Sign Language Recognition using Mediapipe and RNN Models(LSTM and GRU)

M. Kishore Babu[1], M. Venkata Karthik Reddy[2], M. Lavanya[3],N. Sai Raghu Vardhan[4], M. Vijay Kumar[5]

[1]Assistent Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology (Autonomous), Guntur, AP

[2,3,4,5] UG Students, Department of CSE, Vasireddy Venkatadri Institute of Technology (Autonomous), Guntur, AP

[1]kislatha@gmail.com

[2]karthikmedagam@gmail.com

[3]lavanyaml.1702@gmail.com

[4]sairaghu241@gmail.com

[5]vjaymogili@gmail.com

Abstract:

This project aims to bridge the communication gap between regular individuals and the differently abled by utilizing MediaPipe framework and RNN models like LSTM and GRU. The integration of MediaPipe, a popular real-time hand tracking library, with Recurrent Neural Networks like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) to enhance the accuracy and efficiency of sign language recognition. The project begins by gathering recorded video samples of different signs. The proposed system utilizes MediaPipe to extract key hand and facial landmarks and their temporal sequences from video data. These landmarks are then fed into both LSTM and GRU, a type of recurrent neural network's designed to capture temporal dependencies in sequential data. We trained both these models with 15 different signs. Between these two models GRU model stands out as the best with an accuracy between 96%-99% where LSTM model gets an accuracy between 90%-94% when we tried to train with different number of signs.

.

Keywords — **Mediapipe, LSTM, GRU, RNN, Hand gestures, Sign Language.**

I.INTRODUCTION

This Have you ever wished you could have a real-time conversation with someone who uses sign language? Sign language recognition (SLR) is a field of technology that's working towards making that possible! Imagine you're watching a news report or a video call, and someone is signing. SLR technology aims to bridge the communication gap by "reading" those signs and converting them into something everyone can understand, like text or spoken words. This can be done through software that analyzes videos or images, focusing on hand movements, facial expressions, and even body posture.

Think of it like a special translator for sign language. It watches the way someone signs, just like you'd listen to someone speaking, and figures out what message they're trying to convey. While SLR is still under development, it has the potential to revolutionize communication for people who are deaf or hard of hearing, as well as for anyone wants to break down language barriers.

In the year 2021, it was estimated that there were around 18 million people with hearing disabilities in India. This includes individuals who are Deaf as well as those who are hard of hearing.

Deaf is a disability that impair their hearing and make them unable to hear, while mute is a disability that impair their speaking and make them unable to speak. Both are only disabled at their hearing and/or speaking, therefore can still do much other things. The only thing that sperate them and the normal people is communication. If there is a wat for normal people and deaf-mute people to communicate, the deaf-mute people can easily live

like a normal person and the only way for them to communicate is through sign language.

While sign language is very important to deaf-mute people, to communicate both with normal people and among themselves, is still getting attention from the normal people. We as the normal people, tend to ignore the importance of sign language, unless there are loved ones who are deaf-mute.

So, what is sign language? Sign language is a set of hand gestures, facial expressions and body motion that represent words. Each country has its own sign language. Indian Sign Language (ISL) is the primary sign language used in India. Indian Sign Language (ISL) is a unique and expressive form of communication used by the deaf community in India. It has its own vocabulary, grammar, and syntax. ISL is influenced by local languages and has regional variations. It's an amazing way for people to communicate and connect. It plays a crucial role in education, daily communication, and various aspects of life for individuals with hearing impairments in India.

This study contributes to the ongoing efforts in creating inclusive and technology driven solutions for the deaf and hard of hearing community. The proposed approach not only advances the state of the art in sign language recognition but also emphasizes the potential for integrating cutting edge technologies for the benefit for individuals with hearing impairments.

II. LITERATURE REVIEW

An Speech disorders affect roughly 11.5% of the US population and approximately 18.5 million individuals worldwide have a speech, voice, or language disorder. Figures of WHO state that nearly 466 million people that comprise approximately 5% of the World's population are with such disabilities and out of which 35 million are children. we have presented a novel framework for continuous SLR using Mediapipe Holistic and LSTM Network [3].

In paper [2], The SLR model was trained employing different batch sizes, starting from the default 32, 64 and 128. The model with batch size of 128 performed better in comparison to the others at 250 epochs. The authors of paper [8] introduced an ensemble technique where we train multiple sub-models and average them. Random Forest algorithm is an example where it uses multiple Decision tree algorithms.

