2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) | 979-8-3503-4314-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSITSS60515.2023.10334218

Hand Gesture Recognition using MediaPipe and CNN for Indian Sign Language and Conversion to speech format for Indian Regional Languages

Shivani Deshpande
Department of Computer Science and Engineering,
R V College of Engineering,
Bengaluru, India
shivanid.scs21@rvce.edu.in

Dr. Rajashree Shettar
Professor and Dean (PG and Research) Circuit Programs,
R V College of Engineering,
Bengaluru, India
rajashreeshettar@rvce.edu.in

***Abstract*— In recent times, deep learning techniques have made remarkable progress across various domains and applications. However, Indian Sign Language (ISL) recognition, audio generation translation, still present significant challenges from a developmental perspective. In this paper, introduction of a novel method to create a broad framework for real-time ISL recognition, translation tasks. To enhance recognition accuracy, leveraged the power of the MediaPipe library and employ a hybrid Convolutional Neural Network model for extracting pose details and generating text. On the other hand, for producing sign gesture audio corresponding to spoken sentences, used GTTs, supporting the conversion of the audio into various Indian Regional Languages like Hindi, Kannada, Telugu, Marathi, etc. The proposed mechanism tackles the complexities present in earlier approaches and achieves an impressive accuracy of about 94%. Extensively tested the mechanism during its development phases, and evaluation metrics demonstrate convincing improvements over previous methods. In conclusion, this novel approach, combining MediaPipe, CNN, has significantly improved the recognition and production of sign language gestures. The model's performance on evaluation metrics showcases its potential to enhance communication and accessibility for individuals with hearing and speech impairments.**

Keywords*— Indian sign language, ISL, CNN, gesture sign, text-to-speech sign language method, MediaPipe Holistic, sign language recognition, SLR, MT, GTTS, ISL-CSLRT.*

I. INTRODUCTION

Effective communication is a fundamental aspect of human life, enabling individuals to express their needs and engage with others. Latest research has illustrated the fascinating and distinct realm of sign language utilized across different countries. Sign languages rely on visual cues and skillfully coordinate manual and non-manual components, offering invaluable support to the hard-of-hearing and speech-impaired communities, empowering them to access education, employment, and societal rights. In recognition of its significance, several governments have amended legislation to standardize sign language, benefitting those with hearing and speech impairments. However, understanding and responding to sign language require additional training and knowledge for the general population, leading to a communication gap between the two communities. Fortunately, the latest advancements in neural network techniques have affirmed effective in bridging this gap, incorporating diverse mechanisms and mathematical method. Nonetheless, designing such systems entails addressing significant complexities throughout various phases, including misclassification, self-occlusion, movement epenthesis, ambiguity, noise, and blurred output. This research delves into these challenges, pioneering novel solutions to create a powerful architecture that delivers enhanced performance in sign language recognition communication. Introduced a groundbreaking approach, the MediaPipe Holistic combined with Convolutional Neural Networks (CNN) model, which seamlessly integrates sign language recognition, audio generation and translation to various Indian Regional languages tasks within a unified application, is specifically designed to learn diverse modalities of sign gestures in a signer-independent environment, enabling the mechanism to comprehend intricate relationships between input and output. The empirical results demonstrate the remarkable effectiveness of this proposed work, both in terms of recognition accuracy and visual quality.[1,2]

This approach encompasses the recognition of sign language and the conversion of text into various Regional Indian Languages such as Hindi, Telugu, Kannada, Marathi, and more. This conversion is facilitated through the GTTS module, as it acknowledges that many sign language users may struggle with textual communication. This strategy simplifies communication for individuals who are deaf, hard of hearing, or unfamiliar with sign language. It establishes a robust framework to facilitate effective communication for both the deaf and hard-of-hearing communities, emphasizing its importance.[3]

This article introduces an innovative method for Hand Gesture Recognition that leverages both the advanced capabilities of the MediaPipe library and the effectiveness of CNNs. The system put forth in this study not only achieves real-time identification of hand gestures but also translates these recognized gestures into speech to facilitate Indian Sign Language (ISL) communication. By doing so, the system aims to provide a seamless and inclusive communication experience for the hearing-impaired individuals, allowing them to express themselves more freely and interact with others effortlessly.

