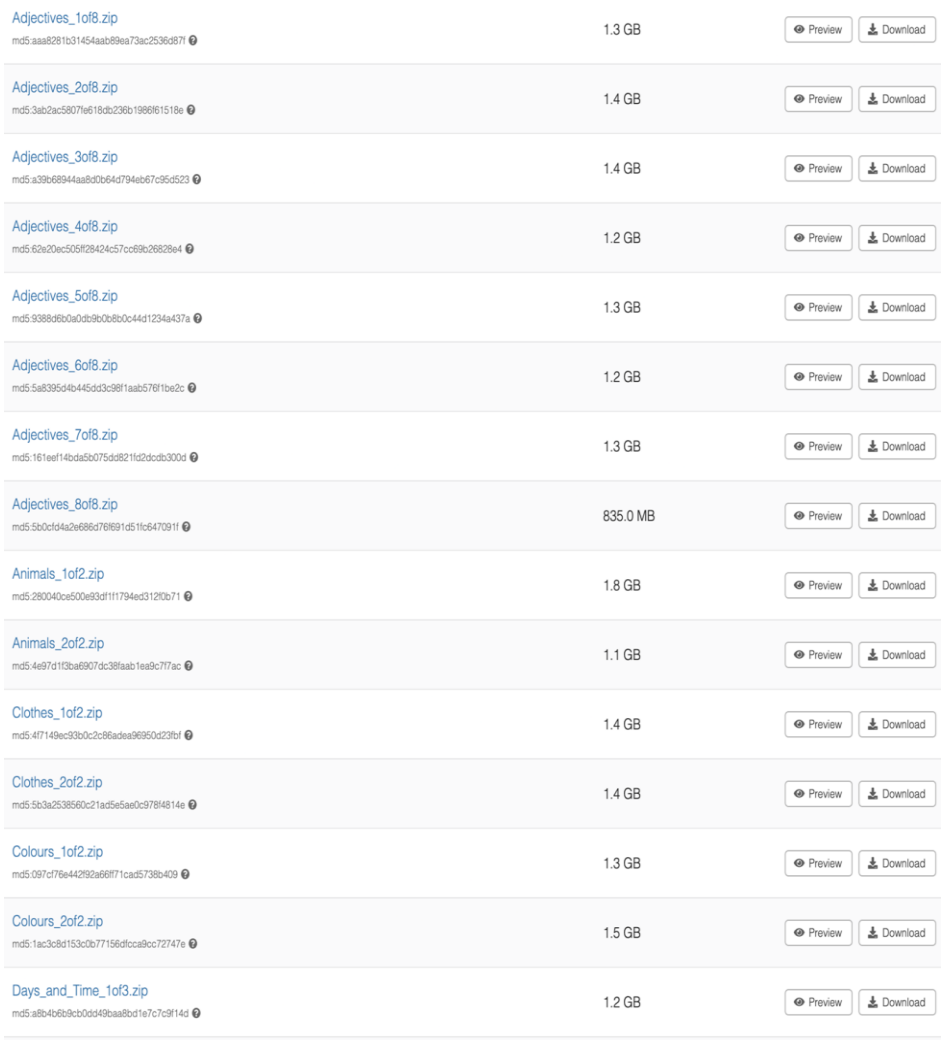
Methodology



Methodology

1. Data Collection and Preparation

- Utilizes both audio and video data for robust multimodal translation
- Audio data sourced from publicly available datasets (e.g., LJ Speech) and ISL-specific datasets
- Video data from INCLUDE, capturing gestures in varied environments for better generalizability
- Audio pre-processing using speech-to-text models (Wav2Vec, DeepSpeech) and feature extraction (MFCC, spectrograms)
- Video pre-processing using frame extraction and pose estimation (OpenPose) for gesture identification
- Data augmentation through noise addition for audio and geometric transformations for video to enhance model robustness



