IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EFFICIENT REAL TIME SIGN LANGUAGE DETECTION BASED ON COMPUTER VISION POWERED BY DEEP LEARNING

Immaraju Giridhar
Master of Technology in Information Technology
Department of IT & CA
Andhra University College of Engineering,
Vishakapatnam,AndhraPradesh

***Abstract*—**The Sign Language Detection system is developed for detecting the alphabets of American Sign Language (ASL). Conventionally, sign languages comprise of finger spelling. For detecting the signs, the Regions of Interest (ROI) are related and tracked employing the Landmarks feature of Media Pipe. Then by using Open CV, we capture the landmarks of the hands, and the key points of landmarks are stored in an NumPy array. Then we can train the model on it by using TensorFlow, Keras and CNN. Finally, the model can be tested in Realtime by ingesting live feed incoming from the webcam. Realtime Sign Detection is one of the essential applications to be used by the deaf and dumb people as it help them to communicate with others effectively.Historically,various methodologies for detecting sign were employed by the Machine Learning Algorithm ,by training it on the image data .However, now we are employing Deep Learning Models to simplify and speed up the process of Realtime sign detection and recognition and producing equal or more accuracy using smaller amounts of data. Keywords—American Sign Language, Realtime, Media Pipe, Landmarks, Key points, OpenCV, NumPy, CNN, TensorFlow, Keras.

I. INTRODUCTION

A. Background

The World Health Organization recognized the conservative number of specially abled people in terms of speech and hearing to be around 450 million. In this context, modern studies are increasingly more focused on making disabled people to communicate with anyone more effectively without being lost in translation. Sign language is thought to be a significant tool of communication for specially abled people. Conventionally sign language has unique meaning attached to each and every gesture, thereby complex thoughts,ideas,words and sentences can be explained with the skillful combination of various basic fingerspelling. Sign language is basically a gesture-oriented language, used predominantly for communication of specially abled people. It is essentially a non-verbal language which is usually used by specially abled people to communicate with more rigor and vigor with each other in the real world.

Humans are social animals who crave for a sense of belonging and purpose, which is satiated by social interactions. Lack of interpersonal communication can damage a person mentally by making them feel alone in the world. In this context a proposal is made to build a system for seamless non-verbal communication, thereby providing a place of interaction to build social connections. The proposed system is built using OpenCV libraries and Media Pipe libraries. The Convolutional Neural Network (CNN) is utilized for a reasonably good accuracy of 91% by consuming minimal data in comparison with similar systems.

B. Motivation

- With the advent of AI and ML revolution, a proposal is made for a meaningful technological value addition to the existing systems by constructing an efficient deep neural network for interpreting non-verbal communication.
- To develop a platform for seamless exchange of non-verbally communicated thoughts, ideas and meaningful conversations.

II. OBJECTIVE AND METHODOLOGY

A. Objective of this Project

The project is developed to build a reliable system for Realtime sign language detection using Media Pipe and Computer Vision. The training of the model is done using CNN.

B. Methodologies and Technologies employed

The Systematic approach to detect sign language to identified sign gestures:

- In First phase we will detect the hands using webcam's live feed, employing OpenCV library.
- From the different coordinates and angles of the hands, landmarks are captured.
- The Key points of the landmarks are stored in the NumPy array so that they can be operated on, in later stages.
- Before capturing the landmarks, the unique label associations of the dataset are leveraged
- Live fed input data is classified in accordance with the labels provided.

The model is trained with the help of CNN (Convolutional Neural Network) layers by providing the key points of hand gesture landmarks as input and labels of the gestures captured as output.

- Finally, the model is tested using the live feed of webcam in Realtime.
- A Web Application is created for the project using python flask.

