**SIGN LANGUAGE RECOGNITION MODEL BASED ON MULTIMODAL INPUTS**

**Submitted**

**By:**

1. **DABBARA PRASHANTH [BU21EECE0100458]**

**Under the Guidance of:**

**Dr. CHINTOO KUMAR**

**Dr. AVISHEK CHAKRABORTHY**

**(Duration: Date/Month/Year to Date/Month/Year)**



**Department of Electrical, Electronics and Communication Engineering**

**GITAM School of Technology**

**GITAM**

**(DEEMED TO BE UNIVERSITY)**

**(Estd. u/s 3 of the UGC Act 1956)**

**NH 207, Nagadenehalli, Doddaballapur taluk, Bengaluru-561203 Karnataka, INDIA.**

**DECLARATION**

**We declare that the project work contained in this report is original and it has been done by us under the guidance of my project guide.**

**Name: DABBARA PRASHANTH**

**Date: Signature of the Student**

****

**Department of Electrical, Electronics, and Communication Engineering**

**GITAM School of Technology, Bengaluru-561203**

**CERTIFICATE**

**This is to certify that DABBARA PRASHANTH bearing (Regd. No.: BU21EECE0100458) has satisfactorily completed Mini Project Entitled in partial fulfillment of the requirements as prescribed by University for VIIth semester, Bachelor of Technology in “Electrical, Electronics and Communication Engineering” and submitted this report during the academic year 2024-2025.**

**Signature of the Guide Signature of HOD**

**Table of contents:**

[**Chapter 1: Introduction 1**](#_heading=h.gjdgxs)

[1.1 Overview of the problem statement 1](#_heading=h.30j0zll)

[1.2 Objectives and goals 1](#_heading=h.1fob9te)

[**Chapter 2 : Literature Review 2**](#_heading=h.3znysh7)

[**Chapter 3 : Strategic Analysis and Problem Definition 3**](#_heading=h.2et92p0)

[3.1 SWOT Analysis 3](#_heading=h.tyjcwt)

[3.2 Project Plan - GANTT Chart 3](#_heading=h.1t3h5sf)

[3.3 Refinement of problem statement 3](#_heading=h.2s8eyo1)

[**Chapter 4 : Methodology 4**](#_heading=h.17dp8vu)

[4.1 Description of the approach 4](#_heading=h.3rdcrjn)

[4.2 Tools and techniques utilized 4](#_heading=h.26in1rg)

[4.3 Design considerations 4](#_heading=h.lnxbz9)

[**Chapter 5 : Implementation 5**](#_heading=h.1ksv4uv)

[5.1 Description of how the project was executed 5](#_heading=h.44sinio)

[5.2 Challenges faced and solutions implemented 5](#_heading=h.2jxsxqh)

[**Chapter 6:Results 6**](#_heading=h.z337ya)

[6.1 outcomes 6](#_heading=h.3j2qqm3)

[6.2 Interpretation of results 6](#_heading=h.1y810tw)

[6.3 Comparison with existing literature or technologies 6](#_heading=h.2xcytpi)

[**Chapter 7: Conclusion 7**](#_heading=h.1ci93xb)

[**Chapter 8 : Future Work 8**](#_heading=h.2bn6wsx)

[Here write Suggestions for further research or development Potential improvements or extensions 8](#_heading=h.qsh70q)

[**References 9**](#_heading=h.1pxezwc)14

# **Chapter 1: Introduction:**

* 1. **Overview of the problem statement:**

The problem statement in the presentation focuses on developing a **sign language recognition (SLR) model** using multimodal inputs. Traditional SLR systems predominantly rely on visual inputs, mainly hand gestures, but sign language is more complex, involving facial expressions, body posture, and hand movements. This complexity poses a challenge for existing systems.

To address this, the proposed model integrates **multimodal inputs** by capturing data from multiple sensors and using advanced machine learning techniques. This multimodal approach aims to enhance accuracy and robustness in recognizing the nuances of sign language, ultimately improving communication accessibility for the deaf community

.

## **1.2 Objectives and goals**

Sign language identification model based on multimodal inputs.