Murakami and Taguchi in the year 1991, published a research article using neural network for the first time in sign language recognition. Mediapipe's state-of-art makes feature extraction easy by breaking down and analyzing complex hand-tracking information, without the need to build a convolutiona neural network from scratch [4].

To bridge the communication gap between the hard-of hearing community and normal people, researchers have proposed a real-time ISL hand gesture recognition system that uses a Microsoft kinetic RGB-D camera for inputting images and applies deep learning techniques to achieve one-to-one mapping between the depth and RGB pixels on training over 45,000 RGB and depth images, while achieving a prediction accuracy of 98.81% [7].

Halder and Tayade [17] used the MediaPipe framework to get multi-hand landmarks and used Support Vector Machine (SVM) for Real-time detection of hand signs. The average accuracy achieved was about 99%. The authors of paper [13] introduced a model which is composed of three layers of modified LSTM units with Rectified Linear Unit (ReLU) activation, followed by a dense layer, and culminating in a SoftMax layer for classifying into multiple categories.In this study, we proposed a MOPGRU model for ISL recognition.

In [10], they used motion trajectory and hand shapes are conventional methodologies for the challenges of DSL recognition. The approach examines the attributes and the characteristics of hand shapes and movement trajectories of hand gestures. In [12], Few researchers have developed tools to convert sign language into text or speech.

Sign language recognition using LSTM and deep learning GRU is a research topic that has received a lot of attention in recent Years.

Using clustering after hand detection, video can be segmented using sign language, and then features such as hand position and orientation and motion trajectory can be extracted. These features can be used as inputs for LSTM and GRU models[12]. In [9], The Convolutional Neural Network (CNN) has been used to extract the spatial features from the signed sequences which are then modelled by the modified LSTM model for recognition. In this section, we have performed an indirect comparative analysis of the proposed SLR framework with some state of the art techniques.

The authors of [11] stated that most automated SLR research is concerned with similar problems, namely the need to interpret hand and body movements associated with sign language characters in a clear and unambiguous manner. Most automated SLR research is concerned with similar problems, namely the need to interpret hand and body movements associated with sign language characters in a clear and unambiguous manner.

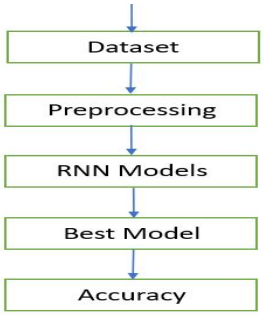
Sahoo [16] used machine learning to work on Indian sign language recognition. This research focused on the static hand gestures in Indian sign language for the numeric values (0-9). A digital RGB sensor was used to capture the images of the signs to build the dataset. The dataset contained 5000 total images, with 500 images for each digit from 0 to 9.

In [9] two classifiers were used based on the supervised learning technique to train the model: Naïve Bayes and kNearest Neighbor. K-Nearest Neighbor technique slightly performed better than the Naïve Bayes classifier in this research as the average accuracy rates of k-NN and Naïve Bayes were 98.36% and 97.79%, respectively.

In [13] Using HMM and 40 signal words, the system was evaluated on 94 and 100 ASL phrases, with accuracy measured at 74.5% and 97.8% for table-based and heat-based systems, respectively.

Das et al. [18] have researched on deep learning-based sign language recognition system with processed static images implemented on American Sign Language gestures. The dataset consisted of 24 labels of static gestures of alphabets from A to Z, excluding J. There were approximately 100 images per class in the dataset captured on an RGB sensor.

III. METHODOLOGIES



Proposed Methodology

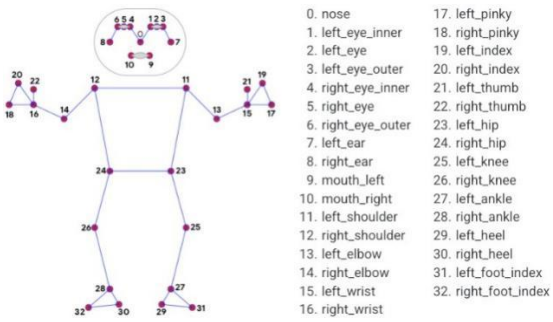
Dataset

We recorded different signs with the help of opencv and we used that data to train both the models LSTM and GRU. Currently this dataset contains 10 signs. We recorded 30 videos per each sign and extracted 30 frames per each video each frame with the resolution 640X480.