In this paper, the architecture is designed for the Hand Gesture Recognition for ISL, detailing the integration of MediaPipe and CNNs for gesture prediction. The paper also displays the preprocessing steps, dataset collection, mechanism training, and the conversion of gestures to speech format. Furthermore, empirical results are presented that demonstrate the system's effectiveness in real-time gesture recognition and speech conversion and assess its accuracy and performance.

II PREVIOUS WORK

The authors of the paper [1] propose a system that utilizes object recognition mechanism to help visually impaired the ones identify objects. Use of a combination of computer vision algorithms and machine learning process to achieve this. Additional to object recognition, the system also generates speech feedback to provide audio information about the object. This enables visually impaired the ones to understand the object's properties and characteristics without needing to see it. The system is implemented as a mobile application that can be installed on a smartphone. This makes it convenient for visually impaired the ones to carry with them and use in a variety of settings. The authors assess the final output of the system using a dataset of 40 images. They report a correctness rate of 93% for object recognition.

The paper [2] presents an overview of the state-of-the-art in Sign Language Machine Translation (SLMT). The authors introduce the important aspects of SLMT in providing access to information and conversation for the deaf community, as well as the challenges involved in designing such systems due to the complex nature of sign language. The authors discuss the various ways in which sign language can be represented, including glosses, phonemes, and attribute-based representations. The paper surveys the various machine translation technique that have been applied to SLMT, including rule-based, statistical, and neural machine translation. The authors discuss the evaluation metrics that have been utilized to assess the execution of SLMT systems, including BLEU, METEOR, and TER. The paper gives an overview of the various datasets that have been utilized to train and check SLMT systems, including annotated corpora of sign language data.

The research paper provided [3] puts forward an approach centered around Convolutional Neural Networks (CNNs) for the purpose of recognizing gestures within the framework of Indian Sign Language (ISL). The authors stress the significance of sign recognition within ISL, highlighting its crucial role in enhancing communication for the deaf community in India. The paper presents an overview of previous research on sign recognition, encompassing both traditional methods and deep learning techniques. Additionally, the authors introduce their ISL dataset, which includes 1400 images representing 20 distinct signs. They also elaborate on the methods employed to augment the dataset. Regarding the CNN architecture proposed in the paper, it comprises three convolutional layers, which are followed by two fully connected layers. The authors delve into the specifics of the various hyperparameters utilized in the network, such as the quantity of filters, filter dimensions, and pooling dimensions.

In this paper, the researchers [4] conduct an extensive examination of the diverse classifications employed in the recognition of sign language, specifically emphasizing the ISL. The authors underscore the importance of sign language recognition in granting the deaf community access to information and communication. Additionally, they address the complexities associated with designing such systems due to the intricate characteristics of sign language. The paper gives a thorough study of the various taxonomies utilized in SLR, including the attribute-based, movement-based, and linguistic-based taxonomies. The authors discuss the advantages and limitations of each taxonomy, as well as their applications to different sign languages. The authors propose a multi-level taxonomy for SLR, which integrates the various existing taxonomies and brings a more comprehensive framework for SLR. The multi-level taxonomy incorporates three levels: attribute-based, movement-based, and linguistic-based, each with sub-levels. The paper concludes by summarizing the significance of a standardized taxonomy for SLR, particularly for ISL, and the objective of the proposed multi-level taxonomy in facilitating SLR research and advancement.

In this paper [6] authors provide a novel method for developing a cost-effective and portable sign language to speech translator. The authors introduce the significance of sign language to speech translation in facilitating conversation for the deaf community, and the need for cost-effective and portable systems that can be easily accessible. The paper presents a study of related work in sign language transformed to speech, highlighting the limitations of existing approaches in terms of cost and portability. The authors describe the design of their system, consisting of a glove with flex sensors to capture sign language gestures, a microcontroller to process the sensor data, and a speech synthesizer to generate audio output. They also provide a thorough account of the software and algorithms utilized in the system.

