[1]Guruprakash B,

[2]Nagarajan Gurusamy,

[3]Ramnath M,

[4] Sumathi S,

[5] Mariappan E,

[6] Saravanan T

ISL Sign Language Recognition Using LSTM-Driven Deep Learning Model

Abstract: - In the development of the "ISL Sign Language Recognition Using LSTM-Driven Deep Learning Model," our methodology integrates advanced computer vision and deep learning techniques to achieve high accuracy in gesture recognition. The system leverages MediaPipe, an open-source library developed by Google, which facilitates real-time hand tracking and gesture detection. MediaPipe provides a reliable framework for extracting precise spatial information from video frames, identifying key landmarks on the hands, and distinguishing complex movements associated with Indian Sign Language (ISL). This pre-processing step ensures that the raw video data is converted into structured data, capturing the essential features necessary for gesture recognition. The processed data is then fed into a Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN) well-suited for sequential data. LSTM networks are designed to address the vanishing gradient problem common in traditional RNNs, making them ideal for modeling temporal dependencies. By using the LSTM's unique memory cell structure, the system effectively retains information over longer sequences, which is critical for understanding the continuity and nuances of hand movements in sign language. The training phase employed a dataset of commonly used ISL signs. Each sign was captured in multiple video samples to ensure variability and robustness in the model. The LSTM model was trained to classify the sequential hand gestures, adjusting weights through backpropagation and optimizing the network using stochastic gradient descent. During training, data augmentation techniques were applied to enhance the model's ability to generalize, preventing overfitting and increasing accuracy. The model's performance was evaluated through a series of tests, measuring precision, recall, and overall accuracy. The results demonstrated that the LSTM-driven model outperforms traditional methods by effectively capturing and interpreting intricate hand movements, achieving superior recognition rates. This research contributes to assistive technology by improving the accessibility of communication for individuals with hearing impairments and promoting inclusivity in society.

Keywords: Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Indian Sign Language (ISL), *MediaPipe Holistic Key Points, OpenCV.*

I. INTRODUCTION

It can be very challenging to communicate with those who are deaf. When deaf and mute people utilize hand motions in sign language, it can be challenging for non-disabled people to understand what they are saying. Systems that can recognize different symptoms and explain them to regular people are therefore necessary. For the deaf and dumb, it is essential to develop specialized sign language applications so that they might converse with those who are unable to understand gesture language with ease.

The primary aim of our research is to lower the transmission gap that exists between able-bodied and deaf or dumb gesture language users. Creating a vision-based system is our primary goal [2] that can identify gestures and motions in image or video referrals. By using the models, both geographical and temporal features have been learned. With the help of spatial data from the video series, the LSTM model of the RNN was trained. With good performance, the suggested method offers an effective means of translating gesture language into text so that the users get an idea

[1]*Corresponding author: Department of Artificial intelligence and Machine Learning, Sethu Institute of Technology, Tamil Nadu, India.

[2] Author: Department of Artificial Intelligence and Data Science, Shri Shanmugha College of Engineering and Technology, Salem, Tamil Nadu, India.

[3,5] Author: Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India.

[4,6] Author: Department of Information Technology, P.S.R Engineering College, Tamil Nadu, India.

about the gesture. The technology has a wide range of uses, such as teaching young children how to use computers using sign language comprehension.

A. Motivation and Background

To reliably recognize and understand the presented symbol sequences, algorithms and methods for sign language recognition are being developed. Lots of SLR strategies treat the issue like they would a gesture recognition (GR) challenge. To correctly classify a specific signal from a collection of likely signs, research has therefore focused on identifying favorable attributes and differentiation strategies. Conversely, sign language consists of more than just a set of precise movements.

B. What is gesture recognition?

Within the fields of computer science and language technology, gesture recognition involves employing mathematical algorithms to interpret a person's touch. A subfield of computer science is computer vision. Although gestures can originate from any bodily movement or position, the most common places to observe them are on the hands and face. In the field, facial and hand touch detection for emotional recognition is currently very popular. With basic touch, users may operate or communicate with devices without coming into contact with them. Numerous techniques that employ cameras and computer vision algorithms to translate the signal language have been developed [2].