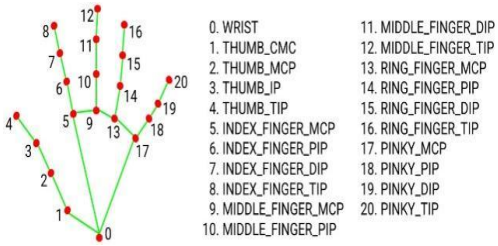
1. MediaPipe Hand Tracking Module

The initial stage of the proposed methodology involves leveraging the robust hand tracking capabilities of MediaPipe. The MediaPipe hand tracking module precisely captures the dynamic features of sign language gestures from video input. This module excels in recognizing hand positions, shapes, and movements, providing a detailed representation of the signer's expressions.

Full body landmarks



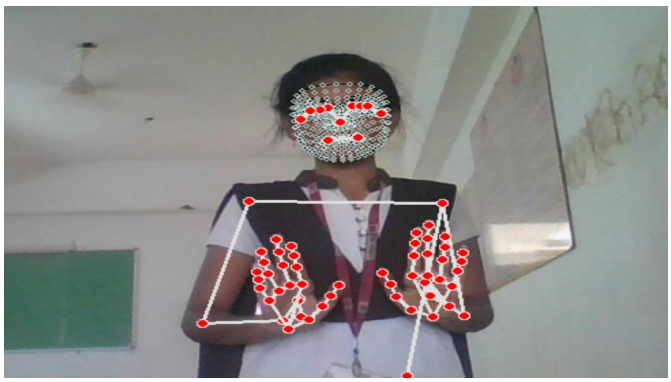
Hand Landmarks



2. Feature Extraction from MediaPipe

Extracted features from the MediaPipe module, including hand landmarks and trajectories, serve as the input for subsequent stages in the sign language recognition pipeline. These features encapsulate spatial information (x, y, z) crucial for recognizing dynamic gestures. Coordinates x and y are for identifying the movement of landmarks and z is the distance of landmarks from the screen. Here are the landmarks we are extracting from videos.

Landmarks identification with MediaPipe



3. Models Used

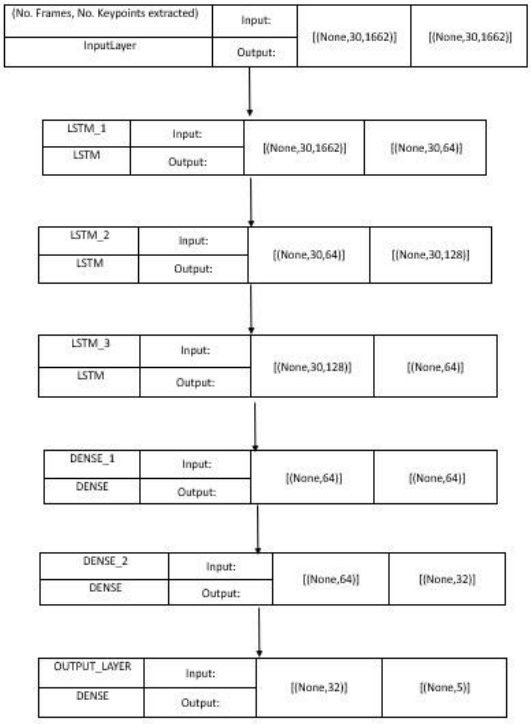
1.Long Short Term Memory (LSTM)
2. Gated Recurrent Unit (GRU)

