VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Project Report on

Gesturely: Sign Language to Sentences

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

By

Piyush Chugeja D12B / 11

Sakshi Kirmathe D12B / 25

Deven Bhagtani D12B / 06

Project Mentor

Dr. (Mrs.) Nupur Giri

University of Mumbai

(AY 2023 - 24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Certificate

This is to certify that the Mini Project entitled **"Gesturely: sign language to sentences"** is a bonafide work of **Piyush Chugeja (D12B - 11), Sakshi Kirmathe (D12B - 25), & Deven Bhagtani (D12B - 06)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering".**

Dr. (Mrs.) Nupur Giri
Mentor, Head of Department

Dr. (Mrs.) J. M. Nair
Principal

Mini Project Approval

This Mini Project entitled **"Gesturely: sign language to sentences"** by **Piyush Chugeja (D12B - 11), Sakshi Kirmathe (D12B - 25), & Deven Bhagtani (D12B - 06)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1. ……
(Internal Examiner name & sign)

2. ……
(External Examiner name & sign)

Date: 01st April 2024

Place: Chembur, Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Piyush Chugeja - D12B / 11)

(Sakshi Kirmathe - D12B / 25)

(Deven Bhagtani - D12B / 06)

Date: 01st April 2024

Acknowledgement

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to the Head of the Computer Department **Dr. (Mrs.) Nupur Giri** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to our Principal **Dr. (Mrs.) J.M. Nair**, for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult to finish this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Index

Abstract

This project aims to improve communication within educational settings for individuals with hearing impairments. Sign language, notably Indian Sign Language (ISL) in India, serves as a primary mode of expression for the deaf community. The form of expression among the deaf, relies on a rich vocabulary of gestures involving fingers, hands, arms, eyes, head, and face. Our research endeavors to develop an algorithm capable of translating ISL into English, focusing initially on words within the education domain. Through the integration of advanced computer vision and deep learning methodologies, our objective is to create a system capable of interpreting ISL gestures and converting them into written text. The project involves the creation of a comprehensive dataset, with 50 words and over 2500 videos. Our vision is to empower the deaf community with real-time translation capabilities, promoting inclusivity and accessibility in communication.

List of abbreviations

Sr no.	Short form	Abbreviated form
1	ISL	Indian Sign Language
2	ASL	American Sign Language
3	BSL	British Sign Language
4	DGS	Deutsche Gebärdensprache
5	CNN	Convolutional Neural Network
6	RNN	Recurrent Neural Networks
7	LLM	Large Language Model
8	NLP	Natural Language Processing
9	TTS	Text to Speech
10	LSTM	Long Short Term Memory
11	VideoMAE	Video Masked Autoencoders
12	SVM	Support Vector Machines

List of figures

List of tables

Chapter I: Introduction

1.1 Introduction

Communication is vital for human interaction, yet individuals who are mute or deaf face challenges connecting with the hearing community. Sign language serves as a vital means of communication for the deaf community, employing gestures, facial expressions, and body movements to convey messages. In India, Indian Sign Language (ISL) is widely used and recognized as the primary mode of communication. As per the 2011 Census, the total population of deaf persons in India numbered about 50 lakh.[1] Despite its importance, there remains a significant challenge in facilitating effective communication between sign language users and those unfamiliar with sign language. This communication gap underscores the need for innovative solutions that can bridge linguistic barriers and enhance inclusivity. There has been relatively limited focus on ISL due to the scarcity of large annotated datasets. We aim to address a new translation dataset focused on ISL, with a particular emphasis on the education domain. Additionally, we introduce a deep learning model for classifying gestures.

1.2 Motivation

Our motivation for undertaking this project stems from our desire to apply our knowledge and skills to drive positive change. By applying what we've learned, we aim to bridge the communication gap between the hearing and speech-impaired communities through technology. We aim to develop a system that can accurately interpret sign language gestures and translate them into understandable text, thereby facilitating seamless communication between sign language users and the broader community.

1.3 Problem Statement

The core issue we aim to address in this project is the pervasive communication barrier that exists between sign language users and those who do not comprehend sign language. This deficiency in a common language results in impediments to effective interaction in a range of scenarios, including social exchanges, job interviews, educational environments, and day-to-day conversations. Existing methods of translation often fall short in accurately capturing the nuances of sign language gestures, leading to misunderstandings and communication breakdowns. Our goal is to develop a robust system that can overcome these challenges and provide accurate translations in real-time.

[1] https://islrtc.nic.in/history-0

1.4 Existing Systems

Several existing systems attempt to address the translation of sign language gestures into text or speech. These systems vary in complexity and effectiveness, with some relying on simple gesture recognition techniques, while others employ more advanced machine learning algorithms. However, many of these systems still struggle to accurately interpret the subtleties of sign language, resulting in inaccurate translations and limited usability.

1.5 Lacuna of existing systems

Despite the progress made in sign language translation technology, there remains a significant gap in the accuracy and reliability of existing systems. Many systems fail to account for the diversity and complexity of sign language gestures, leading to errors and inaccuracies in translation. Furthermore, while ASL benefits from systems like the SignAll SDK[2], facilitating smoother communication for its users, ISL lacks comparable existing systems. This disparity underscores the pressing need for effective ISL translation systems to address the communication challenges faced by the deaf and hard of hearing community in India.

1.6 Relevance of the project

The relevance of this project lies in its potential to significantly improve the accessibility and inclusivity of communication for sign language users. By developing a more accurate and reliable system for translating sign language gestures into text, we aim to empower individuals with hearing impairments to communicate more effectively in a variety of settings. This project has the potential to positively impact the lives of millions of people worldwide by breaking down communication barriers and fostering greater understanding and connection.

[2] https://developers.googleblog.com/2021/04/signall-sdk-sign-language-interface-using-mediapipe-now-available.html

Chapter II: Literature survey

A: CISLR: Corpus for Indian Sign Language Recognition [1]

Authors: Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi.

The authors devise the CISLR dataset to address ISL's word-level recognition requirements. Their dataset, encompassing over 4700 words across diverse topics, is coupled with a proposed model for word recognition from ISL videos. The methodology integrates a prototype-based one-shot learner, leveraging ASL resources to enhance ISL predictions. The dataset offers a rich resource for ISL word recognition, enabling diverse research applications. The use of ASL resources enhances ISL prediction accuracy, addressing resource limitations. Although the dataset caters to ISL word recognition needs, its one-shot learner model may pose limitations in scalability and generalization.

B: Hand Gesture Identification using Mediapipe [2]

Authors: Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, and Jyoti Dange.

The paper explores a sign language recognition system using the Mediapipe framework for hand gesture detection and interpretation. Leveraging machine learning, the framework identifies 21 3D landmarks on the hand, enabling real-time tracking. The system's emphasis lies on recognizing ASL alphabet characters using the KNN algorithm, restricting its applicability to other sign languages or gestures. The Mediapipe-based system facilitates real-time hand gesture interpretation, particularly beneficial for ASL alphabet recognition. However, the system's focus on ASL alphabet recognition limits its applicability to other sign languages or gestures.

C: ISLTranslate: Dataset for Translating Indian Sign Language [3]

Authors: Abhinav Joshi, Susmit Agrawal, Ashutosh Modi

Authors of this paper introduce the ISLTranslate dataset, consisting of 31,222 ISL-English pairs for enhancing ISL translation systems. They propose the Pose-SLT model for ISL-to-English translation, integrating pose estimation models and transformer architecture. While prioritizing pose over images for quicker inference, their gloss-free approach differs from our focus on translating glosses for a more detailed understanding. The ISLTranslate dataset introduced in provides a significant resource for ISL translation systems, incorporating linguistic priors. The Pose-SLT model offers efficient real-time translation, prioritizing pose estimation for faster inference. However, The gloss-free approach adopted in the ISLTranslate dataset and Pose-SLT model may overlook nuances in ISL communication, particularly regarding gloss-specific translations.

Chapter III: Requirement Gathering for the Proposed System

3.1 Introduction to requirement gathering

Requirement gathering is the crucial initial step in any project, where we gather and document the needs, expectations, and objectives from stakeholders. It involves communicating with end-users, clients, and other relevant parties to understand the problem at hand, its scope, and the desired outcomes. This process lays the foundation for the entire project, guiding decisions and ensuring alignment between the final product and stakeholder expectations.

3.2 Functional requirements

Functional requirements specify the specific functionalities and features that the system must provide to meet the needs of stakeholders and users. [4] Functional requirements for our system are:

- Gesture recognition: The system should be able to accurately recognize Indian Sign Language (ISL) gestures from video input in real-time.
- Gloss classification: The system should classify ISL gestures into specific glosses or words, allowing users to understand the meaning conveyed by the gestures.
- Translation functionality: The system should translate ISL glosses into English text or vice versa, enabling communication between ISL users and non-ISL users.
- Feedback mechanism: Users should have the option to provide feedback on the accuracy of gesture recognition and translation results to improve system performance over time.
- Accessibility features: The system should be accessible to users with disabilities, providing alternative input methods or interfaces as needed.

3.3 Non functional requirements

Non-functional requirements define the quality attributes and constraints that the system must adhere to [4]. In our project, non-functional requirements includes:

- Performance: The system should process ISL gestures and generate translated sentences with minimal delay, providing real-time translation capabilities to users.
- Reliability: The system should be available and reliable, ensuring uninterrupted translation services for users without frequent downtime or service disruptions.

- Scalability: The system should be able to handle a growing number of users and increasing translation requests over time, scaling resources such as servers and computational power as needed.
- Usability: The system's user interface should be intuitive and easy to use, allowing users to input ISL gestures and receive translated sentences without requiring extensive training or technical knowledge.

3.4 Hardware, Software, Technology and tools utilized

- Hardware
  - Standard computing hardware including processors, memory, storage device
- Software
  - Python as primary programming language
  - Mediapipe Pose framework for coordinates extraction [5]
- Technology
  - Tensorflow to implement neural networks
  - Google Gemini

3.5 Constraints

- Limited Data Availability: Annotated ISL datasets were scarce, posing challenges during the training and evaluation of our model, potentially affecting its performance and generalization capabilities.
- Computational Resource Constraints: We faced limitations in computational resources, such as processing power and memory, impacting the efficiency and scalability of our system, especially during training and inference stages.
- Model Accuracy Challenges: Achieving high accuracy in gesture recognition and translation tasks proved challenging due to the complexity and variability of ISL gestures, leading to errors in translation and misinterpretation.
- Real-time Performance Concerns: Ensuring real-time performance of the system, particularly during gesture recognition and translation, was essential for providing timely and seamless communication support to users. Delays or latency issues could hinder the user experience.
- Language Variability Considerations: ISL encompasses a wide range of regional and cultural variations, resulting in variability in gestures and expressions. Adapting the model to effectively recognize and translate these variations posed a significant challenge.