III. LITERATURE SURVEY

A. Brief History of Sign Language:

Sign language has been systemically used by First World societies since the 1600s. This language is utilized by early practitioners as a visual kind of non-verbal language or as a means of communication using hand gestures. Sign language is a combination of legacy hand gestures, mimics, hand signs, symbols and finger spellings. Also, various hand positions are employed to convey the distinct letters, words and characters.

B. Sign Language Interpretation

- Generally, single frame-based hand gesture recognition systems using the web camera, preprocess the input images by transforming them into grayscale images, and then uses Gaussian filter for noise removal from the webcam captured images. To obtain a binary image from the grayscale image, thresholding, a segmentation technique is employed which helps one to gauge the pixel intensity, The final stage involves contour extraction for hand gestures' detection and recognition.

- However, this system is implemented by employing Media Pipe landmark extraction, which requires no image data preprocessing for sensing the hand movement accurately without any background noise or skin complexion noise, as only the hand landmarks irrespective of image background and hand imperfections are captured during the process of data collection.

C. Technology Review

Technology review exhibits the research done about technologies that are being used frequently in recent times, on the system. This review covers the advantages and limitations, basically the scope of technologies employed by authors of the system. The camera module is also implemented with the help of the OpenCV. User needs to capture his video hand gesture movement and then it is sent to the server, where in the backend, The CNN model has been trained to predict the meaning of the hand sign.

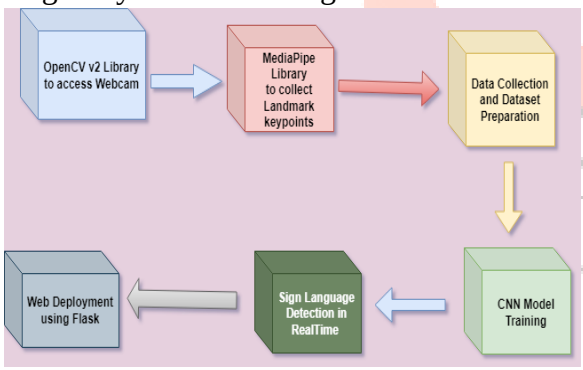
IV. SYSTEM DESIGN

System design shows different phases of the system

A. Preprocessing Layer

- The webcam live feed i.e., continuous stream of frames data is given as input. This stream is divided into a number of frames and fed to the system. Then, the frames are pre-processed by employing an adopted algorithm according to system requirements. The landmarks of hand in the video frame are captured using Media Pipe and the coordinates of key points of the landmarks are stored in NumPy array, to be used for label prediction in final stage.

Fig 1. System Flow Diagram.



- Data Loading and Data Splitting: All the web camera frame landmark data from the landmark preprocessing module is procured and retrieved and this data is split into training and testing dataset in 70:30 ratio.
- Training Phase: In this phase, the model is trained by feeding a training dataset to it, the pre-trained models are further used for training to address the key point classification.
- Trained Model: The pre-trained model is the outcome of the training phase. This can be directly employed for the detection process in subsequent testing phases.

B. Prediction Layer

- Pre-processing Module:

The video stream from the web user is given as input in this layer. This stream of frames will undergo the same set of landmark preprocessing steps as the detection phase.

- Load Trained Model:

A pre-trained model from the training phase is used to predict the meaning of hand sign to be interpreted.

- Final Detection:

In this phase, the model gives the predicted fingerspelling of the alphabet relating to hand sign shown.

V. PROPOSED WORK

A. System Implementation

This proposed system is implemented using flask web framework where the web user ready to converse through signs is the client and the procured dataset acting as database to be accessed by server which is trained on the built gesture prediction CNN model.

B. System Modules

Real-time hand gesture involves:

1. Input webcam live feed streaming.
2. Hand Detection using Media Pipe.
3. Landmark preprocessing using developed algorithm.
4. Sign recognition and label prediction using the developed CNN model.
5. Display of predicted text using flask web framework.

The system is modularized into the following modules:

1.Input Image from Webcam:

In the first phase, Real-time gestures of the web user are captured by making use of a web camera. In this module, the video stream will be captured in real-time and will be yielded through flask API, where the landmarks will be preprocessed.