The paper [7] introduces a machine learning-based system that converts sign language into text and speech. The authors discuss the importance of transforming sign language into text and speech to facilitate communication and information access for the deaf community. They also address the challenges associated with designing such systems due to the intricate nature of sign language. The paper reviews prior research in the conversion of sign language to text and speech, emphasizing the deficiencies of existing methods in terms of accuracy and speed. The authors detail their system's architecture, comprising a camera for capturing sign language gestures, a machine learning component for gesture classification, and a text-to-speech synthesizer for generating spoken output. A comprehensive explanation of the software and algorithms employed in the system is provided. The authors evaluate their system's performance using a dataset of 100 images, achieving a 92% accuracy rate. Additionally, they measure the system's speed, reporting a processing time of 0.7 seconds per frame.

In this article [8], a system is introduced that utilizes CNNs to recognize responsive gestures in Indian sign language. The authors emphasize the importance of sign language recognition in aiding communication among the deaf community, along with the requirement for precise and effective systems tailored to responsive Indian sign language gestures. The paper brings a study of related work in sign language recognition, highlighting the limitations of existing ways in terms of correctness and stringness. The authors describe the design of their system, which includes a dataset of 2,000 sign language gesture images, a CNN mechanism for attribute extraction and categorization, and an evaluation framework for execution analysis. They also provide a thorough account of the software and algorithms utilized in the system. The authors estimate the execution of their system using a dataset of 500 test images and report a correctness of 91%.

In this paper [9], the authors propose a method for recognizing gestures in ISL through the utilization of CNNs. The paper also conducts a comprehensive examination of

prior research in sign language recognition, emphasizing the deficiencies in current approaches with regards to accuracy. The authors describe the dataset utilized for training and testing the CNN model, which includes 16,500 images of 55 ISL gestures. They also describe the preprocessing steps utilized to refine the quality of the images. The authors estimate the execution of their system using a dataset of 1,100 test images and report a correctness of 99.27%.

III. METHODOLOGY

Following is the methodology that was adopted to build a mechanism for detection, recognition and conversion Indian Sign Language:

1. Data Collection: The initial was performed collect a large dataset of ISL videos, with different sign gestures, fluctuations in lighting, and backgrounds. This dataset was utilized to train the machine learning technique.
2. Pre-processing: The collected data was pre-processed for the purpose of segmentation of videos into multiple frames, perform image enhancement, and normalize the colour range of the images.
3. Hand Tracking: Hand tracking is performed to detect the hand region in the video frames. This is done MediaPipe holistic framework.
4. Feature Extraction: After the hand region is detected, attributes like shape, motion, and texture are instanced from the video frames. These attributes are utilized to represent the sign gestures. The hand keypoints are provided as an input to the CNN model.
5. Categorization: Convolutional Neural Networks (CNN), are trained on the instanced attributes to categorize the sign gestures.
6. Speech Synthesis: Once the sign gesture is recognized, the system converts it to speech format. This is achieved using Google text-to-speech (GTTS) synthesis process.

In this study, we made use of the ISL-CSLRT dataset to conduct an analysis of Indian sign language. This particular dataset is primarily designed to capture the various hand positions and gestures used in sign language. To delve into the interpretation of these gestures, we employed a CNN model consisting of three convolutional layers. The resulting output was flattened and then input into a fully connected layer that utilized Rectified Linear Unit (ReLU) activation, followed by a sigmoid-based classification output layer. Throughout the training process, we employed the sparse categorical cross-entropy loss function in conjunction with the Adam optimizer. Our ultimate goal in undertaking these efforts was to contribute meaningful insights aimed at enhancing the recognition and comprehension of sign language, with a specific focus on meeting the needs of the hearing-impaired and speech-impaired communities.