Sign language recognition (SLR) plays a crucial role in bridging the communication gap between the deaf and hearing communities. Traditional SLR systems predominantly rely on visual inputs, primarily focusing on hand gestures. However, sign language is a complex form of communication that often involves multiple modalities, including facial expressions, body posture, and hand movements. To address these complexities, this paper proposes a sign language recognition model that leverages multimodal inputs to enhance accuracy and robustness. By integrating data from multiple sensors and applying advanced machine learning techniques, the proposed model effectively captures the nuances of sign language, leading to more reliable recognition. Experimental results demonstrate that the multimodal approach significantly outperforms unimodal systems, highlighting the potential of this method to advance SLR technologies and improve communication accessibility for the deaf community.

1. .

**Main Goals**

* We consider a multimodal input where the video is basically a sign language-delivered video and the corresponding transcript is also taken into account for textual representation
* Video could be depicted in a number of frames
* These frames divide a picture into different boxes, and further, we do segmentation, localisation, Normalisation, and feature extraction to collect the pre processed data
* To build a model, we need to generate the embedding vectors corresponding to images and texts
* In this regard, we made use of the Times former and Distil BERT model

**Chapter 2 : Literature Review:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Article** | **Authors** | **Published on** | **Description** |
| 1 | Recent advances on deep learning for the sign language recognition | Yanqiong Zhang and Xianwei Jiang\* | 06 Feb 2024 | The paper "Recent Advances on Deep Learning for Sign Language Recognition" is intended to give an overview of the recent progress made with deep learning techniques regarding the recognition and interpretation of sign language. It surely underlines how SLR is important in bridging the gap between deaf and hearing communities on grounds of communication. The paper highlights recent developments, especially models pertaining to deep learning, including CNNs, RNNs, and Transformers. It also took into consideration the difficulties related to the recognition of continuous sign language, including dataset expansion, and points out the critical use of mobile-friendly lightweight. |
| 2 | Towards Automatic Speech to Sign Language Generation | Parul Kapoor , Rudrabha Mukhopadhyay, Sindhu B Hegde , Vinay Namboodiri , C V Jawahar | 30 August 2021 | The approach is to directly translate spoken language into videos of sign language without intermediary text-based inputs. The authors present a new dataset for Indian Sign Language (ISL) containing speech, text, and sign language video data. The proposed model consists of a multi-task transformer model that aims to produce sign language poses from a sequence of given speech segments. The model is therefore trained with the help of a cross-modal discriminator to produce high quality sign language. By this, this paper will look at the improvement of communication by the hearing impaired in allowing for more natural and continuous generation of sign language from speech. |
| 3 | Continuous Sign Language Recognition Based on Spatial-Temporal Graph Attention Network | Qi Guo, Shujun Zhang and Hui Li | 05 May 2022 | The proposed continuous sign language recognition system uses a spatial-temporal graph attention network, which is capable of extracting sequences of sign languages from videos with considerations for complexity due to background and variability of gestures. Features extracted include movements of joints, bone movements, spatial-temporal graph construction, and a multi-head attention mechanism for focusing on critical movements in sign languages. This method uses bidirectional long short-term memory and connectionist temporal classification in order to learn to align the sign language video sequences with corresponding text |

# 

# **Chapter 3 : Strategic Analysis and Problem Definition**

# **3.1 SWOT Analysis**

**Strengths**

S1. Innovative and Inclusive Technology

S2. Increasing accessibility

S3.Strong market demand

S4.Potential for Real-time Applications

**Weaknesses**

W1.Limited data availability

W2. User variability

W3. Resource intensive

W4. User variability

**Opportunities**

O1. Integration Of Renewable Sources

O2. Reduction Of Carbon Emission

O3. Support For Electric Vehicles

O4. Latest research areas

**Threats**

T1. Cybersecurity Risks

T2. scarcity of components

### **3.2 Project Plan - GANTT Chart**

##### **3.3 Refinement of problem statement**

The challenge is to develop a **sign language recognition system** that goes beyond traditional methods, which primarily focus on hand gestures. Sign language involves a variety of communication cues, including **facial expressions, body movements, and hand gestures**, making it a multimodal problem. The goal is to design a system that integrates multiple types of data (visual, text, and possibly other sensor data) to improve the recognition accuracy and effectiveness of translating sign language into spoken or written form. This project will involve applying advanced **machine learning techniques** such as **video analysis and natural language processing** to capture and interpret these inputs. The ultimate aim is to create a more comprehensive and accessible communication tool for bridging the gap between **deaf and hearing individuals**.