Chapter IV: Proposed Design

4.1 Block diagram of system

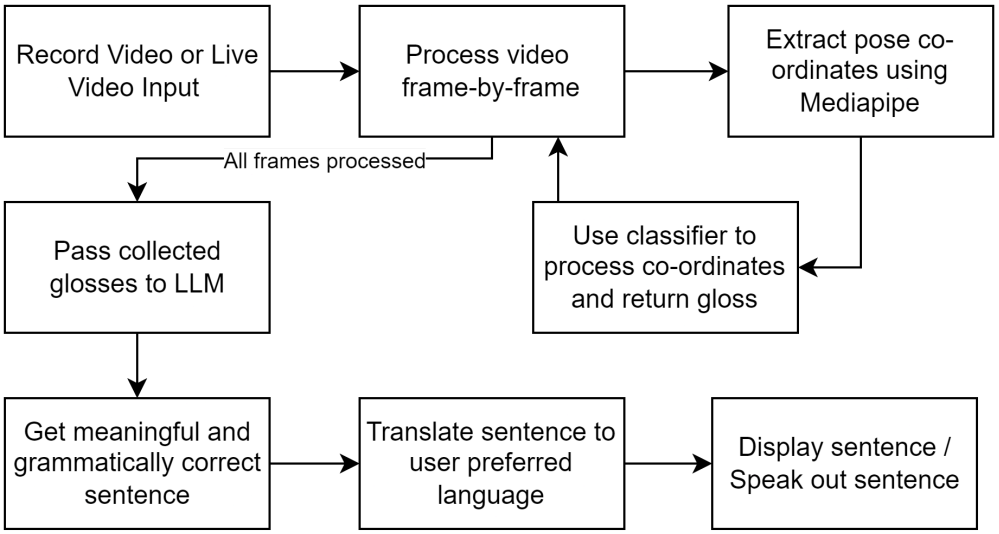


Figure 4.1 Block diagram of system

- Record Video or Live Video Input: The process starts by capturing video input, either recorded or live.
- Process Video Frame-by-Frame: Each frame of the video is analyzed individually.
- Extract Pose Coordinates Using Mediapipe: Mediapipe extracts pose information, including hand gestures and body movements.
- Classifier Processes Coordinates and Returns Gloss: A classifier interprets the pose coordinates and identifies corresponding glosses (individual signs).
- Pass Collected Glosses to LLM: The glosses are fed into a Large Language Model (LLM).
- Get Meaningful and Grammatically Correct Sentence: LLM generates a coherent spoken language sentence based on the glosses.
- Translate Sentence to User's Preferred Language: The sentence can be translated into the user's chosen language.
- Display Sentence / Speak Out Sentence: Finally, the translated sentence is either displayed visually or spoken aloud.

4.2 Modular design of the system

The modular design of our system encompasses several key components, each responsible for specific functionalities to ensure the overall effectiveness and efficiency of the system.

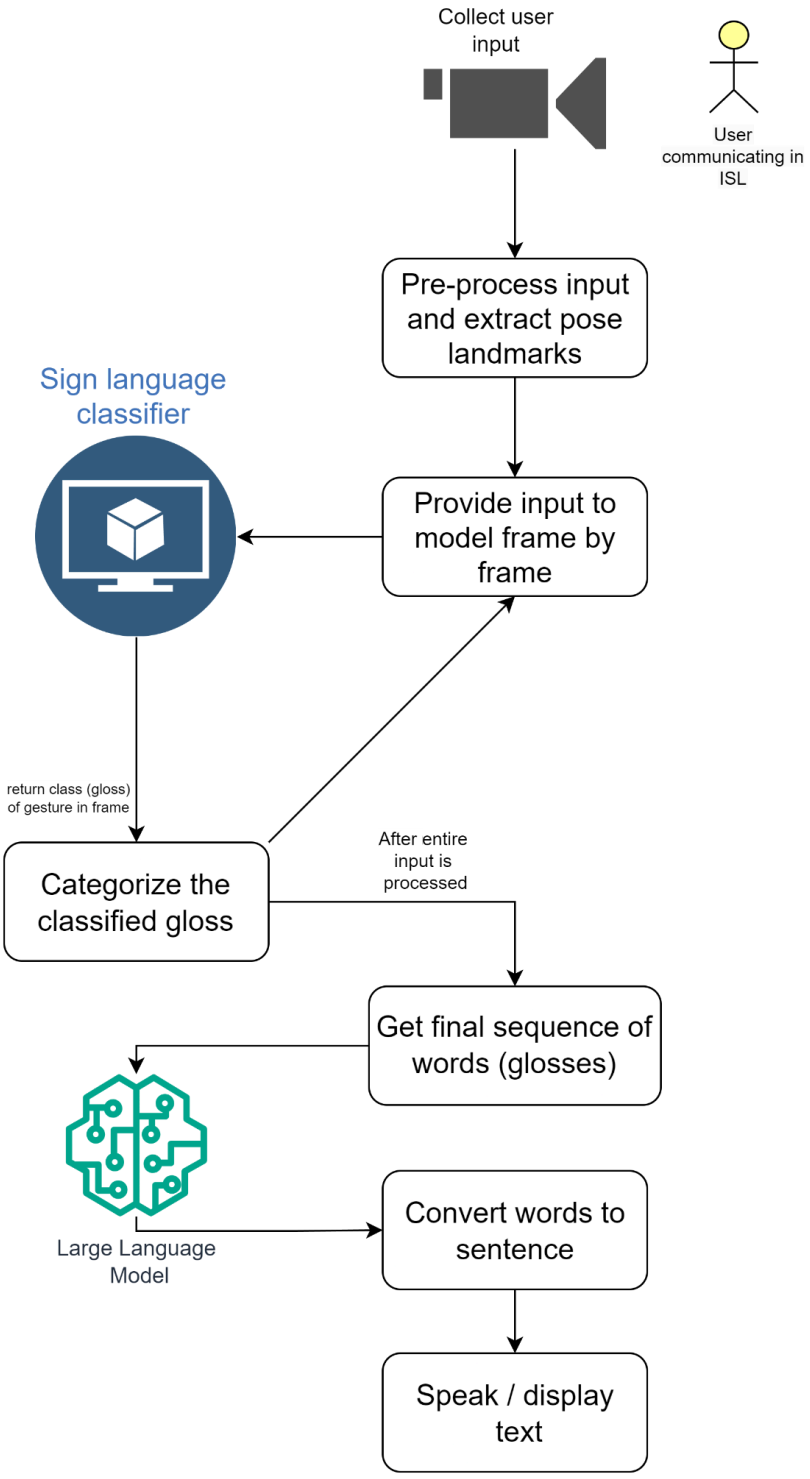


Figure 4.2 Architecture of system

- Input Module: This module is responsible for collecting user input, which can be in the form of live video streams or pre-recorded videos containing ISL gestures. It interfaces with external devices such as cameras or video files to capture the input data.
- Preprocessing Module: Upon receiving the input data, the preprocessing module performs necessary data cleaning and formatting tasks. This includes extracting pose

landmarks using Mediapipe and filtering out irrelevant information to prepare the data for further processing.

- Feature Extraction Module: Once the data is preprocessed, the feature extraction module extracts relevant features from the input data. It utilizes machine learning algorithms to identify distinctive patterns and characteristics in the pose landmarks, which are crucial for accurate gesture recognition.
- Classification Module: The classification module is responsible for predicting the corresponding gloss or word associated with each gesture. It employs a deep learning model, such as a Conv1D or LSTM neural network, trained on labeled ISL data to classify gestures into predefined categories.
- Translation Module: After classifying the gestures, the translation module translates the recognized glosses into English sentences. It leverages language processing techniques and possibly external APIs, such as Google's Gemini API, to generate grammatically correct and contextually relevant translations.
- Output Module: Finally, the output module presents the translated sentences to the user through a user-friendly interface. It may display the translated text on a screen, read it aloud using text-to-speech technology, or provide other means of conveying the information to the user effectively.

4.3 Project Scheduling & Tracking using Timeline / Gantt Chart

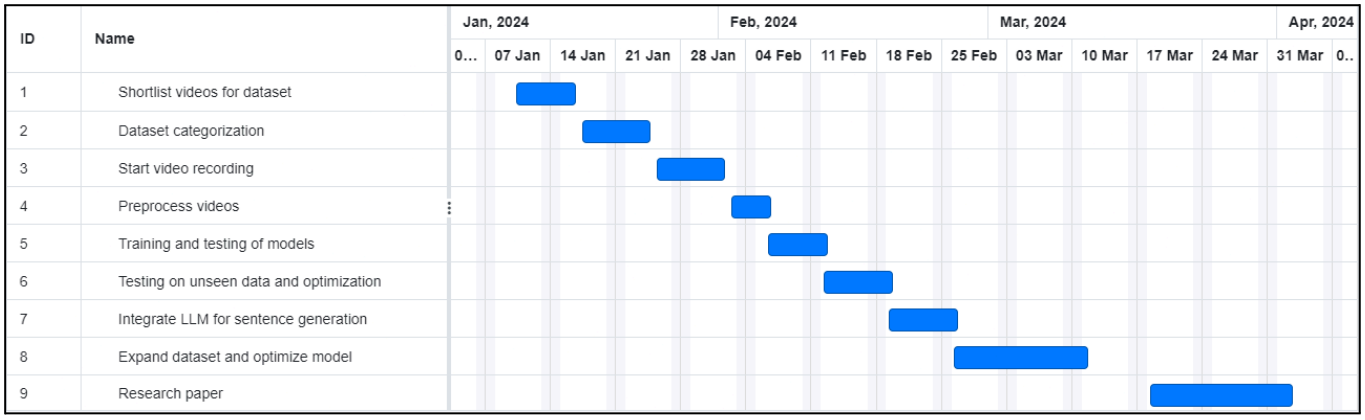


Figure 4.3 Gantt chart

Chapter V: Implementation of Proposed System

5.1 Methodology employed for development

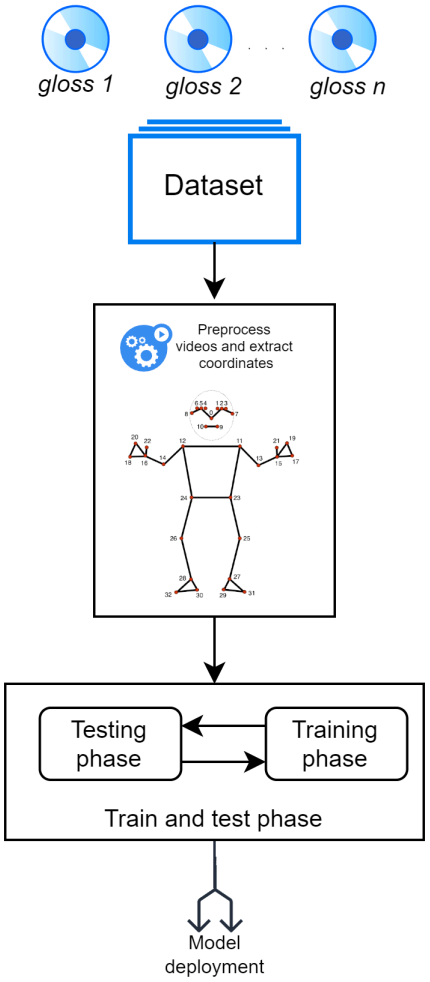


Figure 5.1 Methodology followed

- Data Collection: Our dataset comprises a self-created collection of Indian Sign Language (ISL) videos, influenced by CISLR and featuring 57 categories. Emphasizing the education domain, we curated over 2500 videos, each representing 50+ glosses. Multiple videos per gloss were included to enrich the training data and ensure diversity.

- Pre-processing and Feature Extraction: Each video undergoes frame-by-frame processing, where pose coordinates are extracted from each frame as shown in Fig. 5.2. Mediapipe provides us with 33 coordinates detailing the human body from head to toe. However, as depicted in Fig. 5.3, body parts below the waist remain unseen, making coordinates below the waistline unusable. This discrepancy in data could introduce inconsistencies. To rectify this issue, all unused coordinates per frame, specifically coordinates 25 to 32 (8 coordinates), are filtered out before feeding the data to the model.