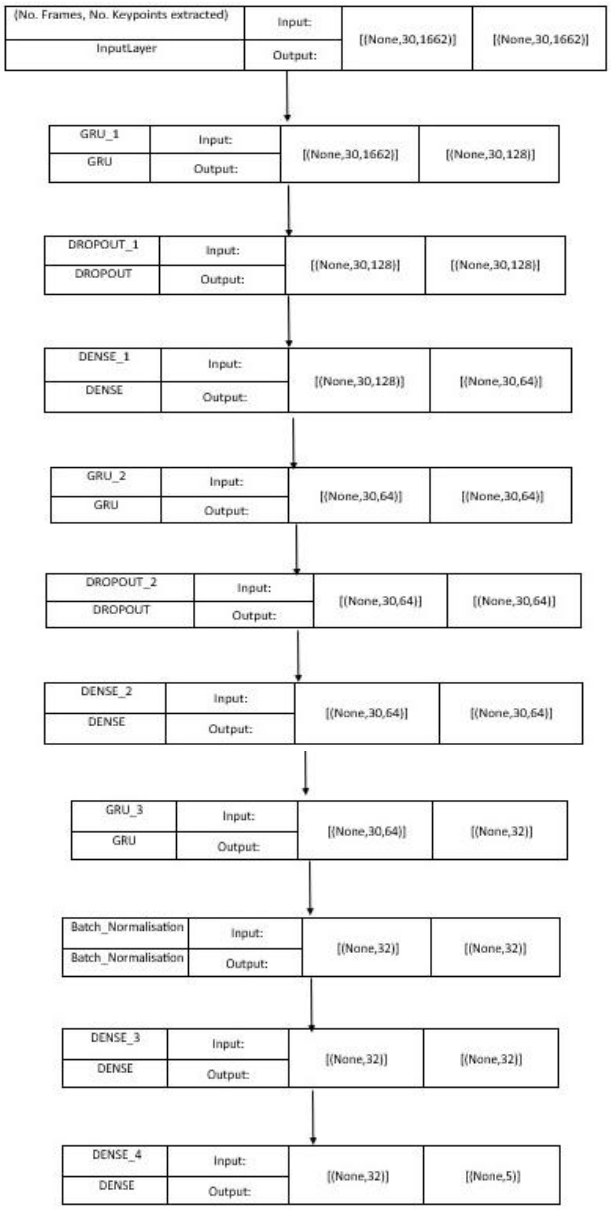
The inputs for the models are keypoints (landmarks) extracted with the help of MediaPipe. The total number of keypoints including with face are 1662 and without face are 258 for each frame. Each sequence length is the frame count of the video. We extracted 30 frames per video.

1. Long Short Term Memory(LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem that commonly occurs in traditional RNNs. The vanishing gradient problem arises when training deep networks, making it challenging for the model to learn long-term dependencies in sequential data.

LSTM Layer:
An LSTM layer is like a special kind of memory chip. It not only processes the current information but also considers what it "remembered" from the previous data. This "memory" is carefully controlled by gates within the LSTM that decide what information to remember, forget, or update. By remembering important parts of the sequence and using them along with the current data, LSTM layers can handle long-term dependencies and make accurate predictions.





2. Gated Recurrent Unit (GRU) :

GRU is one of the RNN networks which is particularly good at understanding sequencial data. It can store important details from the previous parts of the sequence. This model used special gates (update, reset) to control what information flows through the memory. This helps it focus on the relevant parts of the sequence and avoid getting overloaded with data.

Dense Layer:

A dense layer takes all these data points and connects them to every single neuron in the next layer of the network. It's like merging all the separate roads into a big intersection. This allows the network to find complex relationships between the data points that wouldn't be possible if they were kept isolated.

Dropout:

A dropout layer randomly "drops out" a certain percentage of neurons from the neural network during training. This forces the network to rely on different pathways and connections to learn. It's like taking a detour - the network has to find alternative routes to process the information, preventing it from becoming too dependent on any one set of connections.

Batch Normalization:

Batch normalization acts like a coach who helps the team perform consistently. It takes each batch of data the network processes and standardizes it, essentially evening out any extremes. This makes the training process smoother for the network. Think of it like adjusting the difficulty of practice sessions - not too easy, not too hard - so all the neurons learn at a similar pace. By ensuring the data has a consistent scale, batch normalization helps the network converge faster and learn more effectively. It's like creating a stable training environment where the entire team can perform at their best.

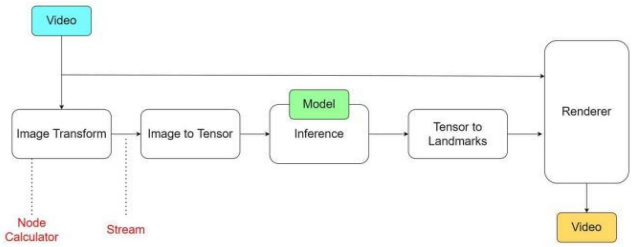
4. Training on Dataset

We recorded different signs with the help of opencv and we used that data to train

both the models LSTM and GRU. Currently this dataset contains 10 signs. We recorded 30 videos per each sign and extracted 30 frames per each video each frame with the resolution 640X480.

5. Integration of models with MediaPipe

The final stage involves the integration of the extracted features from MediaPipe with the LSTM and GRU networks. The fusion of spatial information from hand tracking with the temporal understanding provided by LSTM creates a holistic sign language recognition system. This unified framework aims to achieve a more comprehensive interpretation of sign language expressions, catering to the multifaceted nature of communication in sign language.