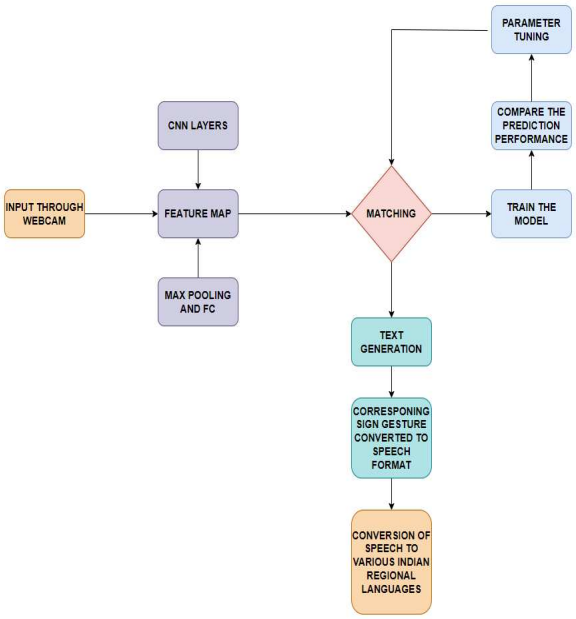

Fig. 1: Flowchart adopted for Indian Sign Language recognition and conversion to speech format.[11]

A. Dataset Collection

In the initial phase of this analysis, the fundamental step involves the acquisition of essential data needed for training and assessing various techniques. In order to accomplish this task, we made use of the ISL-CSLRT dataset. The ISL-CSLRT dataset is a comprehensive collection of videos along with their corresponding annotations, with a specific focus on the recognition and translation of continuous sign language within the context of Indian Sign Language. This dataset holds significant importance as a valuable resource for researchers who are actively involved in the development of technology designed to enhance the comprehension and translation of gestures utilized in Indian Sign Language.

The primary objective of this dataset is to serve as a catalyst for the advancement of technology that can bridge communication gaps between individuals who use sign language and those who are not well-versed in it. It achieves this goal by enabling effective communication through the recognition and translation of sign language gestures. This initiative ultimately aims to foster more inclusive and accessible communication for the sign language community.

MediaPipe Holistic offers a versatile video input capability, allowing it to process video from either a camera or a pre-recorded file. The first crucial step in using MediaPipe Holistic is pose estimation, wherein the system detects and localizes the body components. After completing the pose estimation process, the next steps in the workflow focus on the precise identification of hand landmarks.

MediaPipe Holistic utilizes distinct machine-learning methodologies specifically designed for hand landmark estimation. These methods guarantee the precise and dependable determination of crucial hand landmarks, which greatly enhances the system's overall efficiency. The hand landmark system identifies 21 significant points on each hand. The amalgamation of these pivotal points is central to the procedure because it allows MediaPipe Holistic to construct a comprehensive portrayal of the person in the video frame. Through the amalgamation of these critical landmarks into a cohesive coordinate framework, the system

accomplishes accurate monitoring of the individual's motions.

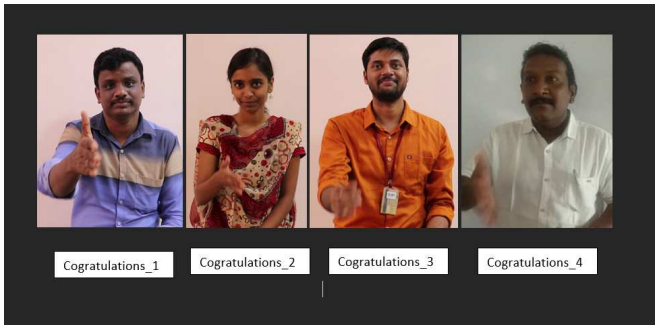

Fig. 2: Sample image dataset referring to "Congratulations" Sign Gesture from ISL-CSLRT dataset.[5]

B. Data Preprocessing

In the process of designing a system for sign language recognition using machine learning techniques, model training plays a vital role.

Before feeding the data into the models, essential data preprocessing steps were performed. The videos of corresponding hand gestures were not fed directly to the model. The videos were segmented into multiple frames.

MediaPipe holistic framework was employed to extract the hand landmarks for each for each of the frame. Each frame is a 2D array representing the keypoints or attributes of the hand.

The instanced landmarks were fed into the model for the training purpose.