C. Sign Language

Sign languages, also referred to as sign language, are visual languages where meaning is expressed by cues. In sign language, both non-sign language and sign language objects are expressed. With their syntax and lexicon, sign languages [13] are completely functional natural languages. Despite several striking similarities between sign languages, these similarities are neither well-established nor ubiquitous. Linguists believe that both spoken and signed communication are natural forms of language [8], suggesting that they both developed into a fuzzy aging process that lasted longer and changed over time without planning. Sign language should not be confused with body language, which is a kind of nonverbal communication.

II. Literature Survey

Translation of gestural language into text is one of the main responsibilities of advanced computer activities. SLR aims to develop algorithms that facilitate the interpretation of sign language, which will be especially useful for those who utilize gesture language as a means of communication with the disabled. [1] MediaPipe with perceptual computing capabilities, including computer vision and machine learning. One of the features offered by MediaPipe is hand tracking and gesture recognition, which is used for sign language recognition. [2] To recognize sign language, a design module or algorithm that combines the advanced computing tasks of a computer is used. Without a speech interface, it can be extended to include interaction between humans and computers. This system is part of the transdisciplinary content, and the method is part of the gesture recognition. [3] The model presented in this research uses deep learning to recognize and discern words in a person's motions. Dynamic sign language uses hand and body gestures to represent its vocabulary. This method tackles dynamic sign language detection issues by combining media pipes and RNN models. Using Media Pipe, we were able to eliminate crucial hands, bodies, and facial features to assess the objects' position, shape, and orientation [4].

To recognize sign language using machine learning, models must be trained to decipher and comprehend sign language motions. Space invariant artificial neural networks, which are used for graphical problems, are employed in the research. [5] Using 3D pictures, 3DCNN presents a gesture detection approach that detects the forms and direction of the hand palm. This work contributes to the investigation of image processing methods for sign language interpretation. [6] Shedding light on the saliency-based method of gesture language detection. The study contributes to our understanding of sophisticated computer and knowledge engineering methods for sign language interpretation by concentrating on linear feature extraction for the identification of the alphabet and numbers in Indian Sign Language [7].

Automatic Sign Language Detection: By offering a resource for researchers and practitioners interested in developing and evaluating sign language detection algorithms, [8] A well-known open-source computer vision and image processing library is called the Computer Vision Library. For tasks like object identification, face recognition,

image and video processing, and more, it offers a broad range of tools and functionalities. Additionally, Google created the open-source MediaPipe framework, which offers a number of pre-made solutions for different computer vision workloads. Deep learning and these kinds of novel techniques are applied to projects involving the detection of sign language [9]. First studies on hand gestures for letter recognition: The first studies on hand gestures for letter recognition offer new perspectives on computational science and gesture-based communication. is the recognition of sign language using hand gestures, MediaPipe, and OpenCV [10]. Gesture Recognition using OpenCV: presenting a sign language recognition system utilizing Kinect technology, with a focus on interpreting both numbers and the English alphabet. This paper provides insights into the intersection of machine learning, computer vision, and sign language communication [11].

Gesture Recognition using OpenCV: presenting a sign language recognition system utilizing Kinect technology, with a focus on interpreting both numbers and the English alphabet. This research sheds light on the relationship between sign language communication, computer vision, and machine learning. [12] Introducing a real-time method for 2D hand identification and tracking for sign language recognition. Because it provides insights into real-time applications for sign language interpretation, this work is important for investigating methods in computer vision and system engineering [13.] Recognition of sign language by machine learning. In order to overcome the constraints of KNN approaches, which include large dataset sizes, noise, outliers, class imbalances, and dimensionalities, we employ CNN. It lacks any facilities [14].

Automatic Sign Language Detection: By offering a resource for researchers and practitioners interested in developing and evaluating sign language detection algorithm [15].

III. Methodology and Framework development

The system combines the concept of long short-term memory networks (LSTM) with the RNN neural network to convey the sequences. Although several systems for the translation of gesture language have been developed, their ability to recognize specific sign motions is restricted. Our research proposes a modified long short-term memory model for dynamic gestures that is capable of recognizing a set of related gestures. Long short-term memory (LSTM) networks have been investigated and applied to gesture data classification due to their ability to make gesture associations. The architecture that was created showed promise for LSTM-based neural networks in the conversion of gesture language, with an accuracy of 90% [8].