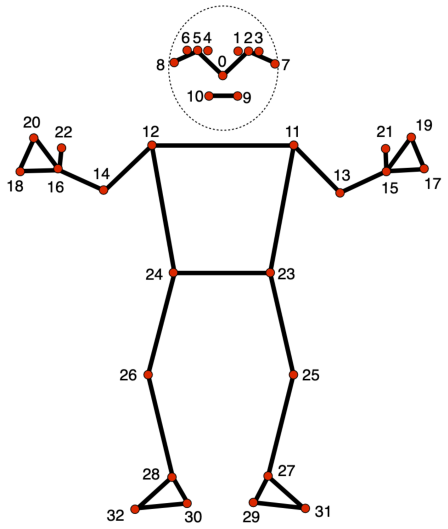


Figure 5.2 Mediapipe pose body landmarks

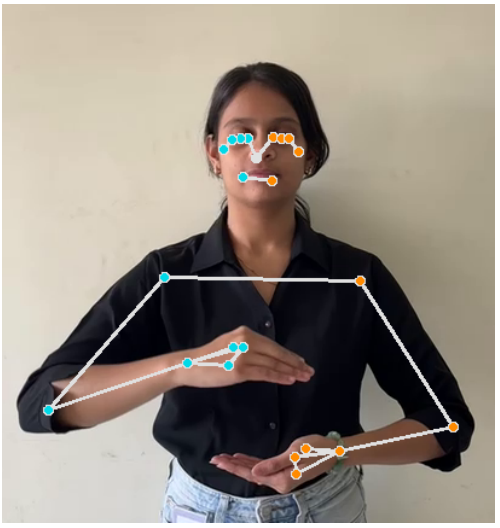


Figure 5.3 Extracting coordinates frame by frame

- Model Development: Following pre-processing and filtering, we trained classification models tailored to ISL recognition. Our approach encompassed a diverse range of models, including variants such as VideoMAE and 3D CNN, each designed to tackle the unique challenges of ISL recognition. Using transfer learning, we fine-tuned pre-trained models like VideoMAE to adapt them to our dataset. Additionally, we developed custom architectures inspired by [6], focusing on distinguishing ISL glosses using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks based on frame-by-frame coordinates. These models were trained using optimization algorithms like Adam, with carefully selected hyperparameters to ensure effective learning. Rigorous experimentation and validation were conducted throughout the model development process to ensure reliability and applicability.
- Evaluation: In evaluation, we assess the performance of the developed model based on various parameters, including accuracy, precision, recall and test loss. We evaluate how well the model works when the data isn't seen or is retrieved real-time. We also test if the model is unbiased and doesn't favor a particular class.

5.2 Algorithms and flowcharts for the respective models developed

Algorithm for fine tuned VideoMAE model

- Initialization Initialize VideoMAE model architecture with pre-trained weights from a large-scale dataset.
- Data Preprocessing: Preprocess ISL videos to extract frames at a sample rate of 4. Resize each frame to 224x224 pixels.
- Model Modification: Fine-tune the pre-trained VideoMAE model on the ISL dataset. Adapt the top layers of the VideoMAE model for ISL gesture recognition.

- Training: Train the modified VideoMAE model with a learning rate of 0.001. Use categorical crossentropy loss to optimize model performance.
- Evaluation: Evaluate model performance using validation split.

Algorithm for SVM model

- Initialization: Initialize SVM classifier with a linear kernel and a regularization parameter (C) set to 1.0.
- Data Preprocessing: Extract features from ISL dataset.
- Training: Train the SVM model on the preprocessed ISL dataset.
- Evaluation: Evaluate model performance using appropriate metrics.

Algorithm for 3D CNN model

- Initialization: Design 3D CNN architecture with 4 hidden neural network layers.
- Data Preprocessing: Preprocess ISL videos to extract pose landmarks from each frame.
- Model Compilation: Compile the 3D CNN model with the Adam optimizer and softmax activation function. Use categorical crossentropy loss for model optimization.
- Training: Train the 3D CNN model for 60 epochs with a batch size of 16. Validate model performance using a validation split of 20%.
- Evaluation: Evaluate model performance on test data.

Algorithm for CNN-LSTM model

- Initialization: Design CNN LSTM architecture with hidden neural network layers.
- Data Preprocessing: Preprocess ISL videos to extract pose landmarks from each frame.
- Model Compilation: Compile the CNN LSTM model with the Adam optimizer and ReLU activation function. Use sparse categorical crossentropy loss for model optimization.
- Training: Train the CNN LSTM model for 10 epochs with a batch size of 32. Validate model performance using a validation split of 20%.
- Evaluation: Evaluate model performance on test data.

5.3 Datasets source and utilization

- Our dataset is a self-created collection of ISL video recordings, to address the unique requirements of our research.

- The dataset is created from the videos provided by CISLR [1], which has 57 distinct categories but a limited number of videos per gloss which makes it difficult to train classification models.
- Within our dataset, we emphasize the education domain, comprising more than 50 glosses distributed across 2500 videos.
- All videos maintain a consistent format, adhering to a 1:1 aspect ratio and recorded with a 720p resolution at 30 frames per second. This standardized recording setup ensures uniformity and quality across the dataset.
- Each gloss is represented by multiple videos, with diverse angles and lighting conditions to improve the training data and enhance model robustness.
- As a self-created resource, our dataset serves as a valuable asset for training and evaluating machine learning models tailored specifically for ISL recognition and translation.
- By using this authentic dataset, our research aims to bridge communication gaps and promote inclusivity for individuals with hearing impairments in real-world settings.

Chapter VI: Testing of the Proposed System

6.1 Introduction to testing

Testing is a critical phase in the development lifecycle, ensuring the reliability, functionality, and performance of the system before deployment. It involves systematically evaluating the system's behavior under different conditions to uncover defects, validate functionality, and ensure that it meets the specified requirements.

6.2 Types of tests considered

- Unit Tests: These tests focus on individual components or modules of the system, verifying their functionality in isolation.
- Integration Tests: Integration tests validate the interactions and interfaces between different components or modules to ensure they work together seamlessly.
- Functional Tests: Functional tests assess the system's behavior against functional requirements, ensuring that it performs as expected from an end-user perspective.
- Performance Tests: Performance tests evaluate the system's responsiveness, scalability, and stability under various load conditions to identify bottlenecks and optimize performance.
- Usability Tests: Usability tests assess the system's user-friendliness, intuitiveness, and accessibility to ensure a positive user experience.
- Security Tests: Security tests identify vulnerabilities and weaknesses in the system's security mechanisms, protecting against potential threats and breaches.

6.3 Various test case scenarios considered

- Recognition Accuracy: Testing the accuracy of ISL gesture recognition against known glosses to ensure precise interpretation.
- Real-time Performance: Assessing the system's responsiveness and performance in processing ISL gestures in real-time scenarios.
- Robustness Testing: Evaluating the system's resilience to variations in lighting conditions, camera angles, and environmental factors.
- End-to-End Testing: Testing the entire ISL translation pipeline from gesture recognition to sentence generation to validate the system's overall functionality and coherence.
- Error Handling: Testing the system's ability to gracefully handle errors, exceptions, and unexpected inputs to maintain stability and reliability.

6.4 Inference drawn from the test cases

Through rigorous testing, we aim to validate the accuracy, reliability, and robustness of the ISL translation system. By identifying and addressing potential issues and shortcomings, we can enhance the system's performance and ensure its effectiveness in real-world scenarios. Testing also provides valuable insights into areas for improvement, guiding future enhancements and optimizations to deliver a more seamless and user-friendly ISL translation experience.

Chapter VII: Results and discussion

7.1 Performance Evaluation measures

Evaluation measures considered:

- Accuracy: This metric measures the proportion of correctly classified instances among all instances.
- Precision (Micro): Micro-precision calculates precision globally by counting the total true positives, false positives, and false negatives. It is particularly useful when dealing with multiclass or multilabel classification problems.
- Precision (Macro): Macro-precision calculates precision for each class independently and then takes the average across all classes.
- Recall (Micro): Micro-recall calculates recall globally by counting the total true positives, false positives, and false negatives. It measures the model's ability to correctly identify all relevant instances.
- Recall (Macro): Macro-recall calculates recall for each class independently and then takes the average across all classes. It provides insight into the model's ability to recall instances from each class.
- Loss: Loss function measures the discrepancy between the predicted values and the actual values in the training data.

	Accuracy	**Precision (micro)**	**Precision (macro)**	**Recall (micro)**	**Recall (macro)**	**Loss**
Values	91.33%	91.33%	90.95%	91.33%	90.71%	0.261

Table 7.1 Performance of main model on test data

7.2 Input Parameters / Features considered

Video dataset

The video dataset was recorded with approximately 50 videos per gesture. Preprocessing on this dataset was done using Mediapipe Pose [5] which gives us 33 coordinates of the human per frame of video. Out of these 33 coordinates, we utilize 25 upper body coordinates, as seen in Fig. 5.3, for classification. Each video in the dataset represents a gloss which is the base output of the model per video.

CNN-LSTM model architecture

- Layers:
  - Conv1D (1-dimensional convolutional layer) with 64 filters and kernel size 3.
  - LSTM (Long Short-Term Memory) with 64 units, returning sequences.
  - LSTM with 64 units.
  - Flatten layer.
  - Dense layer with 128 neurons and ReLU activation.
  - Dense output layer with softmax activation, predicting 26 classes.

- Parameters
  - Optimizer: Adam
  - Activation Function: ReLU
  - Loss Function: Sparse categorical cross entropy
  - Number of Epochs: 10
  - Validation Split: 20%
  - Batch Size: 32

7.3 Graphical and statistical output

Table 7.2 shows a comparison between all trained models and their metrics out of which the custom made CNN-LSTM architecture outperforms all other models. The CNN-LSTM architecture classifies the video based on coordinates.

Model name	*Comparison metrics*											
	Specifications	*train = 0.8, test = 0.2*				*train = 0.75, test = 0.25*				*train = 0.7, test = 0.3*			
		A	P	R	L	A	P	R	L	A	P	R	L
Fine Tuned VideoMAE	learning rate = 0.001, sample rate=4, image resolution=224x224	0.79	0.75	0.74	0.49	0.78	0.74	0.74	0.52	0.75	0.74	0.72	0.52
Support Vector Machine (SVM)	Regularization parameter (C) = 1.0, kernel = 'linear'	0.84	0.84	0.83	-	0.82	0.81	0.81	-	0.80	0.80	0.79	-
Neural network model using 3D CNN	4 hidden NN layers, optimizer='adam', activation='softmax', loss='categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.75	0.74	1.08	0.85	0.73	0.72	1.85	0.80	0.72	0.72	2.01
Neural network model using CNN & LSTM	4 hidden NN layers, optimizer='adam', activation='ReLU', loss='sparse _categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.88	0.87	0.34	0.89	0.89	0.88	0.31	0.86	0.87	0.85	0.39

Table 7.2 Comparison of different models

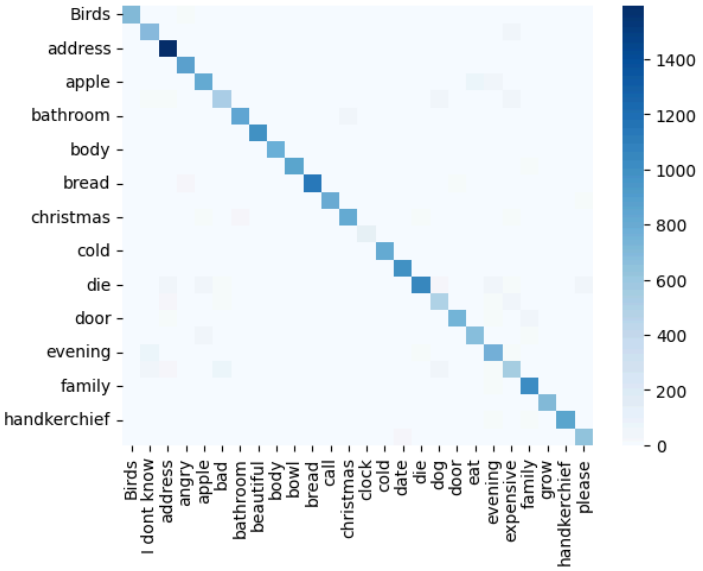


Figure 7.1 Confusion matrix of model

Let's consider a scenario where the system classifies 2 videos into appropriate glosses and builds a sentence in Marathi and Hindi language. The videos are processed frame by frame and the coordinates are extracted. The coordinates are then supplied to the model.