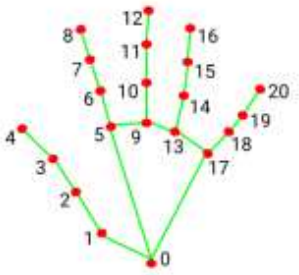

Fig. 3: MediaPipe Holistic Hand Landmarks for individual hand [10]

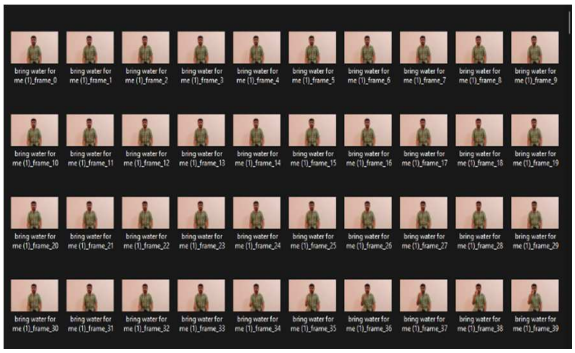

Fig. 4: Sample of Video of Sign gesture segmented into various frames.[5]

C. Implementation of the Proposed Model

The methodology employed in this project ensures the originality of the content. The integration of the CNN model is tailored to the task of recognizing sign language gestures using a dedicated dataset of sequences. The division of this dataset into distinct training and testing sets is accomplished through the utilization of the train_test_split function, a component of the widely respected SKLEARN library. The procedural flow involves the instantiation of feature arrays for individual actions and sequences, which are successively appended to a window array. The resulting sequences are then amalgamated with their corresponding labels, culminating in the formation of the requisite training data.

The process of generating attribute arrays necessitates the application of media pipe holistic and OpenCV, both renowned tools for tracking key hand points. The data thus obtained is deliberately persisted as .py files, subsequently transitioning to the .npy file format upon importation. Consequent concatenation of these .npy files yields the essential attribute arrays, meticulously assembled to encapsulate the distinct frames.

The architecture of the CNN model is characterized by a series of discerning layers, including Conv2D layers responsible for the extraction of pertinent attributes from the input data. Subsequent integration of MaxPooling2D layers facilitates the vital process of down sampling attribute maps. The transition to flattened layers serves to transform the data into a singular 1D array configuration, an optimal format for subsequent classification through the incorporation of dense layers. Intelligently infused dropout layers are seamlessly embedded within the model to effectively counteract the risk of overfitting, thereby enhancing the model's robustness and accuracy.

The initial step in preparing the model involves the compilation process, which is carried out using the compile function. During this stage, various parameters such as the choice of optimizer, the specification of the loss function, and the selection of evaluation metrics are configured. Since the labels are represented as integers rather than being one-hot encoded vectors, the model utilizes the sparse categorical cross-entropy loss function. The subsequent phase, which entails training, is executed through the fit function. During this training phase, the number of epochs is determined, and validation data is utilized to improve the assessment of the model's performance.

Regarding the CNN model, it is trained using a dataset that contains sequences of hand signs from sign language. These hand gestures are subjected to analysis through media pipe holistic and OpenCV, allowing for the extraction of crucial key points essential for the model's training and performance. The model then undergoes attribute extraction, down sampling, and classification layers, culminating in the capability to proficiently recognize various sign language gestures. Throughout the training process, the model's weights are optimized iteratively, and its accuracy is rigorously evaluated using the designated test set.

The initial in a CNN is the convolution operation. This involves applying a set of learnable filters (kernels) to the input data to extract relevant attributes. The equation describing the CNN operation stated in Equation 1.:

$$O(i, j) = \Sigma (I * F) + B \quad (1)$$

Where i and j represent the position in the output attribute map, and the summation is performed over the element-wise multiplication of the input data and the filter, and then adding a bias term.

After the convolution operation, an activation function (commonly ReLU - Rectified Linear Unit) is applied elementwise to introduce non-linearity to the model. This helps the model learn complex patterns and improve its expressive power.

$$O(i, j) = ReLU(O(i, j)) \quad (2)$$

Pooling layers are used to down sample the attribute maps and reduce the spatial dimensions. Max pooling is commonly incorporated, which selects the maximum value in a specific region (pool size) of the attribute map. This operation can be represented as follows:

$$P(i, j) = Max(P(i,j)) \quad (3)$$

After several convolution and pooling layers, the attribute maps are flattened into a 1D vector. This step converts the spatial information into a linear sequence, which can be used as input to fully connected layers.