2.Hand Detection:

After getting the image from the user the python script will detect the hand from that image and crop the image into a square size. The hand detection and tracking are achieved by employing the media pipe framework which is mainly used for building face,pose,hand,palms,legs and feet landmark composition & detection and any other similar body component landmark detection.

3.Landmark Preprocessing:

In this module, the hand landmarks will be preprocessed and converted into floating point numbers to be stored in NumPy arrays. Firstly, Hand landmark coordinates will be given a reference point with respect to the base landmark coordinates of the Hand palm. The distance is calculated for every landmark point coordinate by setting a default reference landmark coordinate. This process constitutes our algorithm. The landmark pre-processed using this algorithm will be trained on the built CNN for the prediction of the label assigned to each of this floating-point comma separated value data created during dataset preparation.

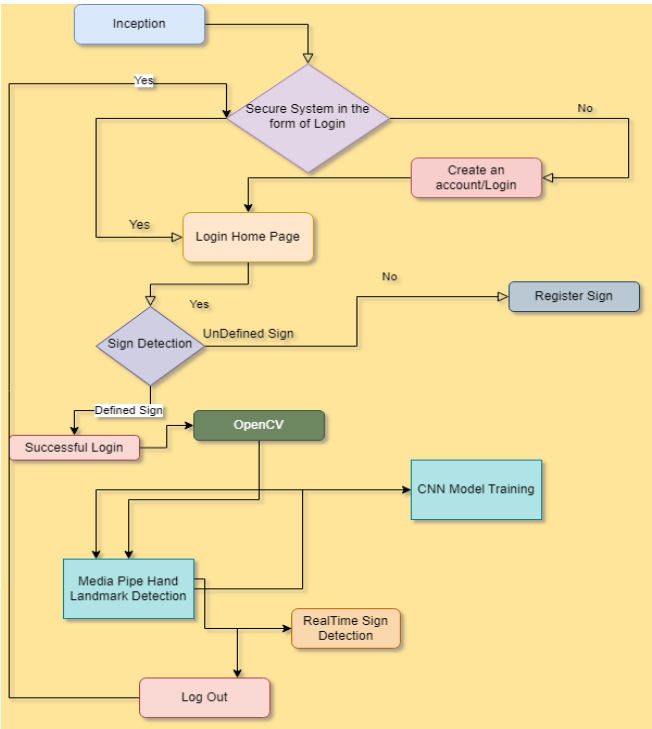


Fig 2 System Architecture Flowchart

- CNN powered Sign detection and label recognition: The preprocessed landmark frame will be fed to the CNN model. Our CNN model is trained using thousands of hand landmarks signs and is capable of predicting hand landmark driven labels with more than 91% accuracy. The output of this model will be a character from A-Z represented by 26 classes. Exploiting this character, a word can be formed, and further a suitable sentence can be produced.
- Display of Predicted Text: This module involves flask web framework to accumulate the labels predicted and store them in webpages for the predicted text to be read by the application web users.

4. Dataset Creation

One of the biggest roadblock neural networks faces is procuring a good dataset that accommodates a diverse and wide variety of landmark data. While researching the work, we observed that there are a few already pretrained landmarks offered by the media pipe library for some hand gestures. But after studying those pretrained landmarks we found out that most of the landmarks have similar gesture formations which don't match our gesture recognition requirements as well as real-time conditions like camera noise and angular and positional coordinates of landmarks composed in a frame. Feeding these noisy and error prone landmark data to the CNN model may provide good accuracy on the test dataset but very poor to worst results on real-time label prediction. To address this dataset issue, the landmark dataset is created with thousands of landmark data relating to various hand gestures. This dataset includes more than 1000 columns of landmark floating-point comma separated value data belonging to each class ranging from
A-Z.