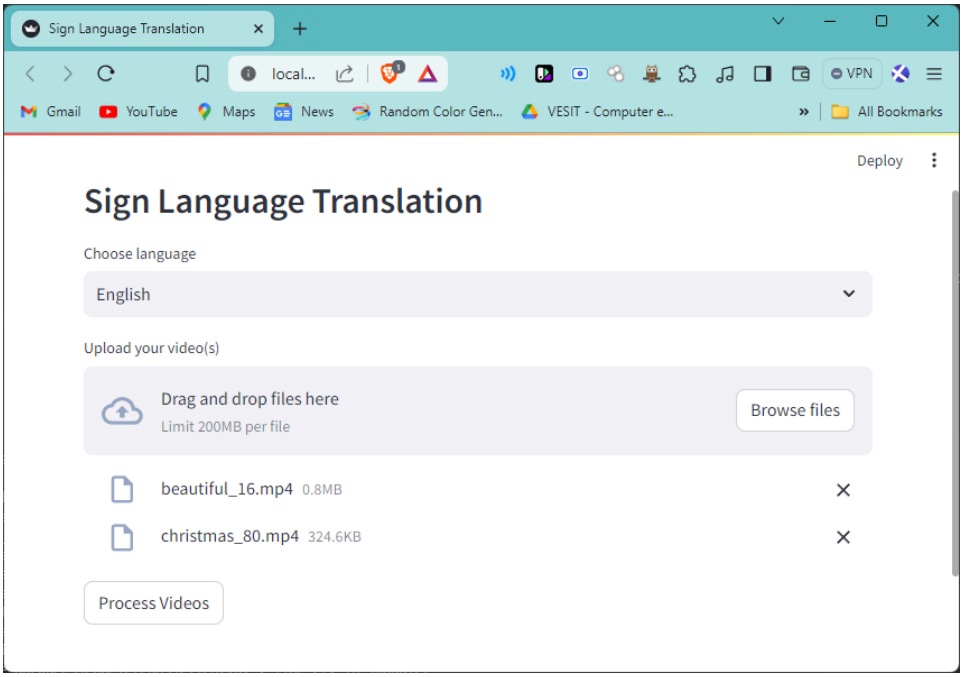


Figure 7.2 Supply videos to model for sentence generation

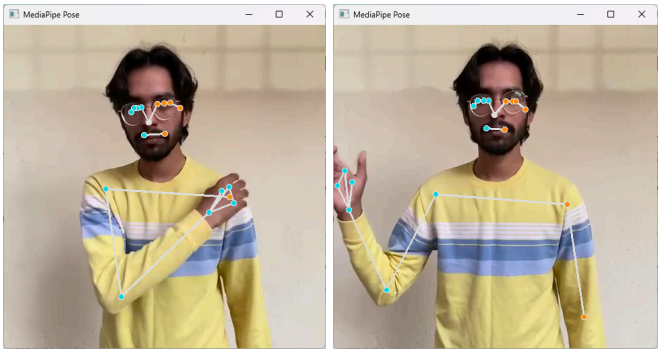


Figure 7.3 Video being processed frame by frame

System's output for given videos:

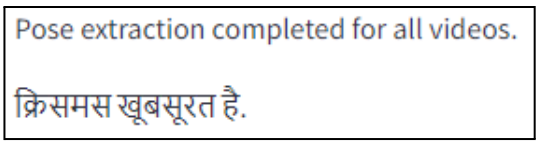


Figure 7.4 Output of system in Hindi

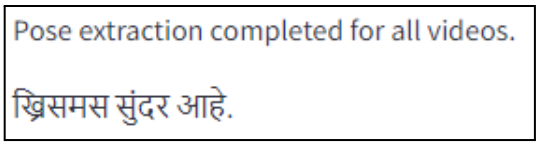


Figure 7.5 Output of system in Marathi

The VideoMAE model also offers promising results but can never accurately predict the gloss of a video with utmost confidence. Consider the following scenario where VideoMAE is given a video of the gloss 'family':

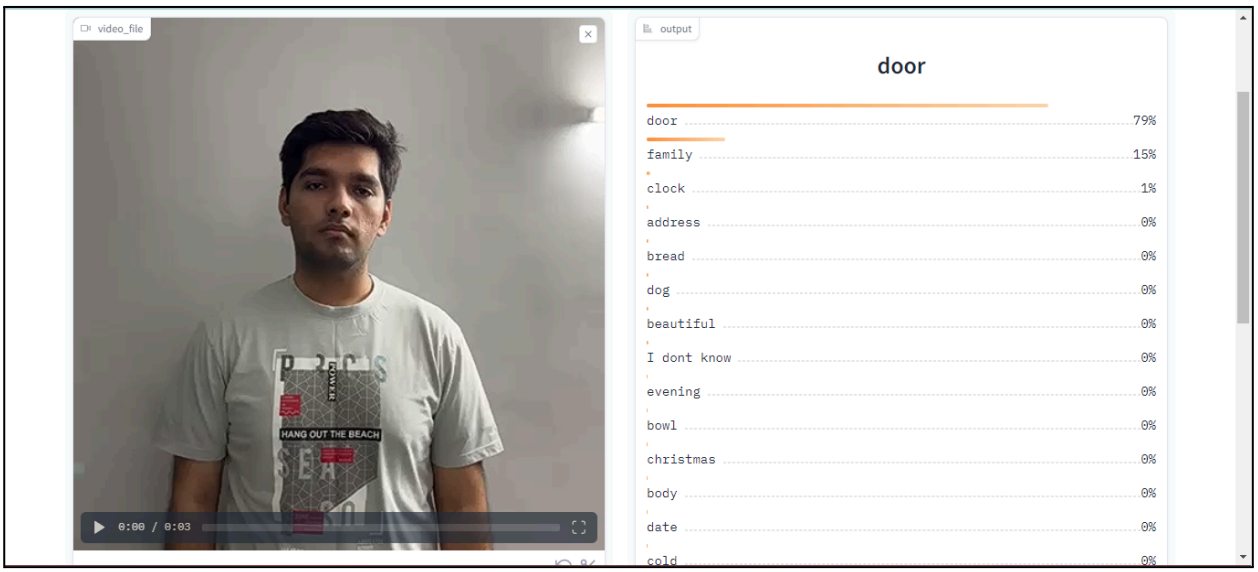


Figure 7.6 Output of VideoMAE model

For the same video, the CNN-LSTM architecture working on Mediapipe Pose coordinates classifies the same video with a confidence score of 87.4% for gloss 'family'.

7.4 Inference drawn

Based on the evaluation results and comparison with existing systems, several key inferences can be drawn regarding the efficacy and applicability of our proposed model. Firstly, our model demonstrates promising performance in accurately recognizing ISL gestures and translating them into text, highlighting its potential to bridge communication gaps between the deaf and hearing communities. Additionally, the accuracy and precision of our model underscore its suitability for real-world applications although many improvements are required. Furthermore, the robustness and versatility of our model, as evidenced by its performance across diverse datasets and scenarios, emphasize its viability for widespread adoption in practical settings. Overall, the positive outcomes from our evaluation reinforce the significance of our proposed model in advancing ISL recognition technology and fostering inclusivity in society.

Chapter VIII: Conclusion

8.1 Limitations

Despite the promising performance of our model, there are several limitations that need to be acknowledged. Firstly, the self created dataset used for training and evaluation may not fully capture the diversity and complexity of real-world ISL gestures. While efforts were made to include a broad range of glosses and variations, the dataset's size and scope may still be limited, potentially affecting the model's generalization ability. Additionally, the reliance on Mediapipe for pose estimation may introduce inaccuracies or inconsistencies in the extracted features, which could impact the model's performance. Furthermore, the computational resources required for training and inference may pose constraints, particularly for real-time applications or deployment on resource-constrained devices. Lastly, the interpretability of the model's predictions may be limited, hindering its usability in certain contexts.

8.2 Conclusion

Our study presents a novel approach to ISL recognition and translation using advanced machine learning techniques. By leveraging a self-created dataset and state-of-the-art models, we developed a robust system capable of accurately recognizing ISL gestures and translating them into text. Through extensive experimentation and evaluation, we demonstrated the effectiveness and reliability of our proposed model, achieving high accuracy and precision metrics. The successful implementation of our model holds significant implications for enhancing accessibility and inclusivity for individuals with hearing impairments, as well as advancing ISL recognition technology.

8.3 Future Scope

Moving forward, there are several avenues for future research and development in the field of ISL recognition and translation. Firstly, expanding and diversifying the dataset to include a wider range of gestures, expressions, and variations would improve the model's robustness and generalization capability. Additionally, exploring alternative pose estimation techniques and feature extraction methods could enhance the accuracy and efficiency of the recognition system. Furthermore, integrating multimodal inputs, such as incorporating facial expressions or hand movements, could enrich the model's understanding of ISL communication. Moreover, investigating real-time deployment options and optimizing the model for resource-constrained environments would facilitate practical applications in assistive technology and communication devices.

References

[1] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[2] Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, and Jyoti Dange. Hand Gesture Identification using Mediapipe. In March 2022 International Research Journal of Engineering and Technology, pages 1199-1202.

[3] Abhinav Joshi, Susmit Agrawal, Ashutosh Modi. 'ISLTranslate: Dataset for Translating Indian Sign Language'. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.05440. arXiv.

[4] Functional and Nonfunctional Requirements: Specification and Types [Online] Available: https://www.altexsoft.com/blog/functional-and-non-functional-requirements-specification-and-types/

[5] Pose landmark detection guide | MediaPipe | Google for Developers [Online] Available: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker

[6] A. Agarwal and M.K. Thakur, "Sign Language Recognition using Microsoft Kinect," Sixth International Conference on Contemporary Computing (IC3), September 2013

Gesturely: a Conversation AI based Indian Sign Language Model

Dr. Nupur Giri
Head of Department, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
nupur.giri@ves.ac.in

Piyush Chugeja
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.piyush.chugeja@ves.ac.in

Sakshi Kirmathe
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.sakshi.kirmathe@ves.ac.in

Deven Bhagtani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.deven.bhagtani@ves.ac.in

***Abstract*—Our project, Gesturely, aims to improve communication within educational settings for individuals with hearing impairments. Sign language, notably Indian Sign Language (ISL) in India, serves as a primary mode of expression for the deaf community. The form of expression among the deaf, relies on a rich vocabulary of gestures involving fingers, hands, arms, eyes, head, and face. Our research endeavors to develop an algorithm capable of translating ISL into English, focusing initially on words within the education domain. Through the integration of advanced computer vision and deep learning methodologies, our objective is to create a system capable of interpreting ISL gestures and converting them into written text. The project involves the creation of a comprehensive dataset, with 50 number of words and over 2500 videos. Our vision is to empower the deaf community with real-time translation capabilities, promoting inclusivity and accessibility in communication.**

***Index Terms*—Indian Sign Language, computer vision, gesture recognition, sign language dataset**

I. Introduction

Communication is vital for human interaction, yet individuals who are mute or deaf face challenges connecting with the hearing community. In India, Indian Sign Language (ISL) is widely used and recognized as the primary mode of communication. It is used by over 5 million deaf people in India. [1] The development of Natural Language Processing (NLP) systems for sign languages such as American Sign Language (ASL) [2], British Sign Language (BSL) [3], and Deutsche Gebärdensprache (DGS) [4] have benefited from the availability of translation datasets. However, there has been relatively limited focus on ISL due to the scarcity of large annotated datasets. This paper aims to address a new translation dataset focused on ISL, with a particular emphasis on the education domain. Additionally, it introduces a deep learning model for classifying gestures.