3.1 Data Collection

A dataset [16] contains the motions and signals. The video camera will be streamed live, and for each frame that detects a movement or gesture inside the designated ROI (region of interest), the information is recorded in a gesture gallery. In the collection, each sign is portrayed by over 30 different video sequences. Every video clip has 30 frames that capture key moments and are saved as Numpy arrays [8]. Throughout the video sequence, the gesture that is being signed must be recognized. Using the media pipe to merge the pose, hand, and face, they are designed for a certain platform. The model that is running processes its input as a 256x256 video frame with a predetermined quality.

1. The hands are equipped with 21 landmarks each.

2. The overall number of Pose Landmarks is 33.

3. Facial Landmarks: There are 468 landmarks in total.

Hand landmarks

Understanding the form and movement of hands can improve the user experience across a variety of technical platforms and professions. A high-precision approach for tracking hands and fingers is Media Pipe Hands. Unlike existing state-of-the-art approaches, which often necessitate powerful desktop workstations for inference, our technology provides many hands-on, real-time performances.

3.2 Methodology

In this research on sign language identification, we develop a sign detector that can recognize custom signs and is easily expandable to recognize a large number of other signs and hand gestures, such as the alphabet and numerals. This project was made using the Python modules OpenCV, MediaPipe [2], TensorFlow, and Keras. To identify the

action of the person who is now being exhibited, the OpenCV feed analyzes the frames of live video captured by a camera [8]. Media Pipe Holistic is used to process the video frames to extract key points from our hands, torso, and face. The prediction algorithm will then receive the pertinent points and start the prediction. Then, in real-time, the system predicts the hand sign being made.

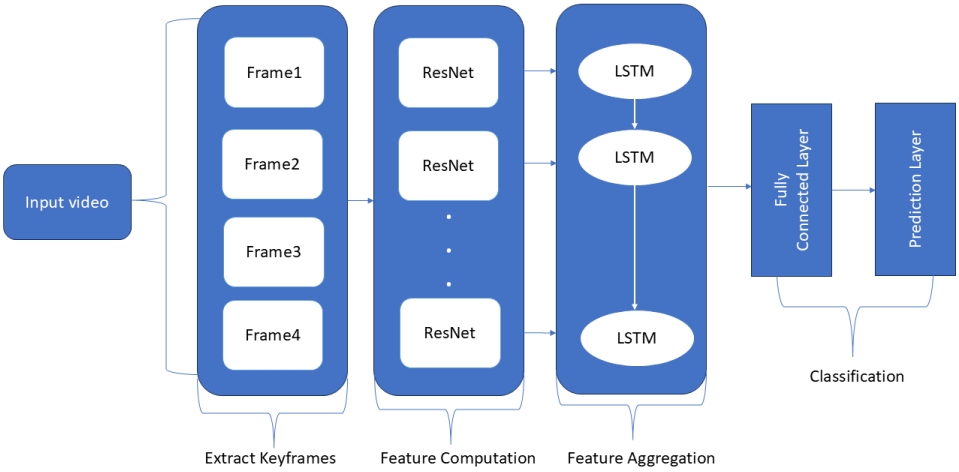


Fig. 1: Proposed System Architecture

In the proposed system architecture depicts the deep learning model for ISL (Indian Sign Language) recognition, combining convolutional and recurrent neural network techniques. Here's a breakdown of each stage in the model pipeline.

3.2.1 Input Video

The input video has been accessed by webcam and its process starts with an input video that contains sequences of hand gestures representing a specific ISL sign.

3.2.2 Extract Keyframes

The video is segmented into keyframes. These frames are critical snapshots that capture the hand positions and movements at different time intervals to represent the full sequence of gestures.

3.2.3 Feature Computation (ResNet)

Each frame is passed through a pre-trained *ResNet* (Residual Network) model to extract spatial features. *ResNet* is a powerful convolutional neural network (CNN) architecture that captures rich, hierarchical image features while addressing the vanishing gradient problem. This step ensures that each frame is represented by a feature vector that encodes meaningful visual information about the hand's position and shape.