5. CNN for Sign Recognition

A Convolution Neural Network (CNN) is a Deep Learning algorithm which is widely used to deal with spatial data like computer vision problems. This algorithm gives significance to numerous distinct aspects or objects in the pixel image/landmark frame and on the basis of that, it can distinguish pixel image/landmarks better compared to other machine learning algorithms. During the traditional period of early methods, filters like gaussian filters etc., are designed manually, with sufficient training, neural networks can learn these nuanced characteristics.

Architecture of CNN: The CNN model developed is trained by feeding thousands of landmarks of sign gestures, which are divided into 26 classes ranging from A-Z. This dataset contains hundreds of megabytes of floating-point landmark co-ordinate data belonging to every class. The convolution neural network model for this project consists of a total of 4 layers and there are 2 layers of fully connected dense layers. The first layer is a convolution input layer with 42 nodes, which is responsible for identifying hand landmark features like point coordinates and their angles with respect to default base landmark coordinate. A rectifier linear unit (RELU) is employed to eliminate any negative values on the map and replace them with 0. After convolution layer, there is a dense layer with 40 nodes which is followed by two fully connected Dense layers with 20 and 27 nodes.

6. CNN Accuracy

The variation between the expected output values and the real time predicted values gives us the accuracy of our CNN model. The model fitted the ingested training dataset with an accuracy of 92%. The accuracy of the employed CNN model on the validation/testing dataset that is used for tuning the hyperparameters turned out to be 92%.

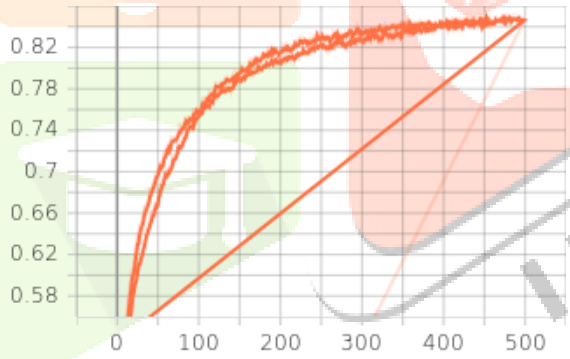


Fig 3 CNN Accuracy shows the accuracy of the model concerning epochs.

7. CNN Loss

It is a methodology to evaluate your modeling of the algorithm with the dataset. In case of wrong prediction, the loss function is expected to be of a higher value. Otherwise, it'll be a lower value. The system employs the "Categorical Cross entropy loss" function in order to train our model where ti and si are said to be ground truth and the CNN score for each class in CC.Generally an activation function (SoftMax/Sigmoid) is applied to the scores prior to the computation of CE Loss.

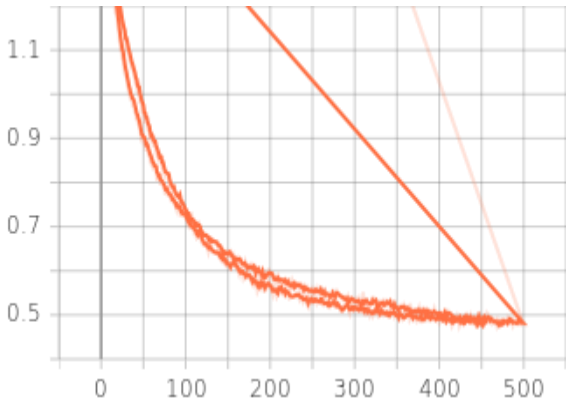


Fig 4 CNN Loss shows the loss of the model concerning epochs.

VI. MODEL DESCRIPTION

The model that is employed is built with TensorFlow and Keras using Convolutional Neural Network which is mostly used to work on spatial data especially for computer vision related projects or any other projects dealing with image data. The CNN model architecture comprises of the following layers:

- CNN Input layer;42 nodes
- CNN Dense layer; 40 nodes
- Fully connected Dense layer; 20 nodes
- Fully connected Output layer; 27 nodes

The final output layer is a fully connected Dense layer with 27 nodes, which is equal to number of labels. In each and every layer, we make use of Relu as activation function and only for output layer we assign SoftMax activation function.