ISL presents unique challenges due to its limited resources and reliance on bodily gestures for communication, which adds complexity to training machine learning models. Annotating sign language at the gesture level, rather than the sentence level, poses scalability issues. Prior research has explored translating signs into gloss representations and then converting them into written language (Sign2Gloss2Text) [4]. *Glosses* are textual labels assigned to signed gestures, helping translation systems in working at a more detailed level of sign translation. [5] However, generating gloss representations for entire signed sentences presents additional challenges in data annotation. Overall, in this research paper, we make the following contributions:

- A comprehensive ISL to English translation dataset which has more than 50 glosses spread across 2500 and more videos. We believe that making this dataset available for the NLP community will help future research in sign languages.
- A deep learning model for ISL to English translation, inspired by SignAll SDK.[1]

II. Literature review

Unlike spoken languages, sign languages rely on body movements like hand shapes, head nods, eye gazes, and facial expressions to communicate. Translating these continuous movements into written text is quite challenging, opening up new opportunities for research in sign language translation. Many studies have looked into recognizing sign language, using different techniques like gloves, Microsoft Kinect sensors to track hand movements, classify frames based on segmentation masks or utilizing Mediapipe pose estimation pipeline[2].

A. Agarwal and M. K. Thakur [6] use Microsoft Kinect sensors to recognise sign language. They make use of depth images that were captured using the sensor and a gesture is

[1] https://developers.googleblog.com/2021/04/signall-sdk-sign-language-interface-using-mediapipe-now-available.html

[2] https://blog.research.google/2020/12/mediapipe-holistic-simultaneous-face.html

viewed as a sequence of frames. T. Pryor et al. [7] developed SignAloud, which incorporates a pair of gloves equipped with sensors. These gloves track hand position and movement, enabling the conversion of gestures into speech. These hardware solutions are reliable and provide good accuracy but are generally expensive and not portable. Our system eliminates the need of external hardware by using any embedded camera.

In [8], Joshi et al. address the lack of resources for ISL in sign language processing. They present a dataset designed for word-level recognition in ISL from video recordings, with over 4700 words covering diverse topics. To overcome ISL's resource limitations, they use a prototype-based one-shot learner, leveraging ASL resources to improve ISL predictions. Ketan Gomase et al. [9] discuss the development of a sign language recognition system using the Mediapipe framework, which leverages machine learning to detect and interpret hand gestures. The framework identifies 21 3D landmarks on the hand from a single frame, making real-time hand and finger tracking possible. Joshi et al. in [10] introduce the ISLTranslate dataset and proposes a baseline model named Pose-SLT for ISL to English translation, leveraging pose estimation models and transformer architecture.

The dataset we are proposing draws inspiration from CISLR. [8] While CISLR focuses on supporting a one-shot learner model, our dataset diverges from this methodology. We aim to enhance the dataset's utility by including a higher number of videos per gloss, enabling a broader spectrum of research and applications.

III. Proposed Methodology

The primary aim of this study is to develop a framework for translating ISL glosses into English text, thereby improving accessibility and inclusivity for individuals with hearing impairments. To achieve this objective, we propose a multi-step approach that involved data collection, preprocessing, feature extraction, model development, and evaluation. By building upon recent advancements in machine learning and computer vision, our methodology aims at overcoming the challenges associated with ISL translation, including the lack of annotated datasets and the complexity of sign language recognition. In our approach, we follow the methodology laid out in [6], which involves carefully examining videos frame by frame.

A. Data Collection

Our dataset is created from the videos provided by CISLR [8], which has 57 distinct categories. Within our dataset, we emphasize the education domain, comprising more than 50 glosses distributed across 2500 videos. All videos maintain a consistent format, adhering to a 1:1 aspect ratio and recorded with a 720p resolution at 30 frames per second. This standardized recording setup ensures uniformity and quality across the dataset. Each gloss is represented by multiple videos, with diverse angles and lighting conditions to improve the training data and enhance model robustness.

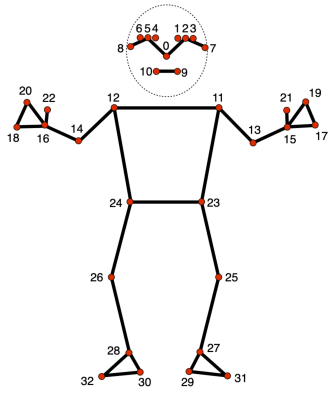


Fig. 1. Mediapipe Pose Landmarks [13]

B. Pre-processing and Feature extraction

Each video undergoes frame-by-frame processing, where pose coordinates are extracted from each frame as shown in Fig. 2. Mediapipe provides us with 33 coordinates detailing the human body from head to toe. However, as depicted in Fig. 2, body parts below the waist remain unseen, making coordinates below the waistline unusable. This discrepancy in data could introduce inconsistencies. To rectify this issue, all unused coordinates per frame, specifically coordinates 25 to 32 (8 coordinates), are filtered out before feeding the data to the model.

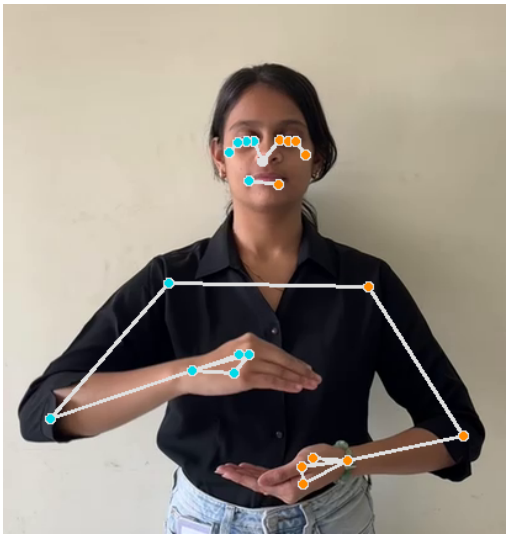


Fig. 2. Extracting coordinates from a frame of video for gloss *'grow'*

C. Model Development

After the coordinates are pre-processed and filtered, classification models will be trained. Following the approach outlined in [6], our focus is on training models capable of accurately distinguishing ISL glosses from the extracted pose data. We incorporate Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks to capture the complex patterns present in sign language gestures. These

models are trained using optimization algorithms like Adam or SGD, with carefully selected hyperparameters to ensure effective learning. Throughout the model development process, we conduct thorough experimentation and validation to ensure the reliability and applicability of our approach.

D. Evaluation

In evaluation, we assess the performance of the developed model based on various parameters, including accuracy, precision, recall and test loss. We evaluate how well the model works when the data isn't seen or is retrieved real-time. We also test if the model is unbiased and doesn't favor a particular class.

IV. Proposed System

The system aims to address the challenge of translating ISL. It works by collecting the user input, processing the frames and classifying them into respective glosses. Once the video is processed and glosses are identified, the sentence is constructed. The proposed system can be seen in Fig. 3, based on the methodology followed in [11] where they leverage *gpt-3.5-turbo* for sentence generation from glosses.

A. Model selection and evaluation

After evaluating various model architectures and their applications, we opted to adopt the model type outlined in [12], which leverages CNN and LSTM. Table I outlines the various classification techniques used, along with their specifications and corresponding test metrics. The Conv1D model was selected as the final model, leveraging classification based on Mediapipe Pose coordinates. Notably, this model achieved the highest accuracy and the lowest test loss score among all the models considered. For the model, labelled as Sign Language Classifier in Fig. 3, we're employing a sequential architecture comprising Conv1D, LSTM, Flatten, and Dense layers. This architecture processes each frame's 25 coordinates provided by Mediapipe, enabling the model to classify each video into specified glosses. The model learns patterns from the sequential data to accurately assign gloss labels to the videos during training.

B. Gesture Classification

User input is obtained either through computer vision or recorded video, wherein the user performs a series of ISL gestures to communicate. The collected input undergoes preprocessing, wherein each frame is analyzed, and the extracted landmarks are forwarded to the classifier model, depicted in Fig. 3. Landmarks generated per frame, acquired using Mediapipe, are refined to include only upper body landmarks (landmarks 0 to 24, as illustrated in Fig. 1). These refined landmarks are then sequentially sent to the model, which classifies the data into glosses.

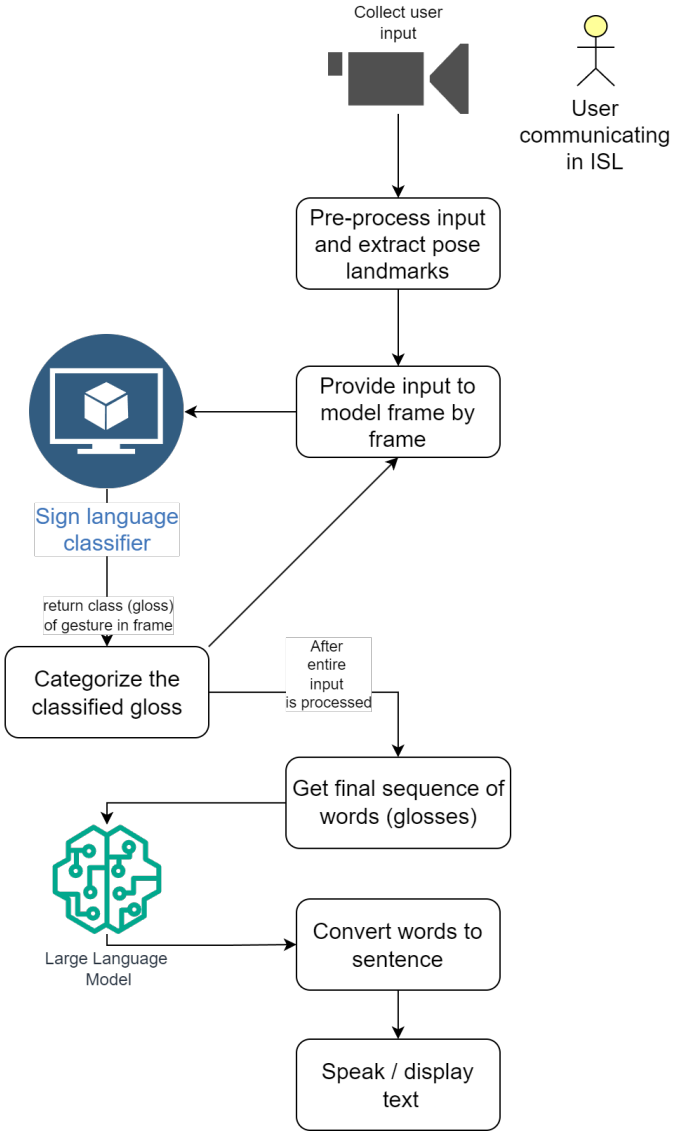


Fig. 3. Block diagram of proposed system

C. Sentence Construction

Following the identification of a sequence of glosses, the list of words is transmitted to a Large Language Model (LLM). Shahin and Ismail [15] explored ChatGPT's ability to convert glosses into sign language. Similarly, we incorporate Google's Gemini[3] to build sentences. The LLM generates a grammatically correct and contextually appropriate sentence that conveys the user's message. Following is the prompt given to Gemini: *You are a sign language translator. You are being given a set of words collected from people using sign language, the words are what sentence they want to speak. Generate a grammatically correct and meaningful sentence using those words. Do not add context to the sentence. Return only the sentence and nothing more.*