Methodology

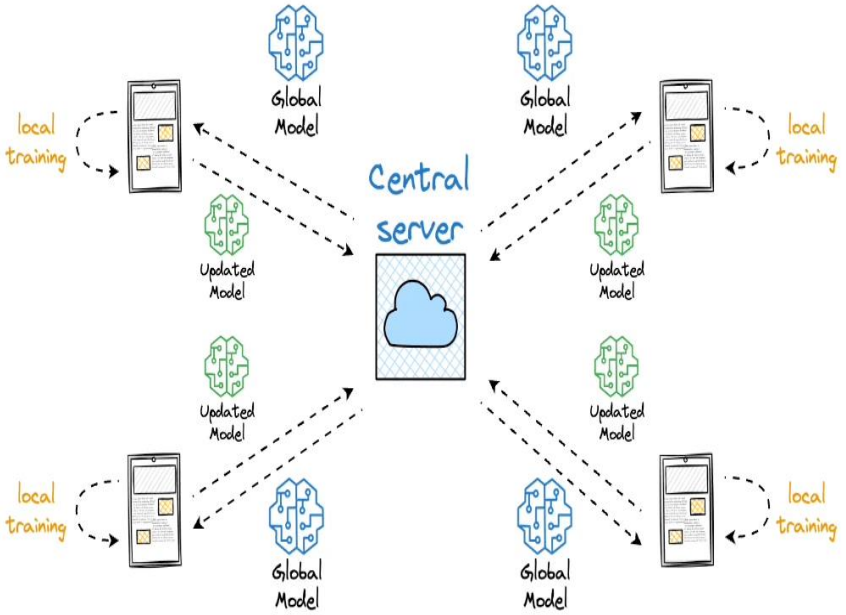
2. Model Development

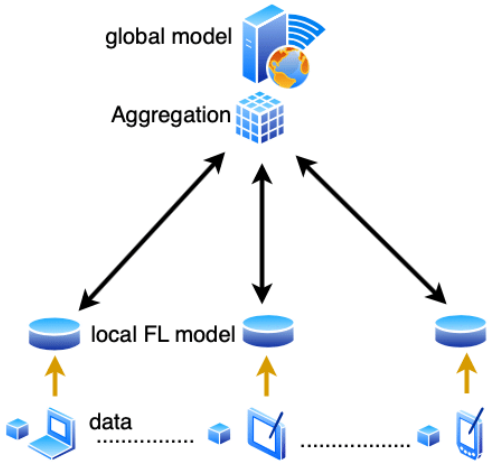
- The system comprises two core models: an audio-to-ISL model and a video-to-ISL model.

- 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.





Methodology

3. Federated Learning Implementation

- Federated learning enables decentralized training on local devices for audio and video data
- Central server aggregates model updates using the Federated Averaging algorithm
- Master server saves the best model weights during training
- Mapping mechanism links textual inputs to pre-saved ISL gesture files for accurate reconstruction
- ISL gesture database (video/animation files) mapped to corresponding text tokens
- Input text is pre-processed to align with ISL grammar for accurate mapping
- Mapped gesture files are retrieved and sequenced to reconstruct the ISL output

```
Received weights from client ISL.
Sent updated ISL weights.
Connected to ('10.127.7.183', 50958)
Received weights from client ISL.
Sent updated ISL weights.
Connected to ('10.127.7.183', 50959)
Client ISL has completed training.
Connected to ('10.127.7.129', 51747)
Received weights from client STT.
Sent updated audio_to_txt weights.
Connected to ('10.127.7.131', 53451)
Received weights from client STT.
Sent updated audio_to_txt weights.
```

```
686/686 [==============================] - 24s 14ms/step - loss: 5.5287 - accuracy: 0.007
Epoch 2/251
686/686 [==============================] - 9s 13ms/step - loss: 5.2297 - accuracy: 0.0128
Epoch 3/251
686/686 [==============================] - 9s 13ms/step - loss: 5.0096 - accuracy: 0.0157
Epoch 4/251
686/686 [==============================] - 9s 13ms/step - loss: 4.8197 - accuracy: 0.0350
Epoch 5/251
686/686 [==============================] - ETA: 0s - loss: 4.6361 - accuracy: 0.0368
Epoch 5: Sent best weights to master server.
Epoch 5: Updated model with global weights.
```

Methodology

4. System Evaluation

- Performance evaluated using accuracy and loss metrics
- k-fold cross-validation on collected datasets
- Simulated federated testing to assess model efficacy in distributed environments
- User testing with ISL practitioners to ensure practical usability
- Evaluation ensures system reliability in real-world scenarios

5. Challenges and Mitigation Strategies

- Challenges include data variability (speech accents, lighting conditions)
- Data augmentation and regularization techniques mitigate performance issues
- Provides a comprehensive framework for robust system development
- Evaluates system effectiveness across diverse conditions

Implementation

1. Local Model Training

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

- Audio-to-Text Model:
  - Model processes speech inputs into text
  - Trained using advanced speech-to-text frameworks
  - Utilizes curated datasets to handle diverse speech patterns and accents
  - Employs feature extraction techniques like MFCC and spectrograms
  - Aims to enhance accuracy in speech-to-text conversion
- ISL-to-Text Model:
  - Model recognizes ISL gestures and translates them into text
  - Trained on ISL gesture datasets for robust recognition
  - Utilizes STCGN+ LSTM model for accurate classification
  - Preprocessing includes pose estimation and frame extraction