IV. IMPLEMENTATION

A diverse dataset is curated, with video sequences annotated to provide ground truth labels. MediaPipe's hand tracking module is then used to annotate hand keypoints in each frame, generating spatial representations for subsequent processing.

1.mediapipe.solutions.drawing_utils.draw_landmarks():

This method helps to visualize the results of running a Mediapipe pose or hand tracking model on that image. This can be done by taking two key inputs:

- The image you want to visualize the landmarks on.
- A list of points (landmarks) detected by the Mediapipe model. These points represent key locations on a face, hand, or body, depending on the model used.
- It draws small circles at each of the landmark points on the image. This makes it easy to see where the model identified important features. Connecting those dots will form a skeleton.

2.mediapipe.solutions.holistic.Holistic():

It combines the skills of three different Mediapipe models:

- Pose detection: This part tracks your body's major joints, like elbows, knees, and hips.
- Face mesh: This one focuses on your face, pinpointing key features like eyes, nose, and mouth.
- Hand tracking: It keeps an eye on your hands, identifying important points like fingertips and palms.

This method is used to extract key features, encompassing hand position, shape, and movement. These features are transformed into spatial-temporal sequences, forming the input for the subsequent LSTM and GRU networks.The LSTM network architecture comprises multiple layers, including LSTM layers for capturing sequential dependencies and dense layers for classification. The model is trained to optimize parameters using our dataset. The GRU network architecture comprises multiple layers, including GRU layers for capturing sequential dependencies, dropout layers for preventing overfitting and dense layers for classification. The model is trained to optimize parameters using our dataset.

In the time of testing our model will detect the landmarks of real-time video and will compare those sequences with the trained landmarks. If there is a match then that sign will be resulted as output text.

These models are for real-time processing, the model will be deployed in future to recognize sign language gestures from live video input. This ensures practical applicability in dynamic environments, facilitating communication in real-time scenarios.

V. RESULTS & DISCUSSION

The version of this We've set up some measurements to figure out how well these programs are doing. One basic measure is accuracy, which tells us if the programs are generally correct in their decisions. We also look at precision and recall, which help us see how accurate the programs are when they say something is an attack or not. We check false alarms (saying there's an attack when there isn't) and true detection rates (spotting actual attacks) to get a better picture. There's a combined measure called the F1 score that gives us an overall idea of how good the programs are at this task.

- True Positive(TP): This appears on the diagonal of the matrix and shows how often the system correctly identified a sign.
- True Negative(TN): Since there's no general "negative" class, True Negatives (TN) wouldn't have a meaningful interpretation in SLR. It wouldn't represent anything relevant in the context of sign recognition.
- False Positive(FP): These occur when the system predicts a sign that wasn't actually used (appears off-diagonal, above the actual sign row).
- False Negative(FN): These occur when the system fails to recognize a sign that was actually used (appears off-diagonal, below the actual sign column).

Accuracy, precision, recall:

The model's performance can be evaluated using the accuracy metric. This measure is determined by dividing the total number of correct predictions by the total number of instances.

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Similarly, precision is used to assess the accuracy of a model's positive predictions. This metric is derived by dividing the number of true positive predictions by the total number of positive predictions made by the model.

$$precison = \frac{(\mathrm{TP})}{(\mathrm{TP+FP})}$$

Lastly, recall is used to measure the effectiveness of a classification model in identifying all relevant instances from a dataset. It is calculated as the ratio of the number of true positive instances to the total number of relevant instances in the dataset.

$$recall = \frac{(\mathrm{TP})}{(\mathrm{TP+FN})}$$

F1-Score:

F1-score gives the overall performance of the classification model.