[3] https://deepmind.google/technologies/gemini/

TABLE I
CLASSIFICATION MODELS SPECIFICATIONS
A - ACCURACY, P - PRECISION, R - RECALL, L - TEST LOSS

Model name	*Specifications*	*Comparison metrics*											
		train = 0.8, test = 0.2				*train = 0.75, test = 0.25*				*train = 0.7, test = 0.3*			
		A	**P**	**R**	**L**	**A**	**P**	**R**	**L**	**A**	**P**	**R**	**L**
Fine Tuned VideoMAE	learning rate = 0.001, sample rate=4, image resolution=224x224	0.79	0.75	0.74	0.49	0.78	0.74	0.74	0.52	0.75	0.74	0.72	0.52
Support Vector Machine (SVM)	Regularization parameter (C) = 1.0, kernel = 'linear'	0.84	0.84	0.83	-	0.82	0.81	0.81	-	0.80	0.80	0.79	-
Neural network model using 3D CNN	4 hidden NN layers, optimizer='adam', activation='softmax', loss='categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.75	0.74	1.08	0.85	0.73	0.72	1.85	0.80	0.72	0.72	2.01
Neural network model using CNN & LSTM	4 hidden NN layers, optimizer='adam', activation='ReLU', loss='sparse _categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.88	0.87	0.34	0.89	0.89	0.88	0.31	0.86	0.87	0.85	0.39

D. Outputs of system

Based on the methodology and flow of system explained in the preceding sections, let us have a look at how the system functions. A basic user interface (UI) shown in Fig. 4 has been developed for users to upload the videos they wish to translate into sentences. As seen in Fig. 4, we upload 2 videos for the glosses *beautiful* and *Christmas*. These videos are then processed using Mediapipe Pose and their coordinates are extracted, as seen in Fig. 5. Once the videos are processed, coordinates are sent to the model and it classifies them into *glosses*. This collection of glosses is sent to an LLM for sentence construction, which is shown to user and converted to speech using Google Text to Speech (gTTS) engine.

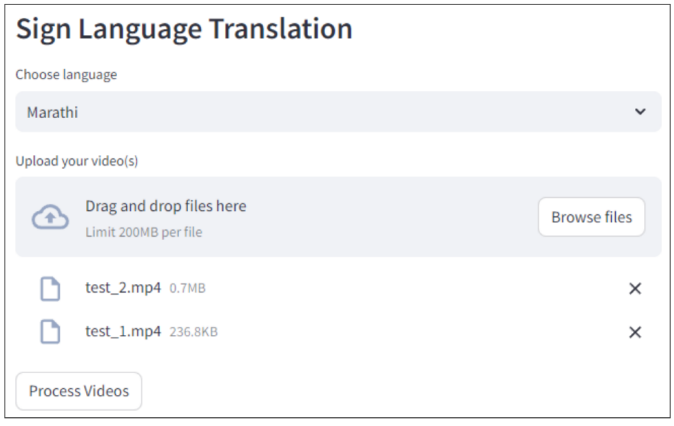


Fig. 4. Uploading 2 videos

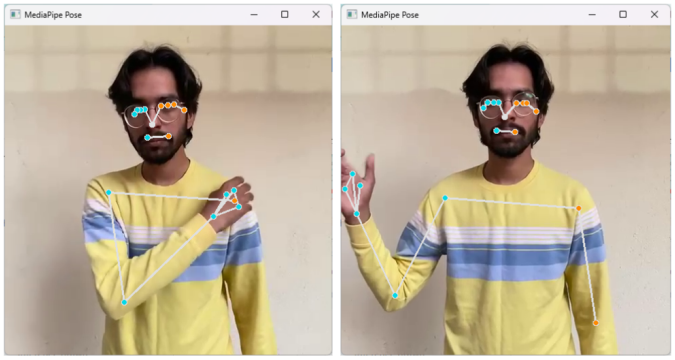


Fig. 5. Processing uploaded videos

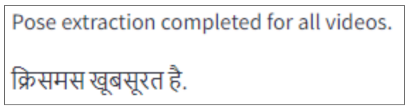


Fig. 6. System outputs sentence in Hindi

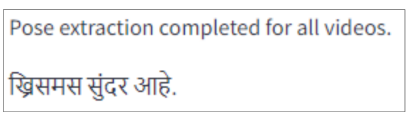


Fig. 7. System outputs sentence in Marathi

V. Conclusion

We present a novel approach to Indian Sign Language (ISL) translation using advanced machine learning techniques. By using Mediapipe for pose estimation and a sequential model architecture comprising Conv1D and LSTM layers, we have developed a system capable of classifying ISL gestures into specified glosses with a 88% accuracy. Our approach, inspired by previous studies, shows promise in connecting the deaf and hearing communities better. Looking ahead, our work paves the way for more improvements in ISL translation tech, which could make life easier and more inclusive for people with hearing challenges.

References

[1] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 1366-1375. https://doi.org/10.1145/3394171.3413528

[2] Li, Dongxu & Rodríguez, Cristian & Yu, Xin & Li, Hongdong. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. 1448-1458. 10.1109/WACV45572.2020.9093512.

[3] Samuel Albanie, GülVarol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In 2018 IEEE/CVFConference on Computer Vision and Pattern Recognition, pages 7784–7793.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, 'Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation', arXiv [cs.CV]. 2020.

[6] A. Agarwal and M.K. Thakur, "Sign Language Recognition using Microsoft Kinect," Sixth International Conference on Contemporary Computing (IC3), September 2013.

[7] MailOnline, "SignAloud gloves translate sign language gestures into spoken English," 2016. [Online]. Available: http://www.dailymail.co.uk/sciencetech/article-3557362/SignAloudgloves-translate-sign-language-movements-spoken-English.html. [Accessed: 06-03-2024].

[8] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[9] Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, and Jyoti Dange. Hand Gesture Identification using Mediapipe. In March 2022 International Research Journal of Engineering and Technology, pages 1199-1202.

[10] Abhinav Joshi, Susmit Agrawal, Ashutosh Modi. 'ISLTranslate: Dataset for Translating Indian Sign Language'. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.05440. arXiv.

[11] O. M. Sincan, N. C. Camgoz, and R. Bowden, 'Using an LLM to Turn Sign Spottings into Spoken Language Sentences', arXiv [cs.CV]. 2024. https://arxiv.org/pdf/2403.10434.pdf

[12] K. B. Tran, U. D. Nguyen and Q. T. Huynh, "Continuous Sign Language Recognition Using MediaPipe," 2023 International Conference on Advanced Technologies for Communications (ATC), Da Nang, Vietnam, 2023, pp. 493-498, doi: 10.1109/ATC58710.2023.10318855.

[13] Pose landmark detection guide — MediaPipe — Google for Developers [Online]. Available: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker [Accessed: 07-03-2024]

[14] Z. Tong, Y. Song, J. Wang, and L. Wang, 'VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training', arXiv [cs.CV]. 2022.

[15] N. Shahin and L. Ismail, 'ChatGPT, Let us Chat Sign Language: Experiments, Architectural Elements, Challenges and Research Directions', arXiv [cs.CL]. 2024.

Gesture

by Sakshi Kamathe

Submission date: 12-Apr-2024 03:20PM (UTC+0530)
Submission ID: 2347441188
File name: Gesturely.pdf (1.36M)
Word count: 2690
Character count: 15389

Gesturely: a Conversation AI based Indian Sign Language Model

Dr. Nupur Giri
Head of Department, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
nupur.giri@ves.ac.in

Piyush Chugeja
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.piyush.chugeja@ves.ac.in

Sakshi Kirmathe
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.sakshi.kirmathe@ves.ac.in

Deven Bhagtani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.deven.bhagtani@ves.ac.in

***Abstract*—Our project, Gesturely, aims to improve communication within educational settings for individuals with hearing impairments. Sign language, notably Indian Sign Language (ISL) in India, serves as a primary mode of expression for the deaf community. The form of expression among the deaf, relies on a rich vocabulary of gestures involving fingers, hands, arms, eyes, head, and face. Our research endeavors to develop an algorithm capable of translating ISL into English, focusing initially on words within the education domain. Through the integration of advanced computer vision and deep learning methodologies, our objective is to create a system capable of interpreting ISL gestures and converting them into written text. The project involves the creation of a comprehensive dataset, with 50 number of words and over 2500 videos. Our vision is to empower the deaf community with real-time translation capabilities, promoting inclusivity and accessibility in communication.**

***Index Terms*—Indian Sign Language, computer vision, gesture recognition, sign language dataset**

I. Introduction

Communication is vital for human interaction, yet individuals who are mute or deaf face challenges connecting with the hearing community. In India, Indian Sign Language (ISL) is widely used and recognized as the primary mode of communication. It is used by over 5 million deaf people in India. [1] The development of Natural Language Processing (NLP) systems for sign languages such as American Sign Language (ASL) [2], British Sign Language (BSL) [3], and Deutsche Gebärdensprache (DGS) [4] have benefited from the availability of translation datasets. However, there has been relatively limited focus on ISL due to the scarcity of large annotated datasets. This paper aims to address a new translation dataset focused on ISL, with a particular emphasis on the education domain. Additionally, it introduces a deep learning model for classifying gestures.

ISL presents unique challenges due to its limited resources and reliance on bodily gestures for communication, which adds complexity to training machine learning models. Annotating sign language at the gesture level, rather than the sentence level, poses scalability issues. Prior research has explored translating signs into gloss representations and then converting them into written language (Sign2Gloss2Text) [4]. *Glosses* are textual labels assigned to signed gestures, helping translation systems in working at a more detailed level of sign translation. [5] However, generating gloss representations for entire signed sentences presents additional challenges in data annotation. Overall, in this research paper, we make the following contributions:

- A comprehensive ISL to English translation dataset which has more than 50 glosses spread across 2500 and more videos. We believe that making this dataset available for the NLP community will help future research in sign languages.
- A deep learning model for ISL to English translation, inspired by SignAll SDK.[1]

II. Literature Review

Unlike spoken languages, sign languages rely on body movements like hand shapes, head nods, eye gazes, and facial expressions to communicate. Translating these continuous movements into written text is quite challenging, opening up new opportunities for research in sign language translation. Many studies have looked into recognizing sign language, using different techniques like gloves, Microsoft Kinect sensors to track hand movements, classify frames based on segmentation masks or utilizing Mediapipe pose estimation pipeline[2].