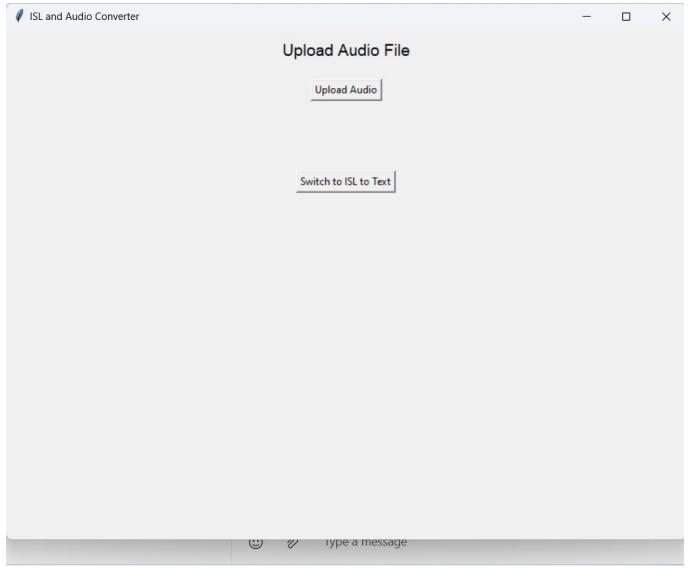
```
target:     <its present manual filing system is obsolete#>
prediction: <its present manual fileng systems is obsely t.>

target:     <it makes no use of the recent developments in automatic data processing which are widely
prediction: <it makes moutic developments in developments in developments in developments in developme

target:     <the secret service and the department of the treasury now recognize this critical need.>
prediction: <the secret service and the department of the treasury now recognize this critical med.>
```

```
Epoch 1/251
686/686 [==============================] - 21s 11ms/step - loss: 5.5051 - accuracy: 0.0080 - val_loss: 5.3185 - val_accuracy: 0.0175 - lr: 1.0000e-04
Epoch 2/251
686/686 [==============================] - 7s 10ms/step - loss: 5.2060 - accuracy: 0.0139 - val_loss: 5.0645 - val_accuracy: 0.0190 - lr: 1.0000e-04
Epoch 3/251
686/686 [==============================] - 7s 11ms/step - loss: 4.9996 - accuracy: 0.0208 - val_loss: 4.9616 - val_accuracy: 0.0175 - lr: 1.0000e-04
Epoch 4/251
686/686 [==============================] - 7s 10ms/step - loss: 4.8163 - accuracy: 0.0303 - val_loss: 4.7404 - val_accuracy: 0.0204 - lr: 1.0000e-04
Epoch 5/251
685/686 [============================>.] - ETA: 0s - loss: 4.6462 - accuracy: 0.0343
Epoch 5: Sent best weights to master server.
Epoch 5: Updated model with global weights.
686/686 [==============================] - 7s 10ms/step - loss: 4.6484 - accuracy: 0.0343 - val_loss: 4.6006 - val_accuracy: 0.0219 - lr: 1.0000e-04
Epoch 6/251
```

Implementation

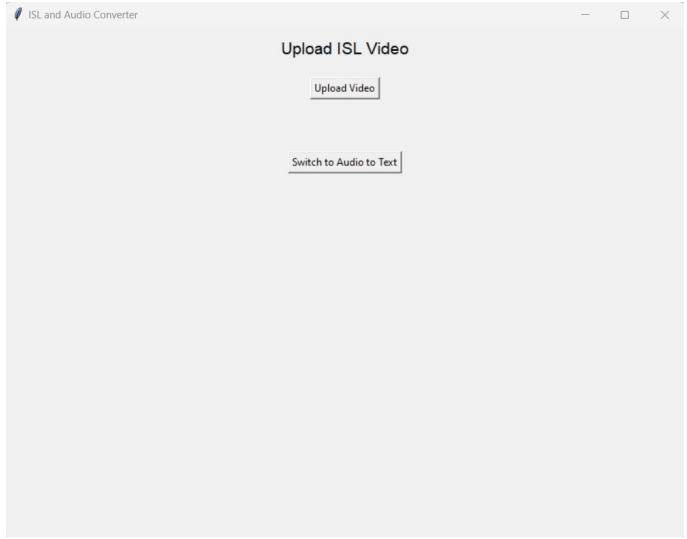
Graphical User Interface (GUI)

- **User-friendly GUI**: Developed to integrate trained models and enhance user experience.
- **ISL-to-Text Translation**: Users input ISL gestures via video; the GUI processes and translates gestures into text in real-time.
- **Audio-to-ISL Translation**: Users provide speech input; the GUI converts audio to text and maps it to ISL gestures for playback.
- **Reconstructed ISL Gestures**: Displayed as videos or animations within the interface for intuitive interaction.