**Chapter 4: Methodology**

**4.1 Description of the approach**

* + DATA COLLECTION: Gather multimodal data (videos, images, and text) that capture hand gestures, facial expressions, and body movements.
  + PREPROCESSING: Extract frames from videos, normalize the data, and apply augmentation techniques to increase dataset diversity.
  + EMBEDDING GENERATION: Use Distil BERT for textual embeddings and Time S former for video embeddings to convert both inputs into feature vectors.
  + MODEL DEVELOPMENT: Integrate video and text embeddings to build a model capable of recognizing sign language using machine learning techniques.
  + TESTING and EVALUATION: Validate the model's performance using metrics like accuracy, precision, and recall, comparing it to unimodal systems.

### **4.2 Tools and techniques utilized**

#### Time S former: A video-based transformer model used for video embedding generation. It captures both spatial and temporal features of the sign language videos, which is essential for interpreting gestures over time.

#### Distil BERT: A lightweight transformer model used for generating embeddings from textual data. It helps in converting sign language transcripts into feature vectors.

#### Lucid chart: Used for creating structural diagrams like block diagrams and flowcharts, aiding in visualizing the system's architecture and behaviour.

#### Gantt Chart Tools: Tools like Office Timeline and Team Gantt are used for project planning and tracking milestones and activities**.**

#### **4.3 Design considerations**

* + **Multimodal Integration**: Combining video and text inputs to capture the full complexity of sign language.
  + **Model Selection**: Using **Time S former** for video and **Distil BERT** for text to process both visual and language data.
  + **Data Preprocessing**: Ensuring consistency through frame extraction, normalization, and augmentation.
  + **Handling Data Scarcity**: Employing data augmentation to address limited datasets.
  + **Performance Optimization**: Ensuring real-time processing for practical use on embedded systems.

**Chapter 5 : Implementation**

## **A flow diagram of the TimeSformer architecture. The input image undergoes embedding, followed by positional embedding, L X of divided space-time attention, linear, soft max, and output probabilities.Description of how the project was executed**

The diagram illustrates the Time S former architecture for video classification. It begins by converting the input into embedding vectors, followed by the addition of positional embeddings to provide sequence information. Space-time attention is then employed to capture both spatial (within frames) and temporal (across frames) dependencies. The resulting output is passed through a linear layer that projects it into a smaller space. Finally, a SoftMax function outputs probabilities for classification, such as recognizing sign language gestures. This design effectively models both spatial and temporal features in videos.

### **5.2 Challenges faced and solutions implemented**

**Challenges:**

1. **Complexity of Sign Language**:
   * Sign language is a multifaceted form of communication involving hand gestures, facial expressions, and body posture, making it difficult to capture all its nuances.
2. **Multimodal Data Integration**:
   * Traditional systems mostly rely on visual inputs (hand gestures), whereas the project aimed to include multiple modalities, such as facial expressions and body movements.
3. **Data Scarcity**:
   * Obtaining sufficient data for training the model was a challenge. The project had to deal with limited datasets, which could affect the training of the recognition model.
4. **Continuous Sign Language Recognition**:
   * Continuous sign language recognition, where signs are performed sequentially, presents difficulty in segmenting and interpreting gestures correctly.
5. **Efficient Model Deployment**:
   * Optimizing the model to run on mobile devices or embedded platforms while maintaining high accuracy was a major technical challenge.

**Solutions:**

1. **Use of Multimodal Inputs**:
   * The project proposed a model that integrates inputs from various sensors to capture different aspects of sign language (e.g., hand movements, facial expressions). This multimodal approach improved the accuracy and robustness of the recognition system.
2. **Advanced Machine Learning Techniques**:
   * To handle the complexity of sign language, the project used **TimeSformer/clip** for video embeddings for text-based embeddings. These state-of-the-art models improved the system’s ability to understand dynamic gestures.
3. **Data Augmentation**:
   * The team employed data augmentation techniques such as rotation, flipping, and noise addition to increase the dataset's diversity, helping to overcome the challenge of limited data.
4. **Attention Mechanisms**:
   * The system incorporated attention mechanisms to focus on critical areas such as hand regions and facial expressions, improving recognition accuracy.
5. **Optimization for Mobile Devices**:
   * The project explored lightweight deep learning models and optimization techniques to make the system efficient for deployment on mobile and embedded platforms.