A. Agarwal and M. K. Thakur [6] use Microsoft Kinect sensors to recognise sign language. They make use of depth images that were captured using the sensor and a gesture is

[1]https://developers.googleblog.com/2021/04/signall-sdk-sign-language-interface-using-mediapipe-now-available.html
[2]https://blog.research.google/2020/12/mediapipe-holistic-simultaneous-face.html

viewed as a sequence of frames. T. Pryor et al. [7] developed SignAloud, which incorporates a pair of gloves equipped with sensors. These gloves track hand position and movement, enabling the conversion of gestures into speech. These hardware solutions are reliable and provide good accuracy but are generally expensive and not portable. Our system eliminates the need of external hardware by using any embedded camera.

In [8], Joshi et al. address the lack of resources for ISL in sign language processing. They present a dataset designed for word-level recognition in ISL from video recordings, with over 4700 words covering diverse topics. To overcome ISL's resource limitations, they use a prototype-based one-shot learner, leveraging ASL resources to improve ISL predictions. Ketan Gomase et al. [9] discuss the development of a sign language recognition system using the Mediapipe framework, which leverages machine learning to detect and interpret hand gestures. The framework identifies 21 3D landmarks on the hand from a single frame, making real-time hand and finger tracking possible. Joshi et al. in [10] introduce the ISLTranslate dataset and proposes a baseline model named Pose-SLT for ISL to English translation, leveraging pose estimation models and transformer architecture.

The dataset we are proposing draws inspiration from CISLR. [8] While CISLR focuses on supporting a one-shot learner model, our dataset diverges from this methodology. We aim to enhance the dataset's utility by including a higher number of videos per gloss, enabling a broader spectrum of research and applications.

III. Proposed Methodology

The primary aim of this study is to develop a framework for translating ISL glosses into English text, thereby improving accessibility and inclusivity for individuals with hearing impairments. To achieve this objective, we propose a multi-step approach that involved data collection, preprocessing, feature extraction, model development, and evaluation. By building upon recent advancements in machine learning and computer vision, our methodology aims at overcoming the challenges associated with ISL translation, including the lack of annotated datasets and the complexity of sign language recognition. In our approach, we follow the methodology laid out in [6], which involves carefully examining videos frame by frame.

A. Data Collection

Our dataset is created from the videos provided by CISLR [8], which has 57 distinct categories. Within our dataset, we emphasize the education domain, comprising more than 50 glosses distributed across 2500 videos. All videos maintain a consistent format, adhering to a 1:1 aspect ratio and recorded with a 720p resolution at 30 frames per second. This standardized recording setup ensures uniformity and quality across the dataset. Each gloss is represented by multiple videos, with diverse angles and lighting conditions to improve the training data and enhance model robustness.

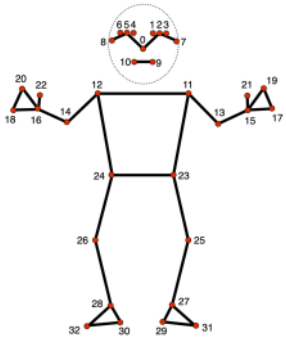


Fig. 1. Mediapipe Pose Landmarks [13]

B. Pre-processing and Feature extraction

Each video undergoes frame-by-frame processing, where pose coordinates are extracted from each frame as shown in Fig. 2. Mediapipe provides us with 33 coordinates detailing the human body from head to toe. However, as depicted in Fig. 2, body parts below the waist remain unseen, making coordinates below the waistline unusable. This discrepancy in data could introduce inconsistencies. To rectify this issue, all unused coordinates per frame, specifically coordinates 25 to 32 (8 coordinates), are filtered out before feeding the data to the model.

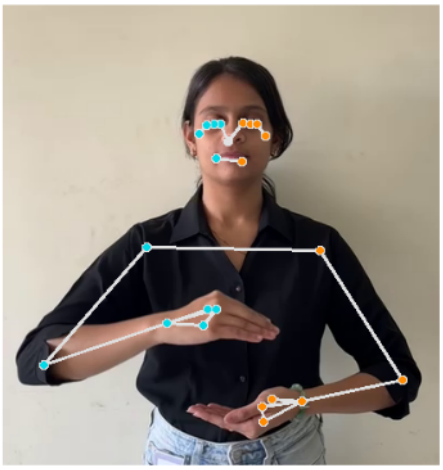


Fig. 2. Extracting coordinates from a frame of video for gloss *'grow'*

C. Model Development

After the coordinates are pre-processed and filtered, classification models will be trained. Following the approach outlined in [6], our focus is on training models capable of accurately distinguishing ISL glosses from the extracted pose data. We incorporate Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks to capture the complex patterns present in sign language gestures. These

models are trained using optimization algorithms like Adam or SGD, with carefully selected hyperparameters to ensure effective learning. Throughout the model development process, we conduct thorough experimentation and validation to ensure the reliability and applicability of our approach.

D. Evaluation

In evaluation, we assess the performance of the developed model based on various parameters, including accuracy, precision, recall and test loss. We evaluate how well the model works when the data isn't seen or is retrieved real-time. We also test if the model is unbiased and doesn't favor a particular class.

IV. Proposed System

The system aims to address the challenge of translating ISL. It works by collecting the user input, processing the frames and classifying them into respective glosses. Once the video is processed and glosses are identified, the sentence is constructed. The proposed system can be seen in Fig. 3, based on the methodology followed in [11] where they leverage *gpt-3.5-turbo* for sentence generation from glosses.

A. Model selection and evaluation

After evaluating various model architectures and their applications, we opted to adopt the model type outlined in [12], which leverages CNN and LSTM. Table I outlines the various classification techniques used, along with their specifications and corresponding test metrics. The Conv1D model was selected as the final model, leveraging classification based on Mediapipe Pose coordinates. Notably, this model achieved the highest accuracy and the lowest test loss score among all the models considered. For the model, labelled as Sign Language Classifier in Fig. 3, we're employing a sequential architecture comprising Conv1D, LSTM, Flatten, and Dense layers. This architecture processes each frame's 25 coordinates provided by Mediapipe, enabling the model to classify each video into specified glosses. The model learns patterns from the sequential data to accurately assign gloss labels to the videos during training.

B. Gesture Classification

User input is obtained either through computer vision or recorded video, wherein the user performs a series of ISL gestures to communicate. The collected input undergoes preprocessing, wherein each frame is analyzed, and the extracted landmarks are forwarded to the classifier model, depicted in Fig. 3. Landmarks generated per frame, acquired using Mediapipe, are refined to include only upper body landmarks (landmarks 0 to 24, as illustrated in Fig. 1). These refined landmarks are then sequentially sent to the model, which classifies the data into glosses.

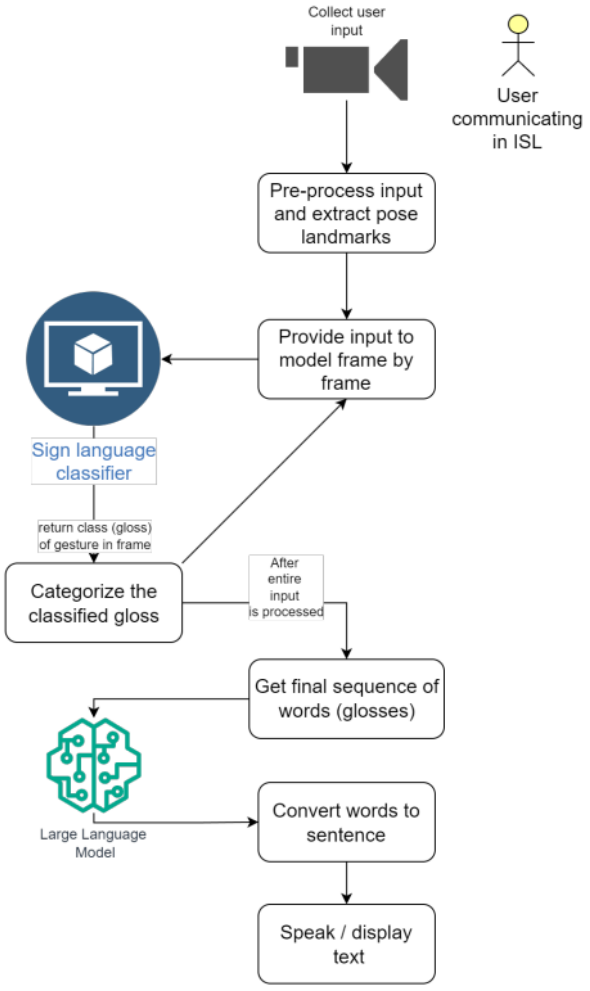


Fig. 3. Block diagram of proposed system

C. Sentence Construction

Following the identification of a sequence of glosses, the list of words is transmitted to a Large Language Model (LLM). Shahin and Ismail [15] explored ChatGPT's ability to convert glosses into sign language. Similarly, we incorporate Google's Gemini[3] to build sentences. The LLM generates a grammatically correct and contextually appropriate sentence that conveys the user's message. Following is the prompt given to Gemini: *You are a sign language translator. You are being given a set of words collected from people using sign language, the words are what sentence they want to speak. Generate a grammatically correct and meaningful sentence using those words. Do not add context to the sentence. Return only the sentence and nothing more.*

[3]https://deepmind.google/technologies/gemini/

TABLE I
CLASSIFICATION MODELS SPECIFICATIONS
A - ACCURACY, P - PRECISION, R - RECALL, L - TEST LOSS

Model name	*Specifications*	*Comparison metrics*											
		train = 0.8, test = 0.2				*train = 0.75, test = 0.25*				*train = 0.7, test = 0.3*			
		A	**P**	**R**	**L**	**A**	**P**	**R**	**L**	**A**	**P**	**R**	**L**
Fine Tuned VideoMAE	learning rate = 0.001, sample rate=4, image resolution=224x224	0.79	0.75	0.74	0.49	0.78	0.74	0.74	0.52	0.75	0.74	0.72	0.52
Support Vector Machine (SVM)	Regularization parameter (C) = 1.0, kernel = 'linear'	0.84	0.84	0.83	-	0.82	0.81	0.81	-	0.80	0.80	0.79	-
Neural network model using 3D CNN	4 hidden NN layers, optimizer='adam', activation='softmax', loss='categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.75	0.74	1.08	0.85	0.73	0.72	1.85	0.80	0.72	0.72	2.01
Neural network model using CNN & LSTM	4 hidden NN layers, optimizer='adam', activation='ReLU', loss='sparse categorical _crossentropy', epochs=10, validation_split=0.2, batch_size=32	0.88	0.88	0.87	0.34	0.89	0.89	0.88	0.31	0.86	0.87	0.85	0.39

D. Outputs of system

Based on the methodology and flow of system explained in the preceding sections, let us have a look at how the system functions. A basic user interface (UI) shown in Fig. 4 has been developed for users to upload the videos they wish to translate into sentences. As seen in Fig. 4, we upload 2 videos for the glosses *beautiful* and *Christmas*. These videos are then processed using Mediapipe Pose and their coordinates are extracted, as seen in Fig. 5. Once the videos are processed, coordinates are sent to the model and it classifies them into *glosses*. This collection of glosses is sent to an LLM for sentence construction, which is shown to user and converted to speech using Google Text to Speech (gTTS) engine.