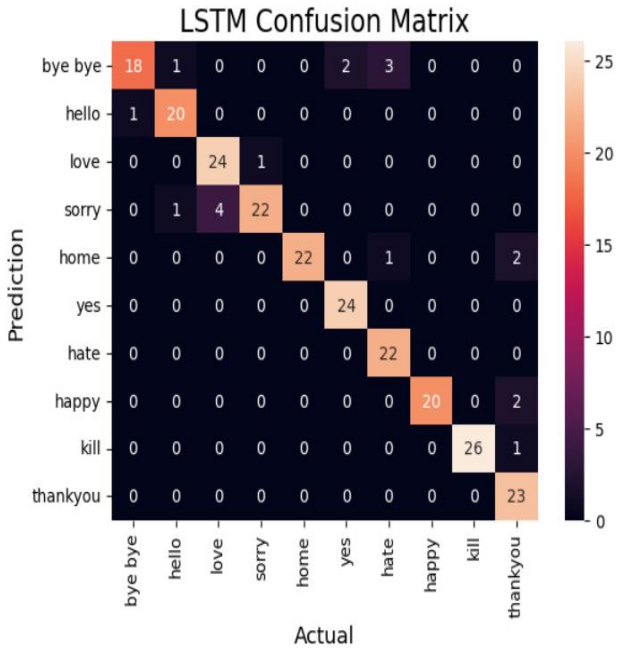
$$f1\_score = \frac{2*(\mathrm{precision*recall})}{(\mathrm{precision+recall})}$$

Classification Report*:*

The Classification_report() method is used to generate the evaluation matrices like Precision,

Recall, and F1-score and heatmap for confusion matrix.The following diagrams shows the confusion matrices and classification reports of LSTM and GRU models.

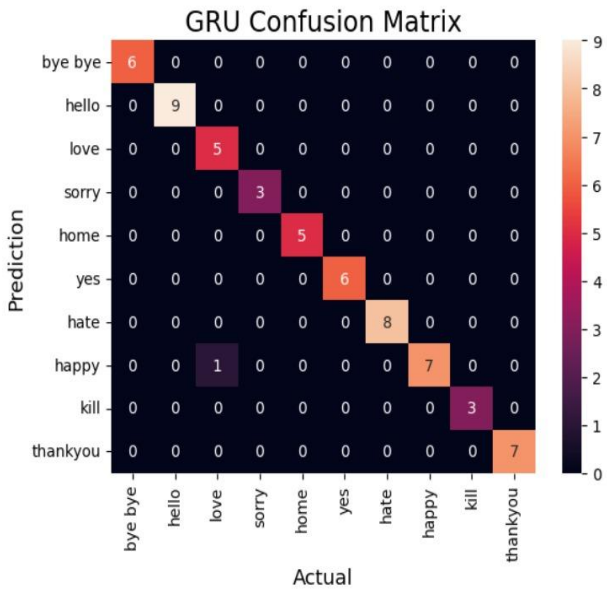
Confusion matrix For LSTM model



Classification Results for LSTM model

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	1.00	1.00	1.00	9
2	1.00	1.00	1.00	5
3	1.00	1.00	1.00	3
4	1.00	0.80	0.89	5
5	0.75	1.00	0.86	6
6	1.00	1.00	1.00	8
7	0.88	0.88	0.88	8
8	1.00	1.00	1.00	3
9	0.78	1.00	0.88	7
accuracy			0.92	60
macro avg	0.94	0.92	0.92	60
weighted avg	0.93	0.92	0.91	60