Integration and Workflow

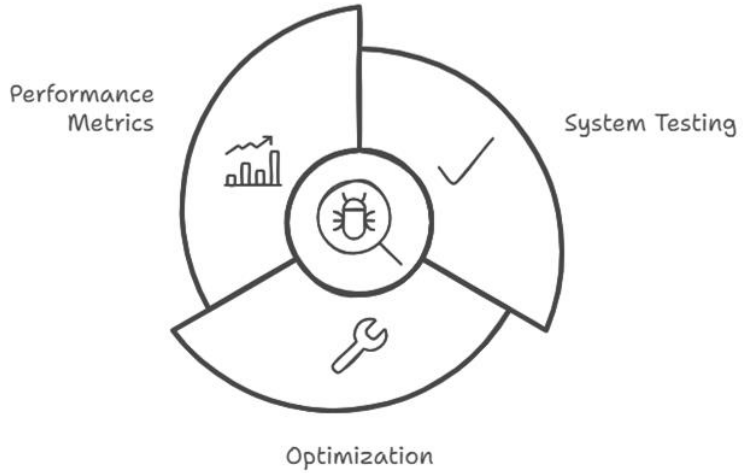
- **Seamless System Workflow**: Connects both models through the GUI for smooth interaction.
- **Audio-to-ISL Translation**: Audio-to-text model generates text, which is mapped to pre-defined ISL gestures.
- **ISL-to-Text Translation**: GUI integrates ISL-to-text model to process gesture inputs and produce real-time text outputs.



Implementation

Performance and Usability

- **Intuitive GUI**: Designed for users of varying technical proficiency to interact effectively.
- **Advanced Model Integration**: Combines machine learning models with an accessible interface for effective ISL translation.
- **Enhanced Accessibility**: Bridges communication gaps for individuals relying on ISL.
- **Paving the Way for Inclusivity**: Contributes to the development of a comprehensive multimodal translation system.