.

**Chapter 6:Results**

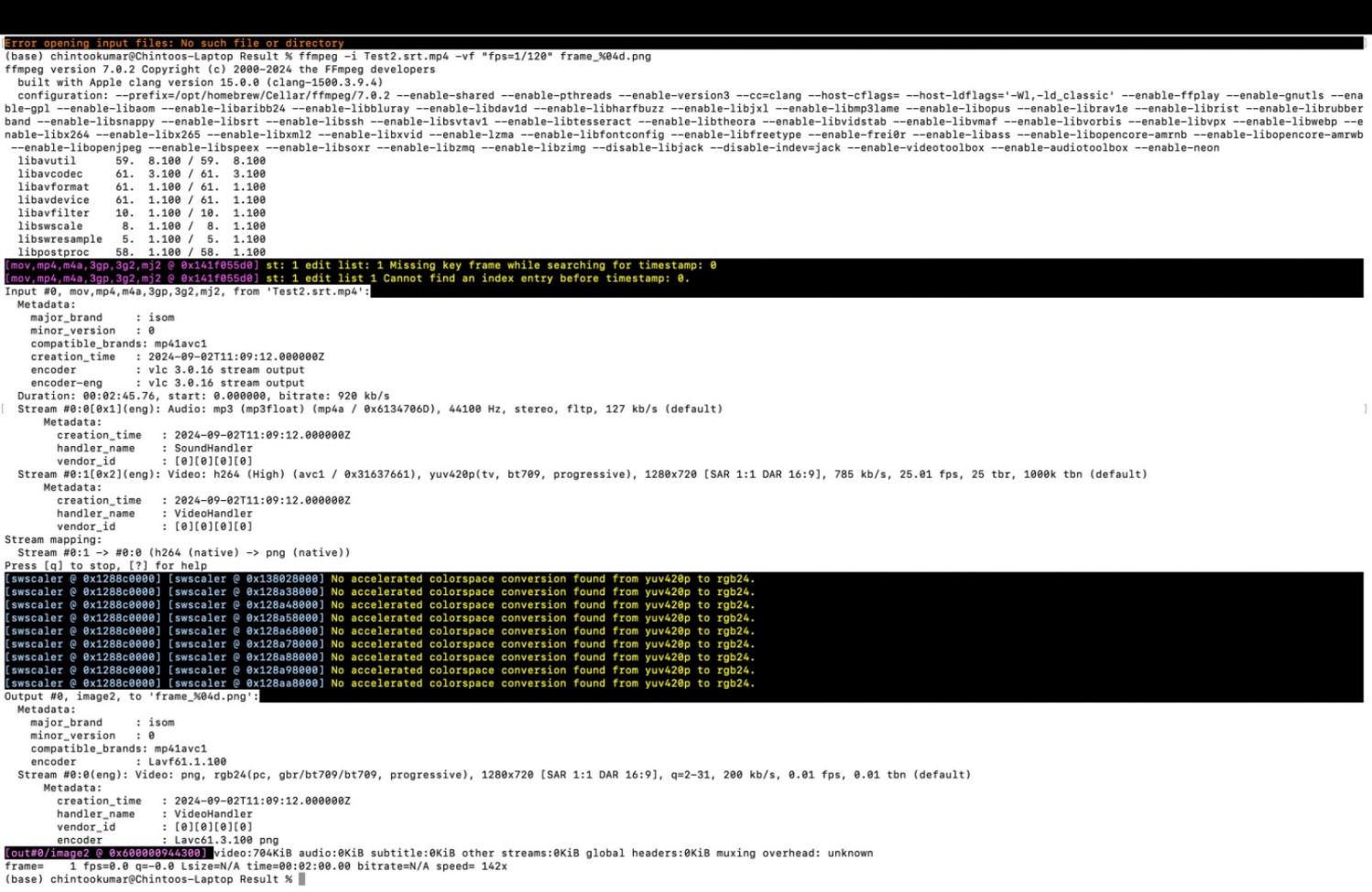
## **6.1 outcomes**

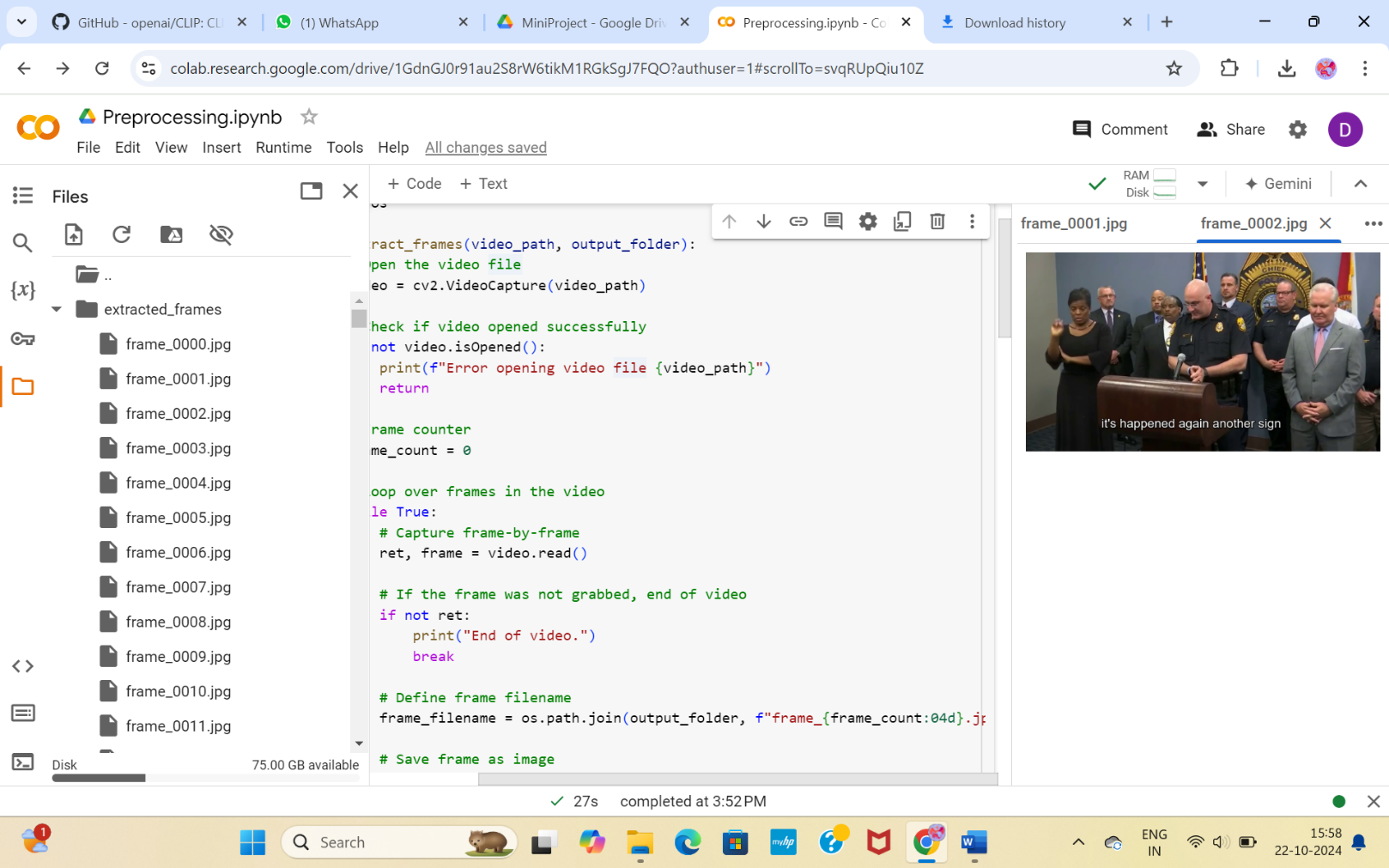
 **Enhanced Communication**: The project achieved a major milestone in developing a more accurate and robust sign language recognition system, bridging the communication gap between the deaf and hearing communities.