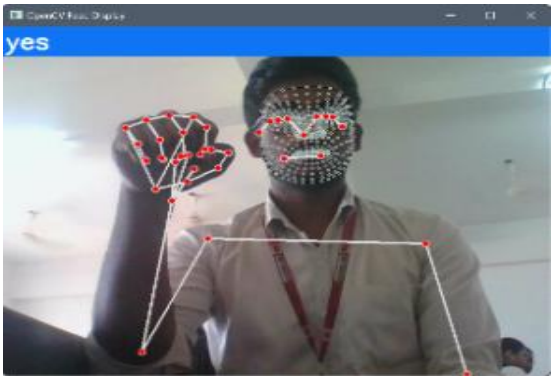
Confusion matrix For GRU model

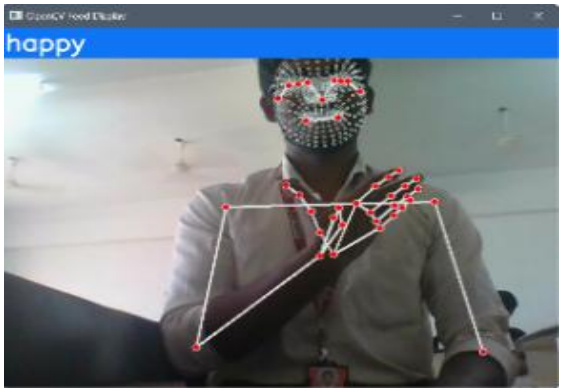


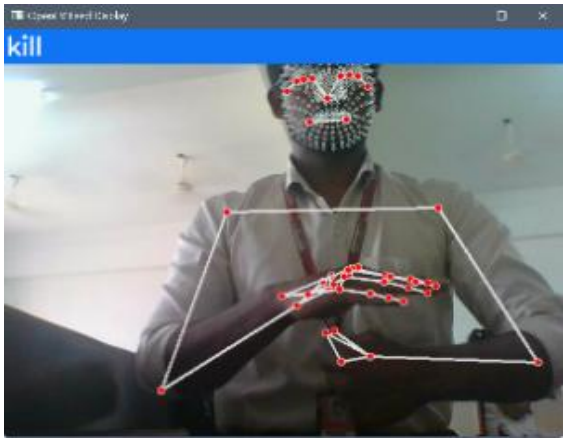
Classification Results for GRU model

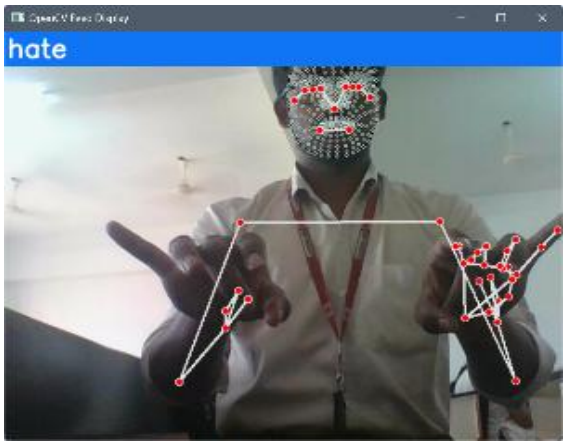
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6
1	1.00	1.00	1.00	9
2	0.83	1.00	0.91	5
3	1.00	1.00	1.00	3
4	1.00	1.00	1.00	5
5	1.00	1.00	1.00	6
6	1.00	1.00	1.00	8
7	1.00	0.88	0.93	8
8	1.00	1.00	1.00	3
9	1.00	1.00	1.00	7
accuracy			0.98	60
macro avg	0.98	0.99	0.98	60
weighted avg	0.99	0.98	0.98	60

Sample Output :









Our tests show that the system combining MediaPipe and GRU is more accurate than MediaPipe with LSTM. It even works well when the lighting changes or the person signing uses different hand movements, showing it's reliable.

In conclusion, the integration of MediaPipe and GRU for sign language recognition offers a powerful solution. The system's accuracy, real-time performance, and adaptability to diverse sign language expressions underscore its potential as a valuable tool for enhancing inclusivity and accessibility

VI. LIMITATIONS

Acknowledging the achievements of the proposed methodology, it is essential to recognize its limitations. The system may encounter challenges in scenarios with rapid hand movements or when presented with signs exhibiting subtle variations. The system is also unable to detect facial expressions accurately. Variability in signing styles among individuals may also pose difficulties.

VII. FUTURE WORK

In upcoming studies, there's an opportunity to delve into incorporating additional elements into sign language recognition. This could involve exploring how facial expressions and body language contribute to a more comprehensive understanding of the context in which sign language is used. Understanding these nuances could significantly enhance the accuracy and richness of sign language interpretation systems, fostering better communication for individuals with hearing impairments. Researchers may also explore user-friendly ways to implement and integrate these additional elements into existing technologies, making them more accessible and effective in real-world scenarios.

VIII. ACKNOWLEDGEMENT

We would like to express our deepest gratitude to all those who contributed to the completion of this article on Sign language recognition using Mediapipe and RNN Models (LSTM and GRU). Special thanks to our mentor for their invaluable guidance and support throughout this project. We are also thankful to the researchers and developers whose work served as the foundation for this study. Additionally, we extend our appreciation to the reviewers for their constructive feedback, which greatly enhanced the quality of this work. Furthermore, we are grateful to the academic institutions and organizations that provided resources and facilities for conducting this research. Finally, we acknowledge the encouragement and understanding of our family and friends, whose

unwavering support kept us motivated during challenging times.