Experiment results and analysis

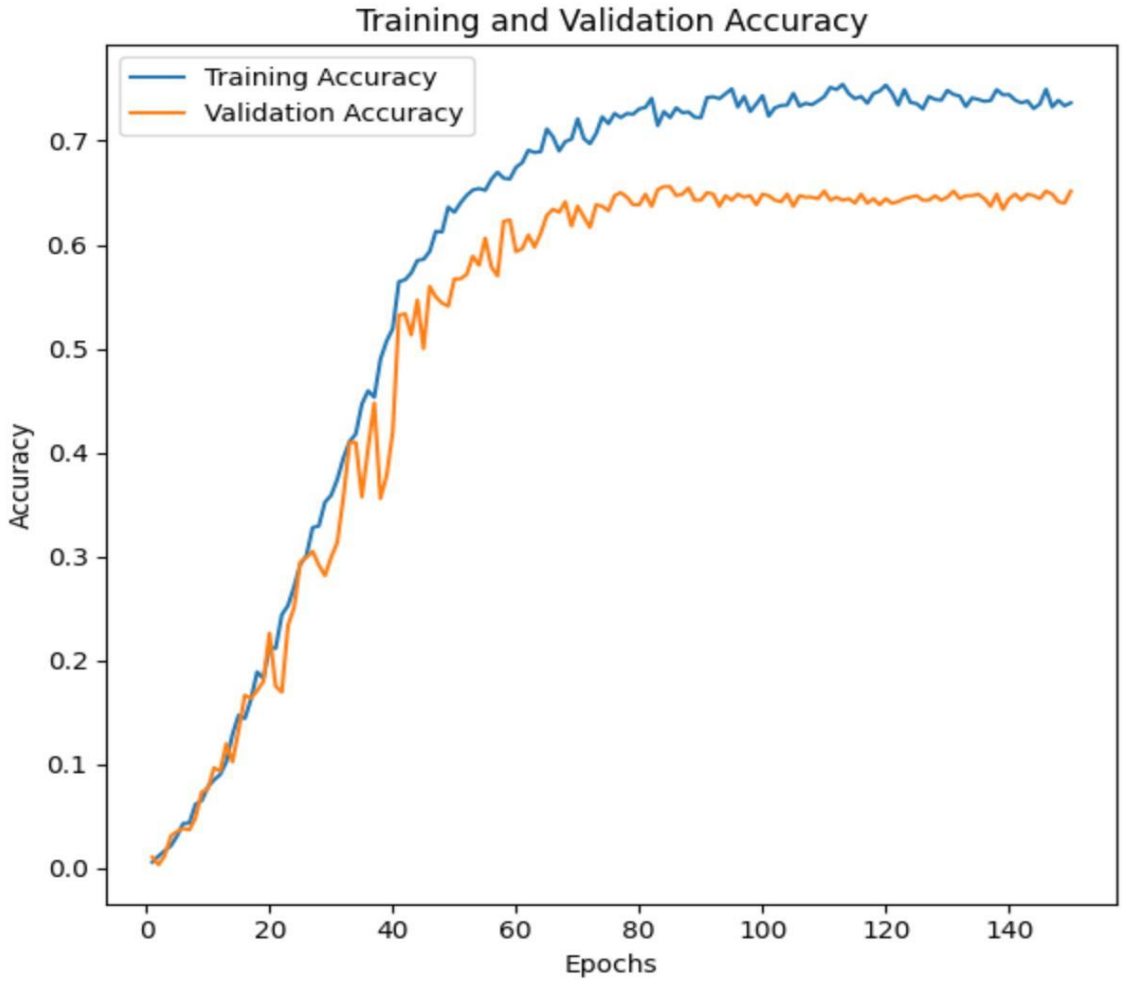
Audio-to-Text Translation Results

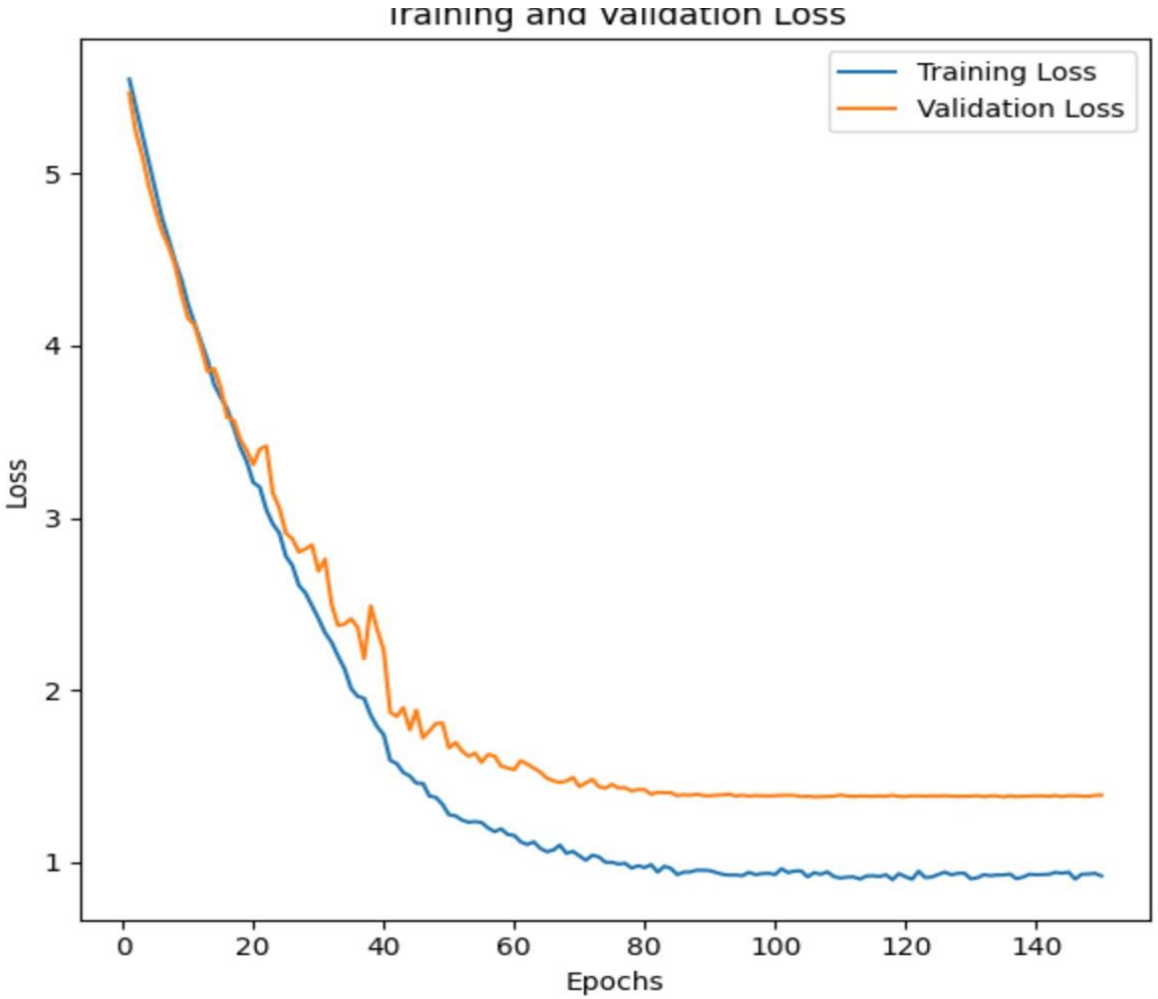
- **Evaluation of Audio-to-Text Model**: Tested with speech inputs of varying accents, speeds, and noise levels.
- **Metrics**: Used word error rate (WER) and transcription accuracy for performance evaluation.
- **Loss**: Achieved an average transcription loss of 0.415, with minor drop in noisy environments.
- **Performance Analysis**: Moderate performance for clear/moderate speech, with minor errors in homophone distinction and accented speech processing.