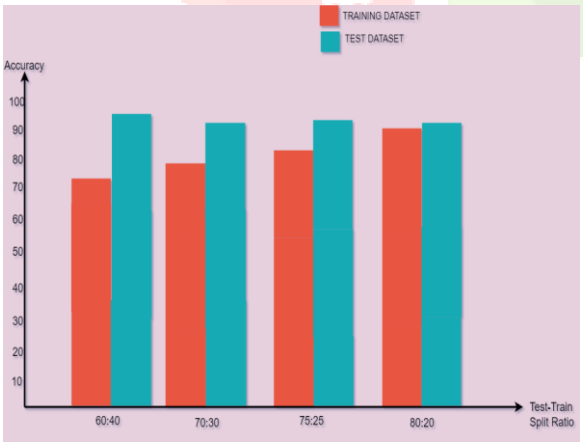


Fig 5. Test/Train Split Ratio Comparison

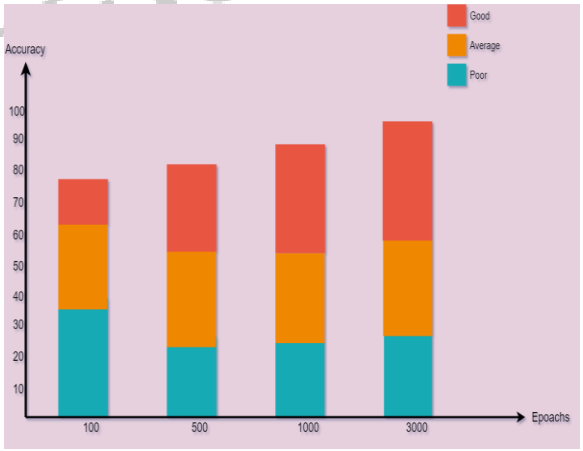


Fig 6. Model Accuracy Comparison

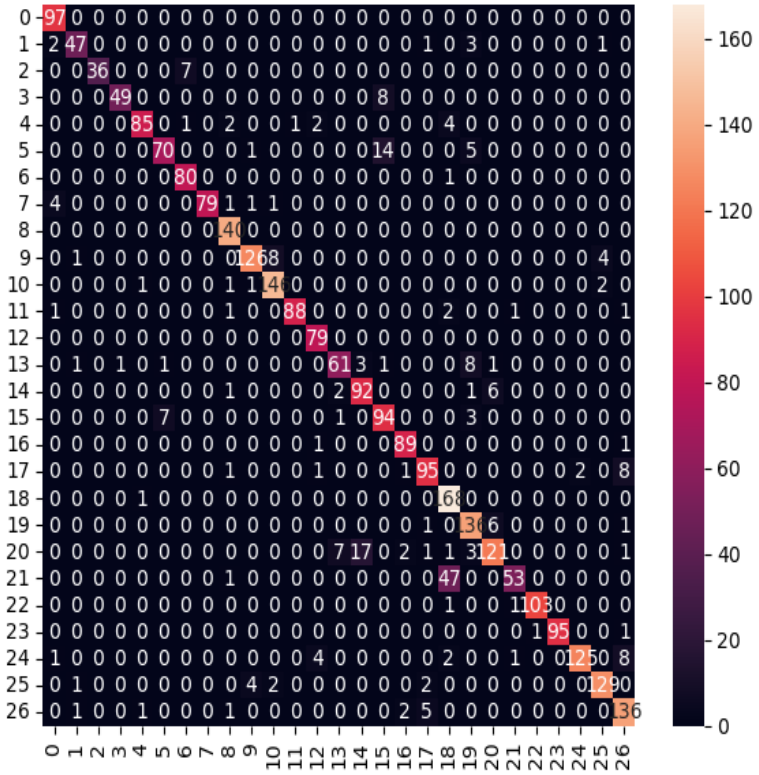


Fig 7. Confusion Matrix

Classification Report

	precision	recall	f1-score	support
0	0.92	1.00	0.96	97
1	0.92	0.87	0.90	54
2	1.00	0.84	0.91	43
3	0.98	0.86	0.92	57
4	0.97	0.89	0.93	95
5	0.90	0.78	0.83	90
6	0.91	0.99	0.95	81
7	1.00	0.92	0.96	86
8	0.94	1.00	0.97	140
9	0.95	0.91	0.93	139
10	0.93	0.97	0.95	151
11	0.99	0.94	0.96	94
12	0.91	1.00	0.95	79
13	0.86	0.79	0.82	77
14	0.82	0.90	0.86	102
15	0.80	0.90	0.85	105
16	0.95	0.98	0.96	91
17	0.90	0.88	0.89	108
18	0.74	0.99	0.85	169
19	0.86	0.94	0.90	144
20	0.90	0.79	0.84	153
21	0.95	0.52	0.68	101
22	0.99	0.98	0.99	105

23	1.00	0.98	0.99	97
24	0.98	0.89	0.93	141
25	0.95	0.93	0.94	138
26	0.87	0.93	0.90	146
accuracy			0.91	2883
macro avg	0.92	0.90	0.91	2883
weighted avg	0.91	0.91	0.91	2883

IX. DEPLOYMENT

We deployed our project using Flask. We also created a login system on our website. This login system lands us on sign recognition webpage to perform our sign detection and obtain predictions.