3.2.4 Feature Aggregation (LSTM)

The extracted feature vectors from ResNet are sequentially fed into an *LSTM (Long Short-Term Memory) network*. LSTMs are designed to handle sequential data, allowing the model to understand the temporal dependencies between frames. This stage aggregates the features across the time dimension, capturing how the hand movements change throughout the video, which is essential for recognizing gestures.

3.2.5 Fully Connected Layer

The output of the LSTM is passed to a fully connected layer, which acts as a dense neural network layer. This layer processes the aggregated features and prepares them for classification.

3.2.6 Prediction Layer

The final output is generated through the prediction layer, which classifies the input video into the appropriate ISL sign based on the learned features. This step outputs the predicted label corresponding to the recognized sign language gesture.

The pipeline begins with the extraction of frames from an input video, followed by feature extraction using ResNet. These features are sequentially processed by the LSTM network, which captures the temporal relationships between the frames. The aggregated information then passes through a fully connected layer, and the final prediction layer outputs the recognized ISL sign. This approach effectively combines spatial and temporal data for robust sign language recognition.

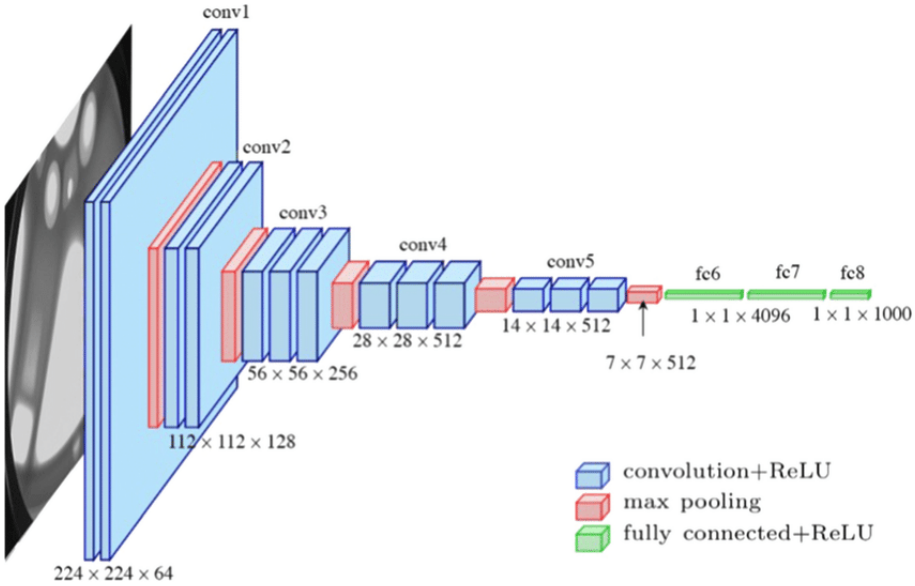


Fig. 2: VGG-16 Network Architecture Overview for Deep Feature Extraction

The provided image represents the architecture of the VGG-16 convolutional neural network (CNN), a widely used deep learning model known for its strong performance in image classification and feature extraction. Here is a breakdown of the different layers and their roles:

i. Input Layer (224 × 224 × 64)

The network takes an image input of size 224 × 224 pixels with three colour channels (RGB). The image is processed through an initial set of filters to start feature extraction.

ii. Convolutional Layers (conv1 to conv5)

conv1: The first set of convolutional layers applies filters to the input image, creating feature maps. Each convolutional layer uses a small 3 × 3 receptive field and a stride of 1, followed by a ReLU activation function to introduce non-linearity. **conv2 to conv5**: The subsequent layers deepen the network, progressively increasing the number of feature maps (e.g., 128, 256, and 512) and enhancing the network's ability to learn complex patterns. The image's spatial dimensions decrease while the depth (number of feature maps) increases.

iii. Max Pooling Layers

The red blocks represent max pooling layers, which follow specific convolutional layers. These layers down sample the spatial dimensions of feature maps (e.g., from 112 × 112 to 56 × 56) by selecting the maximum value in a 2 × 2 window. This reduces computational complexity and provides spatial invariance.

iv. Fully Connected Layers (fc6, fc7, and fc8)

The green blocks indicate fully connected (FC) layers that act as dense layers with ReLU activation. fc6 and fc7: Each has 4096 units, contributing to high-level feature representation. fc8: The final FC layer has 1000 units, corresponding to the number of output classes in the original VGG-16 used for ImageNet classification. It outputs the probability distribution over the classes using a softmax function.