ISL-to-Text Translation Results

- **Evaluation of ISL-to-Text Model**: Tested with static and dynamic gestures under varied lighting and backgrounds.
- **Metrics**: Used gesture recognition accuracy and inference speed for performance evaluation.
- **Accuracy**: Achieved an average accuracy of ~67% for temporal videos.
- **Performance Analysis**: Effective in well-lit environments, but challenges in recognizing subtle hand movements, inconsistent speeds, and inaccurate sign language.

Experiment results and analysis

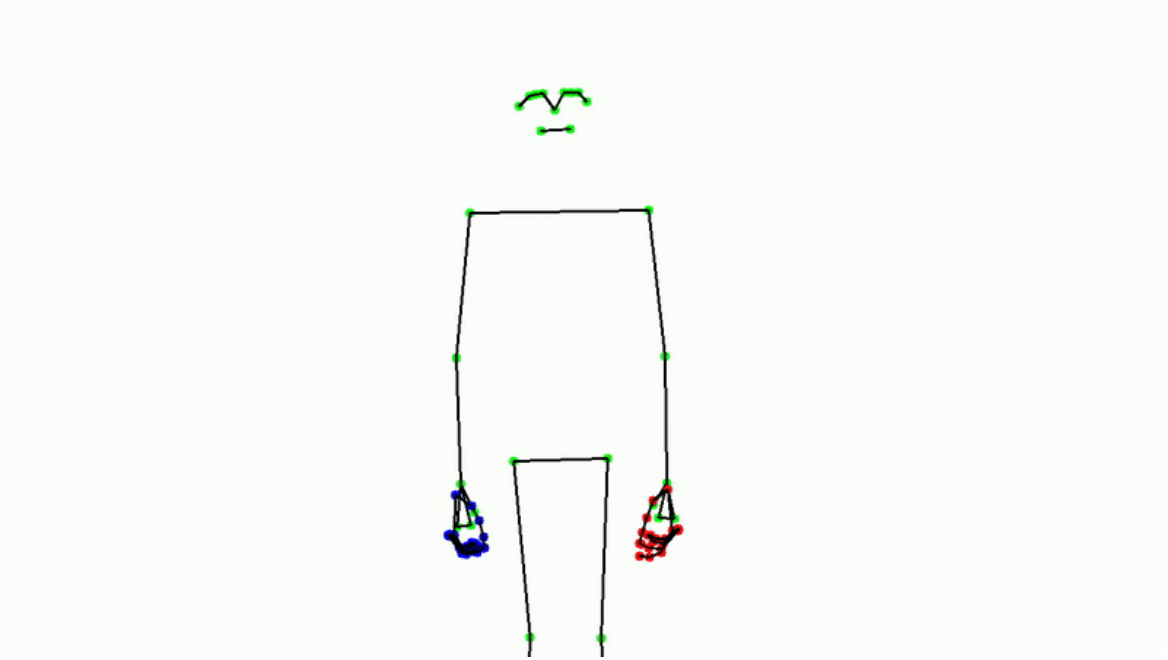


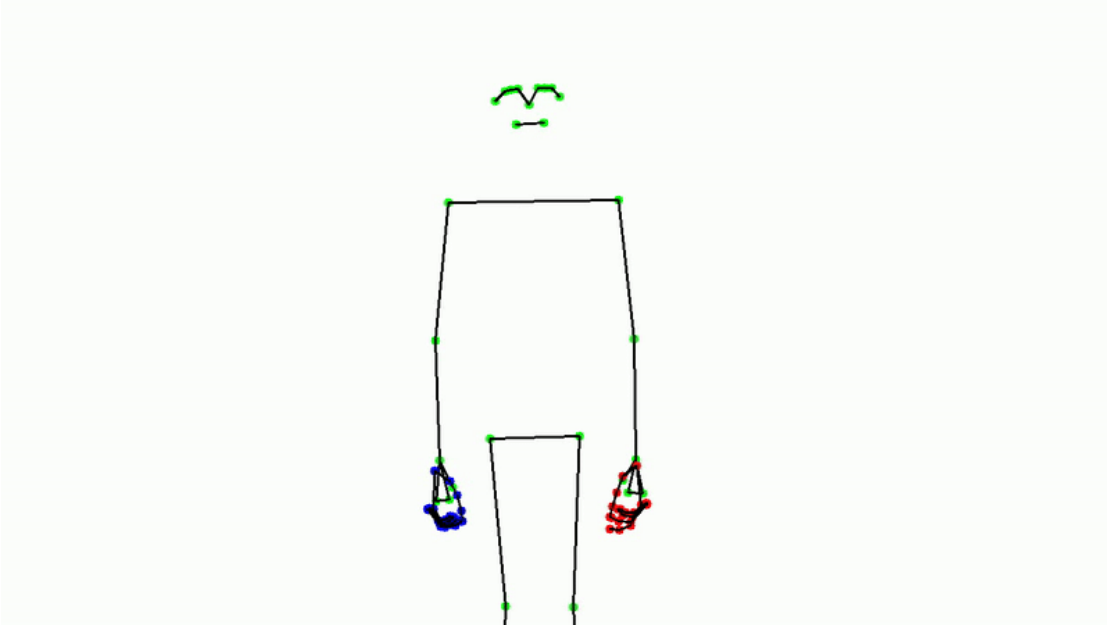


Experiment results and analysis

Text-to-ISL Translation Results

- **Audio-to-ISL Functionality**: Combines audio-to-text transcription with ISL gesture reconstruction.
- **Evaluation**: Conducted using speech inputs and compared ISL output with expected gestures.
- **Algorithm**: Utilizes a brute force approach for reconstructing ISL from pre-saved landmarks.
- **Data**: Based on a database of 6000 common words in sign language.





Testing and Validation

Audio-to-Text Translation Testing