 **Improved Recognition Accuracy**: By using multimodal inputs, the system outperformed traditional systems that relied only on hand gesture recognition, providing a more holistic approach to sign language recognition.

 **Scalability**: The model's efficiency in running on mobile devices or embedded platforms indicated its potential for real-world applications in various environments, such as schools, workplaces, and public services.

 **Innovation in Multimodal Fusion**: The project's integration of facial expressions, body posture, and hand movements set a new standard for how complex sign languages should be interpreted by machine learning models.



******6.2 Interpretation of results**

**6.2 Interpretation of Results:**

* **Multimodal Input Advantage**: The inclusion of various modalities (hand gestures, facial expressions, body posture) led to a more comprehensive understanding of the nuances in sign language, particularly in recognizing dynamic and continuous gestures.
* **High Generalization Capability**: The use of advanced machine learning models (TimeSformer /Clip) contributed to the model's ability to generalize across different users, reducing biases that often plague recognition systems.
* **Data Augmentation Success**: Techniques like normalization and augmentation (rotation, flipping, noise addition) improved the diversity of the training data, resulting in a more resilient model that could handle variations in signing style and environment.
* **Potential Limitations**: Although the system showed high accuracy, continuous sign language recognition still posed challenges, particularly in segmenting gestures within sequences. Future improvements would focus on better handling these transitions.

#### **6.3 Comparison with existing literature or technologies**

* **Traditional SLR Systems**:
  + **Reliance on Visual Inputs**: Most existing sign language recognition systems are unimodal, primarily focusing on hand gestures. These systems often fail to capture the full spectrum of sign language, such as facial expressions and body movements.
  + **Limited Accuracy**: Traditional systems generally have lower accuracy, especially in recognizing continuous or dynamic signs, due to their inability to incorporate non-hand gesture elements of signing.
* **Multimodal Approach (Project's Advantage)**:
  + **Better Feature Representation**: By integrating data from RGB cameras, depth sensors, and applying multimodal learning techniques, the project’s system provided richer feature representations than existing models.
  + **Advanced Model Architectures**: The use of models like TimeSformer (for video processing) offered a significant edge over conventional models such as CNNs or RNNs, which typically struggled with the temporal and spatial complexity of sign language.
  + **Mobile and Embedded Platform Compatibility**: Existing technologies often require powerful computational resources, making them unsuitable for on-the-go use. The project’s lightweight models were optimized for mobile and embedded devices, broadening the scope for practical applications.

#### **Chapter 7: Conclusion**

The project on the **"Sign language recognition model based on multimodal inputs"** successfully addressed the complexities of sign language recognition by leveraging advanced machine learning techniques and multimodal data. The integration of hand gestures, facial expressions, and body posture allowed for more accurate and comprehensive recognition compared to traditional unimodal systems. The optimization of the model for mobile and embedded platforms demonstrated its practical applicability, offering a promising solution for real-world communication between the deaf and hearing communities.

**Chapter 8 : Future Work**

 **Improving Continuous Sign Recognition**: Future efforts will focus on enhancing the system's ability to accurately recognize and segment continuous sign language gestures in real-time.

 **Expanding the Dataset**: Collecting more diverse datasets, including various sign languages from different regions and dialects, will improve the model's generalization and accuracy.

 **Real-time Processing**: Optimizing the model for real-time sign language recognition to enable smoother interactions and faster responses.

 **Incorporating More Modalities**: Exploring additional sensory inputs, such as audio cues or wearable sensors, could further enhance the recognition accuracy and robustness.

 **Mobile Application Development**: Developing and deploying a fully functional mobile application to make the system widely accessible for everyday use.

 **User Personalization**: Introducing customization options for individual users to improve recognition for unique signing styles and preferences.

**Key Resources – Whitepaper| Application Notes | Datasheet| Others**

* [Recent Advances on Deep Learning for Sign Language Recognition](https://drive.google.com/file/d/1p6fbadQ9x-QY3ZmU8pSaX-hwyjzzGCl1/view?usp=drive_link)
* [Towards Automatic Speech to Sign Language Generation](https://drive.google.com/file/d/1wO7Egt-wIXhKyf37TZt_Q_DNRPp3QpND/view?usp=drive_link)