The above fig. 2 architecture is notable for using small convolutional filters (3 × 3) throughout, which enhances feature learning while maintaining computational efficiency. The depth of VGG-16 (16 layers) allows it to capture intricate features at different levels of abstraction, making it highly effective for tasks like image classification and feature extraction in transfer learning.

3.3 Training

TensorFlow is a machine learning platform that is used for Python-based training. To apply transfer learning, the datasets and label files must be transformed into a format that TensorFlow can process. To create tfrecord files, use the training and testing data directories. To begin using transfer learning to train a network, the object detection models need to be downloaded. There must be considerable adjustments made to the configuration files to match the number of classes in the dataset. It takes 2000 training steps to reach high accuracy. An LSTM model made possible by TensorFlow and Keras can predict screen activity, in this case, a sign language motion.

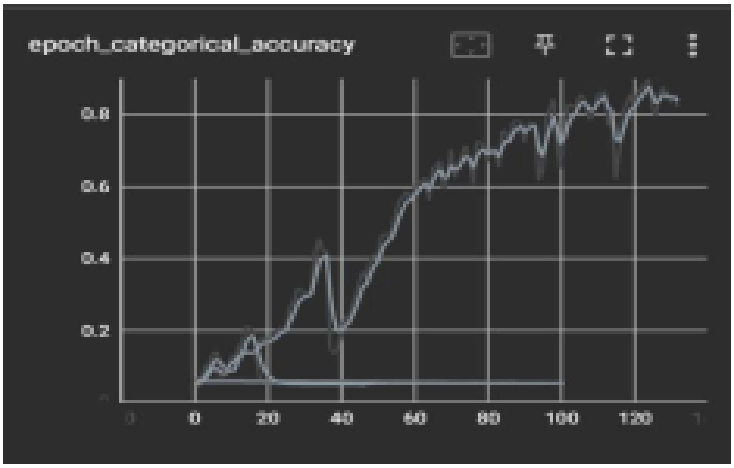


Fig. 3: Accuracy during the training and validation phases

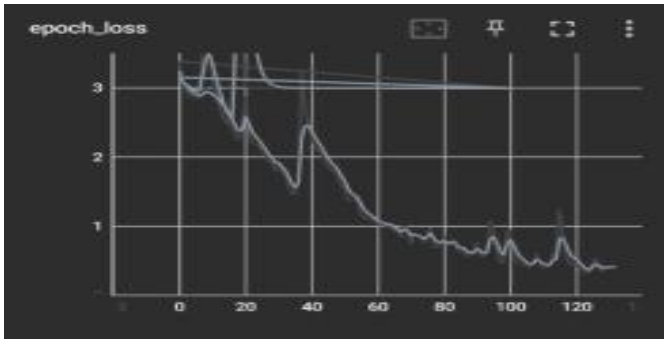


Fig. 4: Epoch loss during the training and validation phases

IV. RESULTS AND DISCUSSION

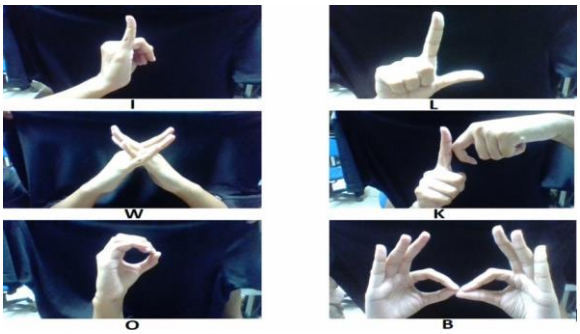
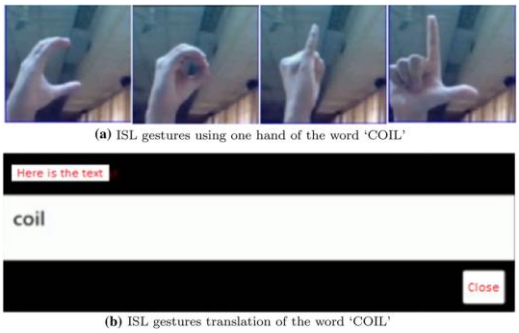