- **Tested on Diverse Speech Datasets**: Includes accents, speech rates, and noise conditions.
- **Primary Evaluation Metrics**: Accuracy (transcription correctness) and loss (prediction error minimization).
- **Loss Curve Analysis**: Steady decline indicating effective error reduction over time.
- **Convergence**: Loss plateaued after a certain number of epochs, indicating optimal model state.

ISL-to-Text Translation Testing

- **Testing with Static and Dynamic ISL Gestures**: Evaluated using a dataset containing both types of gestures.
- **Evaluation Metrics:** Accuracy (gesture recognition correctness) and loss (training error).
- **Loss Graph Analysis**: Significant reduction in loss during early epochs, indicating improved predictions.
- **Convergence**: Loss stabilized after approximately 50 epochs, showing effective learning of ISL-text relationship.

Model: "functional"

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

Total params: 901,268 (3.44 MB)

Trainable params: 300,294 (1.15 MB)

Non-trainable params: 384 (1.50 KB)

Optimizer params: 600,590 (2.29 MB)

Testing and Validation

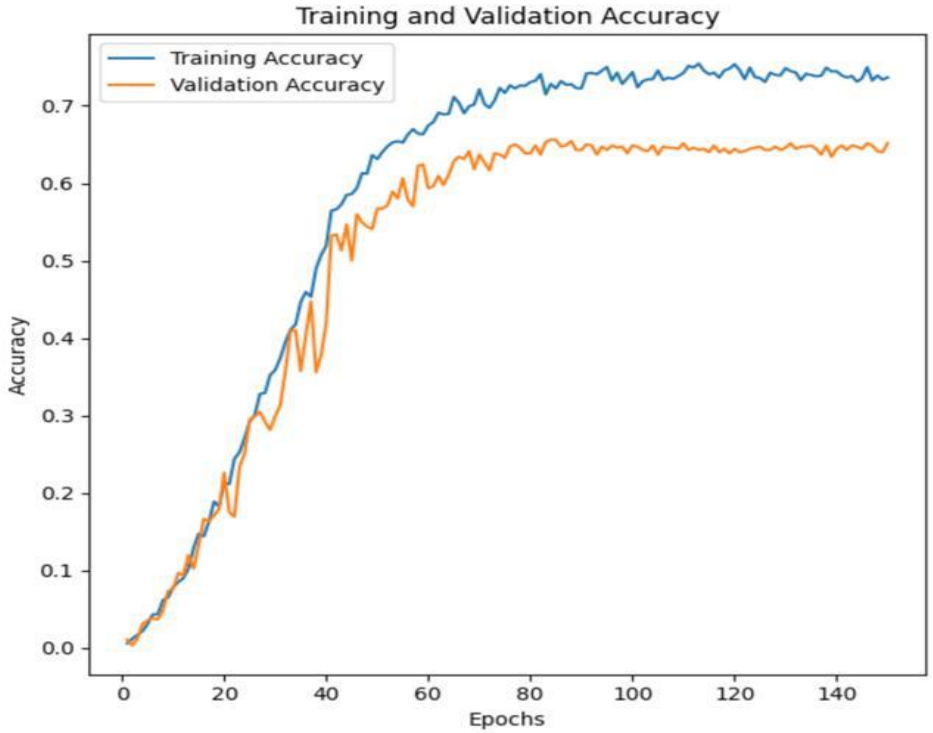
Accuracy Graph for ISL-to-Text Translation