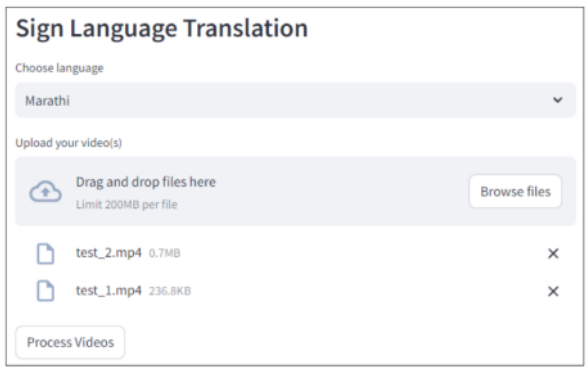


Fig. 4. Uploading 2 videos

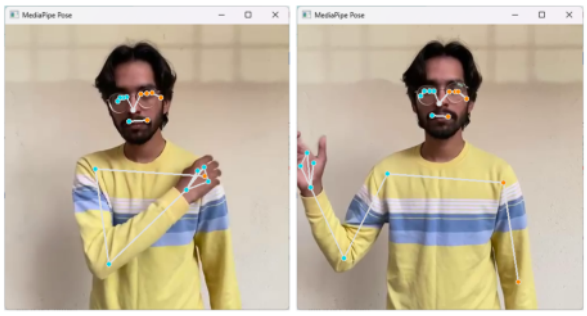


Fig. 5. Processing uploaded videos

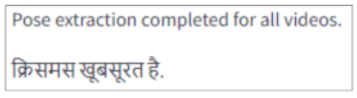


Fig. 6. System outputs sentence in Hindi

Fig. 7. System outputs sentence in Marathi

V. Conclusion

We present a novel approach to Indian Sign Language (ISL) translation using advanced machine learning techniques. By using Mediapipe for pose estimation and a sequential model architecture comprising Conv1D and LSTM layers, we have developed a system capable of classifying ISL gestures into specified glosses with a 88% accuracy. Our approach, inspired by previous studies, shows promise in connecting the deaf and hearing communities better. Looking ahead, our work paves the way for more improvements in ISL translation tech, which could make life easier and more inclusive for people with hearing challenges.

References

[1] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 1366-1375. https://doi.org/10.1145/3394171.3413528

[2] Li, Dongxu & Rodríguez, Cristian & Yu, Xin & Li, Hongdong. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. 1448-1458. 10.1109/WACV45572.2020.9093512.

[3] Samuel Albanie, GülVarol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In 2018 IEEE/CVFConference on Computer Vision and Pattern Recognition, pages 7784–7793.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, 'Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation', arXiv [cs.CV]. 2020.

[6] A. Agarwal and M.K. Thakur, "Sign Language Recognition using Microsoft Kinect," Sixth International Conference on Contemporary Computing (IC3), September 2013.

[7] MailOnline, "SignAloud gloves translate sign language gestures into spoken English," 2016. [Online]. Available: http://www.dailymail.co.uk/sciencetech/article-3557362/SignAloudgloves-translate-sign-language-movements-spoken-English.html. [Accessed: 06-03-2024].

[8] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[9] Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, and Jyoti Dange. Hand Gesture Identification using Mediapipe. In March 2022 International Research Journal of Engineering and Technology, pages 1199-1202.

[10] Abhinav Joshi, Susmit Agrawal, Ashutosh Modi. 'ISLTranslate: Dataset for Translating Indian Sign Language'. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.05440. arXiv.

[11] O. M. Sincan, N. C. Camgoz, and R. Bowden, 'Using an LLM to Turn Sign Spottings into Spoken Language Sentences', arXiv [cs.CV]. 2024. https://arxiv.org/pdf/2403.10434.pdf

[12] K. B. Tran, U. D. Nguyen and Q. T. Huynh, "Continuous Sign Language Recognition Using MediaPipe," 2023 International Conference on Advanced Technologies for Communications (ATC), Da Nang, Vietnam, 2023, pp. 493-498, doi: 10.1109/ATC58710.2023.10318855.

[13] Pose landmark detection guide — MediaPipe — Google for Developers [Online]. Available: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker [Accessed: 07-03-2024]

[14] Z. Tong, Y. Song, J. Wang, and L. Wang, 'VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training', arXiv [cs.CV]. 2022.

[15] N. Shahin and L. Ismail, 'ChatGPT, Let us Chat Sign Language: Experiments, Architectural Elements, Challenges and Research Directions', arXiv [cs.CL]. 2024.

Gesture

ORIGINALITY REPORT

6%	4%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

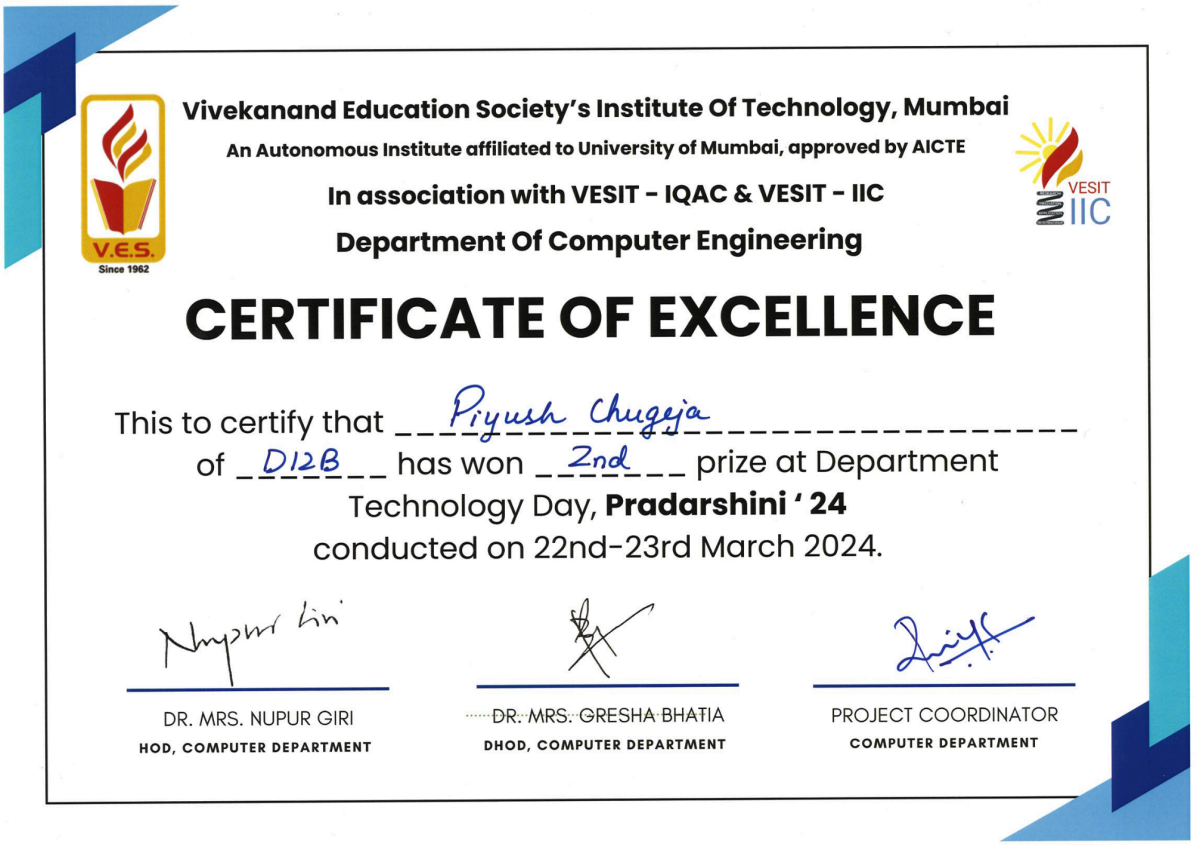
PRIMARY SOURCES

1	Nupur Giri, Rahul Jaisinghani, Rohit Kriplani, Tarun Ramrakhyani, Vinay Bhatia. "Distributed Denial Of Service(DDoS) Mitigation in Software Defined Network using Blockchain", 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019 Publication	1%
2	arxiv.org Internet Source	1%
3	Aakash Deep, Aashutosh Litoriya, Akshay Ingole, Vaibhav Asare, Shubham M Bhole, Shantanu Pathak. "Realtime Sign Language Detection and Recognition", 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 2022 Publication	1%
4	eudl.eu Internet Source	1%
5	github.com Internet Source	1%

6	Md. Mahadi Hasan Sany, Mumenunnesa Keya, Sharun Akter Khushbu, Akm Shahariar Azad Rabby, Abu Kaisar Mohammad Masum. "Chapter 1 An Opinion Mining of Text in COVID-19 Issues Along with Comparative Study in ML, BERT & RNN", Springer Science and Business Media LLC, 2022 Publication	1 %
7	fastercapital.com Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

2. Competition certificate



Vivekanand Education Society's Institute Of Technology, Mumbai

An Autonomous Institute affiliated to University of Mumbai, approved by AICTE

In association with VESIT - IQAC & VESIT - IIC

Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Piyush Chugeja
of D12B has won 2nd prize at Department Technology Day, **Pradarshini '24** conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI
HOD, COMPUTER DEPARTMENT

DR. MRS. GRESHA BHATIA
DHOD, COMPUTER DEPARTMENT

PROJECT COORDINATOR
COMPUTER DEPARTMENT

Vivekanand Education Society's Institute Of Technology, Mumbai

An Autonomous Institute affiliated to University of Mumbai, approved by AICTE

In association with VESIT - IQAC & VESIT - IIC

Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Sakshi Kirmathe
of D12B has won 2nd prize at Department Technology Day, **Pradarshini '24** conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI
HOD, COMPUTER DEPARTMENT

DR. MRS. GRESHA BHATIA
DHOD, COMPUTER DEPARTMENT

PROJECT COORDINATOR
COMPUTER DEPARTMENT

Vivekanand Education Society's Institute Of Technology, Mumbai

An Autonomous Institute affiliated to University of Mumbai, approved by AICTE

In association with VESIT - IQAC & VESIT - IIC

Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Deven Bhagtani of D12B has won 2nd prize at Department Technology Day, **Pradarshini '24** conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI
HOD, COMPUTER DEPARTMENT

DR. MRS. GRESHA BHATIA
DHOD, COMPUTER DEPARTMENT

PROJECT COORDINATOR
COMPUTER DEPARTMENT

3. Project Review Sheet

Industry / Inhouse:
Research / Innovation:

Project Evaluation Sheet 2023-24

Class: D12 B

Title of Project (Group no): Group 2: Gesturely - Sign language to words

Group Members: 11 - Piyush Chugeja, 25 - Sakshi, 6 - Deven.

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation on Skills (3)	Applied Engg & Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	5	5	4	3	5	2	2	2	2	3	3	3	5	1	45

Comments: Classification of dataset might be required to demonstrate sentences

Sanjay Mirchandani
Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation on Skills (3)	Applied Engg & Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	5	5	4	3	5	2	2	2	2	3	3	3	5	1	45

Comments: Need to work on lower body landmark coordinates.
ISL - translation requires to be augmented

Date: 10th February, 2024

Dr. Nupur Giri
Name & Signature Reviewer2

Inhouse/ Industry
Innovation/Resear

Project Evaluation Sheet 2023 - 24

Class: D12 A/B/C
Group No.: 2

Title of Project: Gesterrely - Gestures to Words

Group Members: Deven Bhagtani (6), Piyush Chugeja (11), Sakshi Kiranathe (25)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation on Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	5	2	2	2	2	2	3	3	3	3	3	44

Comments: For comparison of dataset split it into 75:25, 80:20, 70:30

Sanjay M.
Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation on Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	5	2	2	2	2	2	3	3	3	3	3	44

Comments: Keep specifications for similar models.

Date: 9th March, 2024

Dr. Nupur Giri
Name & Signature Reviewer 2