Fig. 5: During the training phases



(a) ISL gestures using one hand of the word 'COIL'

(b) ISL gestures translation of the word 'COIL'

Fig. 6: During the ISL Gesture validation phases

The system appears to be correctly identified each hand gesture and translate the sequence into the English word "coil." This demonstration shows the effectiveness of the gesture recognition algorithm used during the training. If relevant, quantitative metrics such as accuracy, precision, recall, and F1-score can be mentioned to validate the reliability of the model in recognizing the ISL alphabet and forming words and shown below in fig. 7 and fig. 8. Our system can significantly aid communication for individuals who are deaf or hard of hearing, especially in regions where ISL is commonly used. The translation result shown in a UI format suggests that the model's integration within a user-friendly interface is aimed at real-time gesture-to-text conversion, enhancing accessibility. Factors such as varying hand sizes, angles, and lighting conditions might affect recognition accuracy. Our Future work could address these challenges by training the model on a more diverse dataset.

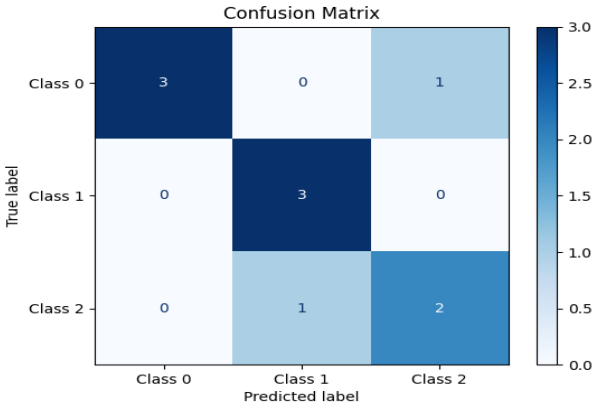


Fig. 7. Confusion Matrix

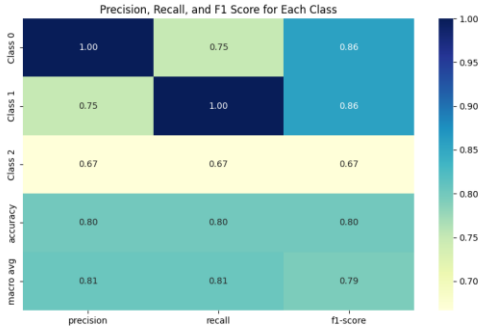


Fig. 8. Precision, Recall and F1 Score

V. CONCLUSION

With a plethora of uses in human-computer interaction, hand gestures are a powerful communication tool. There are several well-established advantages to the vision-based method of hand motion identification. Videos are difficult to analyze since they contain both temporal and spatial data. Using models, we have classified based on both historical and current geographic data. LSTM was used to classify the data based on these two features.

Understanding sign language may be a demanding task. It would be more advantageous, though, to divide this problem into smaller ones, with the method shown here serving as a potential fix for one of them. It was shown that the model frequently incorporated several signs, including W and U. After some consideration, though, it might not have to be flawless because the translation will be more accurate when orthographic correctors or word predictors are used. Examining the reaction and looking for ways to improve the system are the next steps. Range Prospective A model for recognizing words and sentences in a single language can be constructed. Applying a system that can recognize changes in temporal space will be necessary for this. We can close the communication gap for those who are deaf or hard of hearing by developing an inclusive solution. Gesture communication will become a two-way process and replace regular communication in order to improve the system's vital processing components and meet the fundamental criteria. We'll look for any hints about motion. The conversion of the movement pattern into text, words, or sentences and back again into audible speech will also be a focus. Analyzing the reaction and searching for ways to improve the system is the next step.

VI. FUTURE SCOPE

A model can be developed for the identification of words and sentences in a single language. This will necessitate the use of a system that can recognize temporal space changes. For people who are hard of hearing or deaf, we can close the communication gap by developing a comprehensive service. To convert the normal mode of communication into gesture communication and making it a two-way process to establish, the crucial processing parts of the systems needs an updating to meet the necessary requirements. We will look for any hints that pertain to motion. We'll also be focusing on translating the movement pattern into words, sentences, or text, and back again into audible speech.