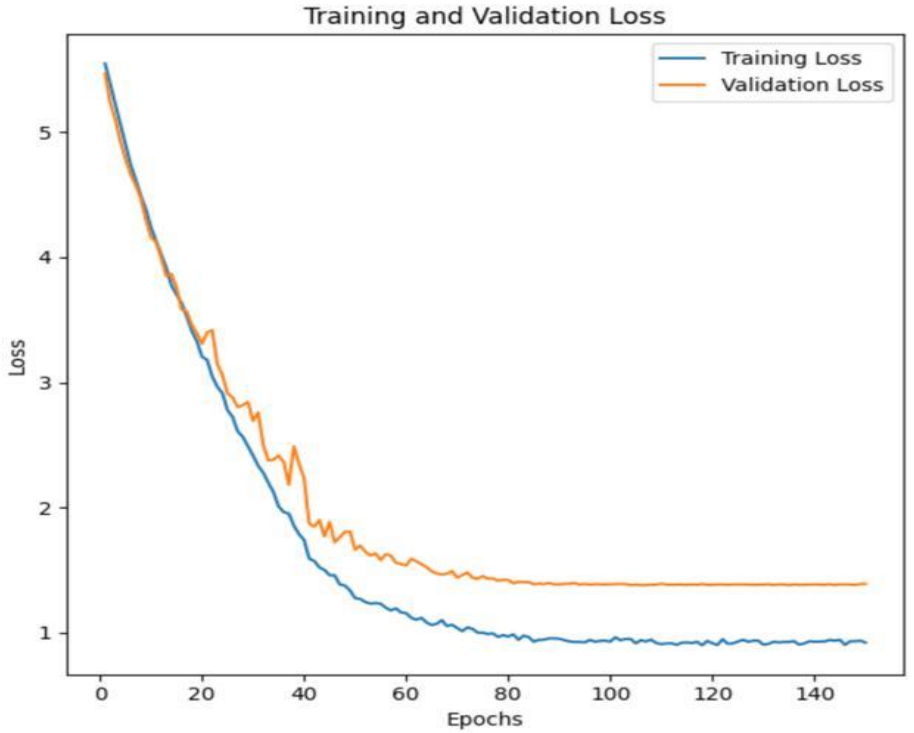
Accuracy Graph: Shows the percentage of correctly recognized gestures per epoch.

Evaluation Metric: Accuracy measures the model's ability to recognize and translate ISL gestures into text.

Progressive Improvement: Accuracy increased over epochs.

Final Accuracy: Model achieved a final accuracy of 67%.





Conclusion

This project introduces a multimodal AI system aimed at translating audio, text, and Indian Sign Language (ISL) while ensuring privacy through federated learning. The system integrates two key models: an audio-to-text model for speech recognition and an ISL-to-text model for gesture recognition. These models are seamlessly embedded into a user-friendly graphical user interface (GUI), allowing for easy interaction and real-time translation. The system demonstrated promising performance, with validation through loss and accuracy metrics showing its ability to generalize across diverse inputs. However, challenges such as handling noisy environments, subtle ISL gestures, and dynamic hand movements were encountered, prompting the application of techniques like data augmentation and model optimization.

Looking ahead, future improvements will focus on refining gesture recognition and enhancing real-time translation capabilities, particularly in handling complex ISL grammar and regional variations. These improvements will make the system more robust and accurate, providing a better experience for a wider range of ISL users. The integration of federated learning not only ensures the privacy of user data during training but also contributes to the system's potential for improving accessibility and inclusivity in communication. With continuous enhancements, this project holds the potential to significantly bridge communication gaps for individuals relying on ISL.

References

- M. Kowsigan, R. Dhawan, and A. Kundu, "An Efficient Speech to Sign Language Conversion and Text Recognition through Live Gesture," 2024.
- J. P. Jeevanandham, G. B. A., and H. A., "Real-Time Hand Sign Language Translation: Text and Speech Conversion," 2024.
- M. Kavitha, A. Chatterjee, S. Shrivastava, and G. Sarkar, "Formation of Text from Indian Sign Language using Convolutional Neural Networks," 2022.
- MediaPipe Framework Study, "Sign Language Recognition Using Neural Networks and MediaPipe Framework," 2023.
- Federated Learning Study, "Federated Learning for Privacy-Preserving Sign Language Translation Systems," 2023.
- Mathurina Chelliah, "The LJ Speech Dataset," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset. [Accessed: 04-Jan-2025].
- 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].
- 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].

Thank You