REFERENCES

1. *S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998. Rachana Patil1,∗, Vivek Patil1,∗∗, Abhishek Bahuguna1,∗∗∗, and Mr. Gaurav Datkhile1,∗∗∗∗ "Indian Sign Language Recognition using Convolutional Neural Network", ITM Web of Conferences 40, 03004 (2021)*
2. *Shamitha S H1 , Dr. Badarinath K2, "Sign Language Recognition Utilizing LSTM And Mediapipe For Dynamic Gestures Of ISL" International Journal for Multidisciplinary Research (IJFMR) Volume 5, Issue 5, September-October 2023.*
3. *Madhura Mirikar*, Komal Singh*, Prof. Dr. Sampada Dhole** "Continuous Sign Language Recognition Using LSTM and Media Pipe Holistic", International Journal of Scientific Research and Engineering Development, Volume 6 Issue 5, Sep- Oct 2023.*
4. *Arpita Haldera , Akshit Tayadeb, "Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning", Vol (2) Issue (5) (2021).*
5. *Ashish B Deharkar "An Approach To Reducing Cloud Cost And Bandwidth Using Tre System", Journal Name, Vol. 11, Issue 3, March 2022.*
6. *Jyotishman Bora, Saine Dehingia, Abhijit Boruah* Anuraag Anuj Chetia, Dikhit Gogoi, "Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning", International Conference on Machine Learning and Data Engineering, Procedia Computer Science 218 (2023).*
7. *Barathi Subramanian1 , Bekhzod Olimov1 , Shraddha M. Naik1 , Sangchul Kim2 , Kil Houm Park3 & Jeonghong Kim1 "An integrated mediapipe optimized GRU model for Indian sign language recognition", Scientific Reports | (2022) 12:11964.*
8. *1Satwik Ram Kodandaram, 2 N Pavan Kumar 3 Sunil G L, "Sign Language Recognition",ResearchGate,DOI: 10.13140/RG.2.2.29061.47845,July 2021.*
9. *Anshul Mittal, Pradeep Kumar, Partha Pratim Roy, Raman Balasubramanian and Bidyut B. Chaudhuri "A Modified-LSTM Model for Continuous Sign Language Recognition using Leap motion", 1558-1748 (c) 2018 IEEE.A.O . Akinwumi , A.O.Akingbesote , O.O.Ajayi,F.O.Aranuwa,"DETECTION OF DISTRIBUTED DENIAL OF SERVICE (DDOS) ATTACKS USING CONVOLUTIONAL NEURAL NETWORKS" in 'Nigerian Journal of Technology (NIJOTECH) ' , vol.41 , Issue-06, November 2022*
10. *Gerges H. Samaan , Abanoub R. Wadie , Abanoub K. Attia , Abanoub M. Asaad , Andrew E. Kamel , Salwa O. Slim , Mohamed S. Abdallah 2,3,* and Young-Im Cho 2,*, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition", Electronics 2022, 11, 3228.*
11. *MUHAMMADAL-QURISHI, (Member, IEEE), THARIQ KHALID, AND RIAD SOUISSI. "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues", IEEE Access, Vol. 6, 2021.*
12. *Pranav Sheth, Sanju Rajora, Sanju Rajora. Sign Language Recognition Application Using LSTM and GRU (RNN).ResearchGate,April 2023.*
13. *Ridwang a,*, Amil Ahmad Ilham b, Ingrid Nurtanio b , Syafaruddin c , Dynamic Sign Language Recognition Using Mediapipe Library and Modified LSTM Method, IJASEIT, Vol.13 (2023) No.6.*
14. *Ahmed Elgohary, Rawan Galal Elrayes (2021) "Egyption Sign Language Recognition Using CNN and LSTM" Computer Vision and Pattern Recognition. https//doi.org/10.48550/arXiv.2017.136.*
15. *Zhibo Wang, Tengda Zhao, Jinxin Ma,Hongkai Chen,Kaixin Liu,Huajie Shao,Qian Wang,Ju Ren,"Hear Sign Language: A Real-time End-to-End Sign Language Recognition System", IEEE , December 20,2020.M. Devendra Prasad. Prasanta Babu V, C Amaranth "Machine Learning DDoS Detection Using Stochastic Gradient Boosting " in 'JCSE International Journal of Computer Sciences and Engineering' , Vol.7, Issue - 4, April 2019*
16. *A. K. Sahoo. (2021) "Indian sign language recognition using machine learning techniques," in Macromolecular Symposia.*
17. *A. Halder and A. Tayad. (2021) "Real-time vernacular sign language recognition using mediapipe and machine learning," Journal homepage:www. ijrpr. com ISSN, vol. 2582, p. 7421.*
18. *A. Das, S. Gawde, K. Suratwala and D. Kalbande. (2018) "Sign language recognition using deep learning on custom processed static gesture images," in International Conference on Smart City and Emerging Technology (ICSCET).*