X. CONCLUSION

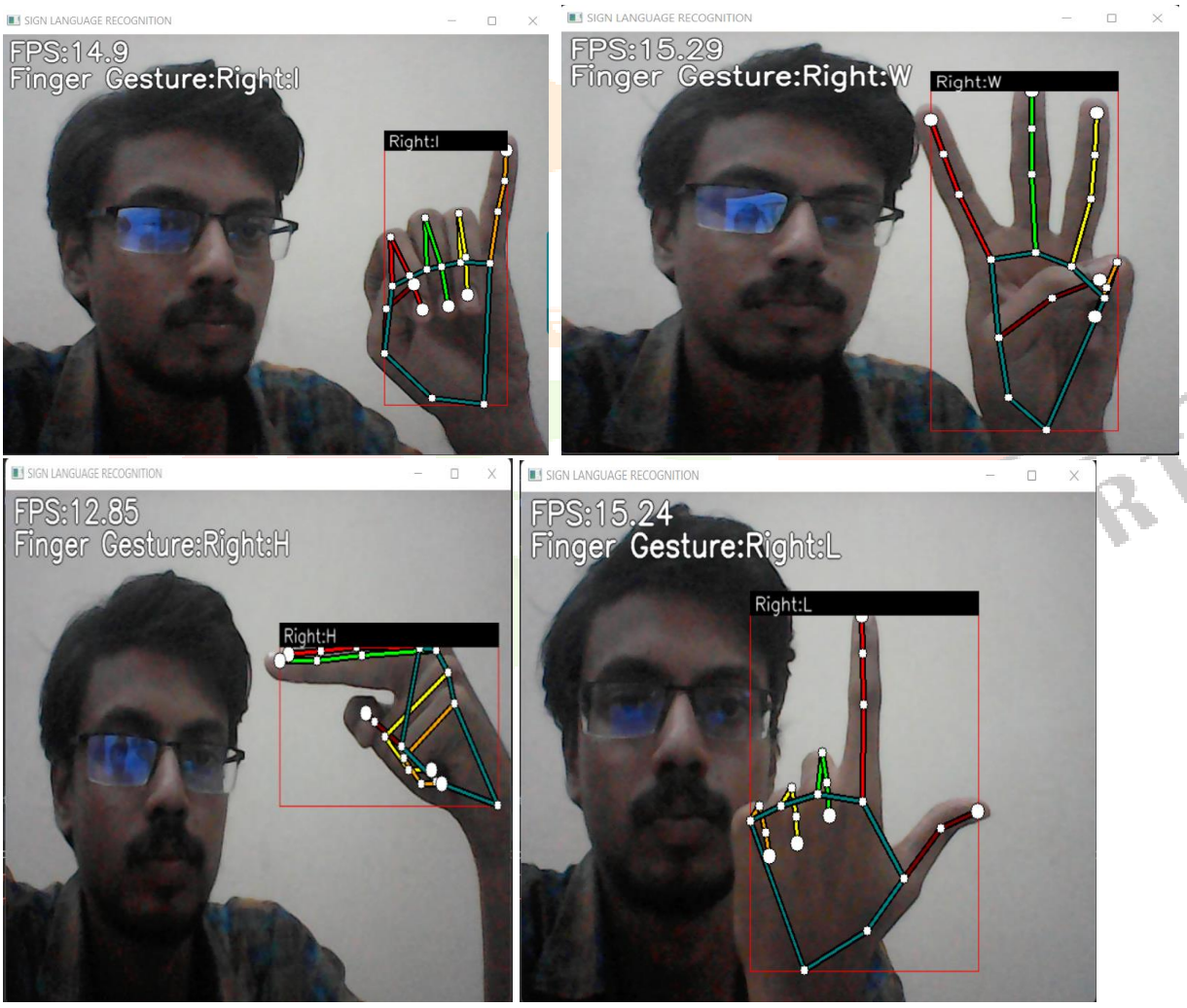
This project of Realtime Sign Language Detection is built by employing Media Pipe Library to capture the landmarks of hands and palm. A CNN model is utilized to train the dataset and predict the sign language gestures. In this system, the CNN model requires less storage as it works on landmarks. The objective of this project is met partially. The program is able to execute and capture live web cam feed within the stipulated time frame and the resulting prediction's accuracy is reasonably good enough to satisfy project requirements and standards. This prototype can be further tested and evaluated on various parameters of sustainability and scalability.

Models Used	Activation Function Employed	Accuracy Obtained	Loss Obtained
Epochs = 100	ReLu/Softmax	0.8965	0.4576
Epochs = 500	Tanh/Softmax	0.9184	0.4382
Epochs = 1000	Tanh/Softmax	0.9279	0.4265

Table1. Model Comparison

Model Used	Activation Function Employed	Accuracy Obtained	Loss Obtained
CNN Model trained on Landmark	ReLU/ Softmax	0.9239	0.4376
CNN Model trained Image Data	ReLU	0.9487	0.3617

Table2. Model Parameter Comparison



REFERENCES

1. https://mediapipe.dev/
2. https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneousface.html
3. https://poloclub.github.io/cnn-explainer/
4. https://www.tensorflow.org/tutorials/images/cnn
5. https://arxiv.org/abs/1702.01923
6. https://www.kaggle.com/datasets/vaishnaviaonwane/indian-signlanguagedataset.
7. https://developers.googleblog.com/2021/04/signall-sdk-sign-languageinterface-using-mediapipe-now-available.html
8. Arpita Halder, Akshit Tayade, "Real-time Vernacular Sign Language Recognition using Media Pipe and Machine Learning" www.ijrpr.com ISSN 2582-7421
9. Shivangi Nagdewani,Ashika Janet al,A review on methods for speech-to-text and text-to-speech conversion,2018.
10. Murat Taskiran, Mehmet Killioglu, Nihan Kahraman, "A Real-Time System for Recognition of American Sign Language by using Deep Learning" IEEE 18044537
11. American Sign Language Recognition with Convolutional Neural Networks" stanford.edu