ACKNOWLEDGMENT

	Guru Prakash.B (Guru Prakash Baskaran) is an Associate Professor of Computer Science and Engineering (Department of Artificial intelligence and Machine Learning) in Sethu Institute of Technology, Virudhunagar, India. He received his B.E degree in Computer Science and Engineering from Madurai Kamaraj University, Tamilnadu, India, in 2003. He has received M.E degree in Network Engineering from Anna University, Tamilnadu, India, in 2005. He has received his Ph.D Degree from Anna University Chennai, Tamilnadu, India, in 2020. His research interests include Multicasting and Wireless Sensor Networks.
	Ramnath M received B.E. degree in Information Technology from Francis Xavier Engineering College, Tamil Nadu, India in 2009 and M.E. degree in Network Engineering from Vel Tech Multi Tech Engineering College, Tamil Nadu, India in 2011. He is currently pursuing Ph.D. at the Department of Information and Communication Engineering, Anna University, Chennai, India. He is working as an Assistant professor in Department of Artificial Intelligence and Data Science at Ramco Institute of Technology, Tamil Nadu, India. He is a life time member of International Association of Engineers (IAENG) and Institute For Educational Research and Publication (FERP).

REFERENCES

[1] Aditi Deshpande; Ansh Shriwas; Vaishnavi Deshmukh; Shubhangi Kale published in 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)

[2] Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning by Panel Jyotishman Bora, Saine Dehingia, Abhijit Boruah, Anuraag Anuj Chetia, Dikshit Gogoi 2023

[3] SIGN FORMER: DeepVision Transformer for Sign Language Recognition by Deep R. Kothadiya; Chintan M. Bhatt; Tanzila Saba; Amjad Rehman (IEEE) 2023

[4] Real-Time Sign Language Recognition System by Sanjukta Sen; Shreya Narang; P. Gouthaman(IEEE) 2023

[5] A.K. Sahoo Indian sign language recognition using machine learning techniques by Macromolecular Symposia (2021)

[6] Muneer AI-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsuaiman, Mohamed A Bencherif, Mohamed Amine Mekhtiche “Hand Gesture Recognition for sign language using 3DCNN” 2020(IEEE)

[7] Real-Time Recognition of Indian Sign Language by H Muthu Mariappan; V Gomathi. Published in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS)

[8] Indian Sign Language Character Recognition by Piyush Mohan, Tanya Sabarwal, T. Preethiya Published in 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (IEEE)2023

[9] Realtime sign Language Detection and Recognition by Aakash Deep, Aashutosh Litoriya, Akshay Ingole, Vaibhav Asare, Shubham Mbhole, Shantanu Pathak published in 2022 2nd Asian Conference on Innovation in Technology (ASIANCON)

[10] S. K. A, S. Kesavan, J. J, A. K. K. S and P. Maddula, "Voice Enabled Deep Learning Based Image Captioning Solution for Guided Navigation," 2023 International conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-6

[11] W. Y. Yeh, T. H. Tseng, J. W. Hsieh and C. M. Tsai, "Sign language recognition system via Kinect: Number and English alphabet," 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, 2016, pp. 660-665. doi 10.1109/ICMLC.2016.7872966

[12] L. K. Phadtare, R. S. Kushalnagar and N. D. Cahill, "Detecting hand palm orientation and hand shapes for sign language gesture recognition using 3D images," 2012 Western New York Image Processing Workshop, New York, NY, 10.1109/WNYIPW.2012.6466652

[13] S. Wu and H. Nagahashi, "Real-time 2D hands detection and tracking for sign language recognition," 2013 8th International Conference on System of Systems Engineering, Maui, HI, 2013, pp. 40-45. doi: 10.1109/SYSoSE.2013.6575240R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[14] N Rajasekhar; M Geetha Yadav; Charitha Vedantam;Karthik Pellakuru; Chaitanya Navapete " Sign Language Recognition using Machine Learning Algorithm" published in: 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)

[15] "Indian Sign Language Recognition System for Dynamic Signs" published in: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) by Arun Singh, Ankita Wadhawan, Manik Rakhra, Usha Mittal, Ahmed AI Ahdal, Shambhu Kumar Jha.

[16] M, RAMNATH (2024), "Indian Sign Language_Dataset", Mendeley Data, V1, doi: 10.17632/yx7kdssfjp.1