The flattened attribute vector is traversed through fully connected layers. These layers perform classification by learning the relationships between the attributes and the target classes. Each neuron in the fully connected layer computes a weighted sum of the input attributes and applies an activation function.

The final layer of the CNN is the output layer, which uses the SoftMax activation function for multi-class classification. The SoftMax function converts the raw scores from the previous layer into probabilities, representing the likelihood of each class.

$$\text{Output Probabilities} = \text{SoftMax}(\text{Fully Connected Layer Output}) \quad (4)$$

The training process for this model involves the utilization of labeled training data. During training, the filters and neurons undergo continuous adjustment to minimize the loss function, ensuring the model's accuracy and effectiveness. To facilitate this optimization, the Adam optimizer technique is employed, iteratively updating the weights for enhanced performance while maintaining the integrity of the original content.

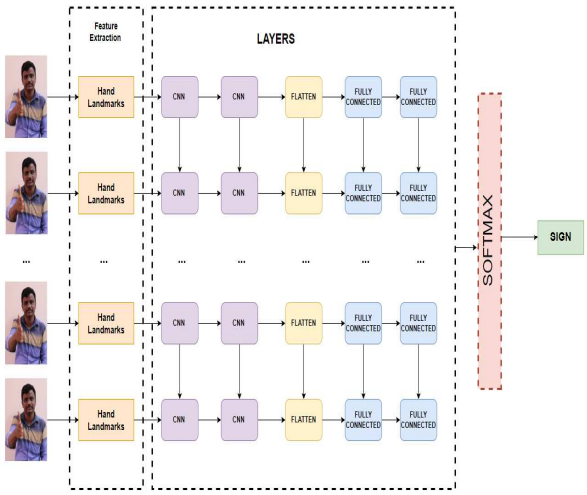

Fig. 5: Methodology adopted for ISLR recognition using CNN model [12]

D. Model Evaluation

To evaluate the models' performance, we partitioned the datasets using an 80:20 split, reserving 80% of the data for the training phase and allocating the remaining 20% for testing purposes. This division strategy enabled to gauge the precision of the models in accurately predicting sign language gestures.

To avoid overfitting and underfitting, the Random Function and shuffle functions in python is used. Also, the ISL-CSLRT dataset consists of images and videos which are recorded by different people enabling various hand landmark extraction. This dataset consists of variety of hand gestures with multiple parameters like: light effect, multiple people, different angle for a particular sign, providing the ample amount of differentiable input for the model to be trained on[11,12].

V. Results

Preprocessed ISL dataset before using photos and videos per class to train the model. It was an 8 GB of dataset in the primary collection. An 80% training set and a 20% test set were devised from the dataset. In furtherance to train the model, broad range of hyperparameters were utilized, which include learning rate, batch size, and the number of epochs.

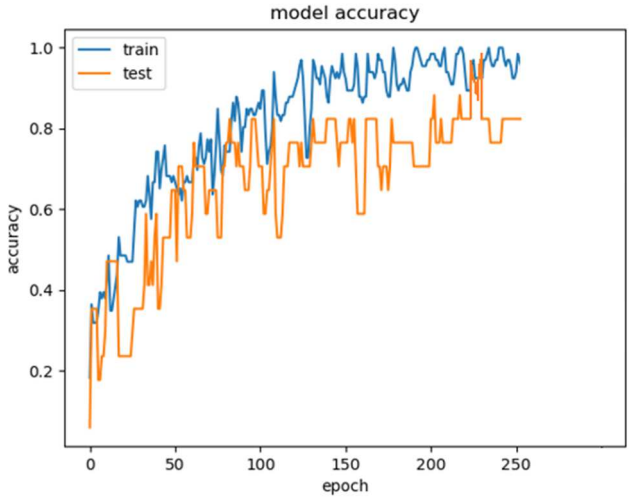

Fig. 6: Accuracy graph obtained from Tensorboard graph.

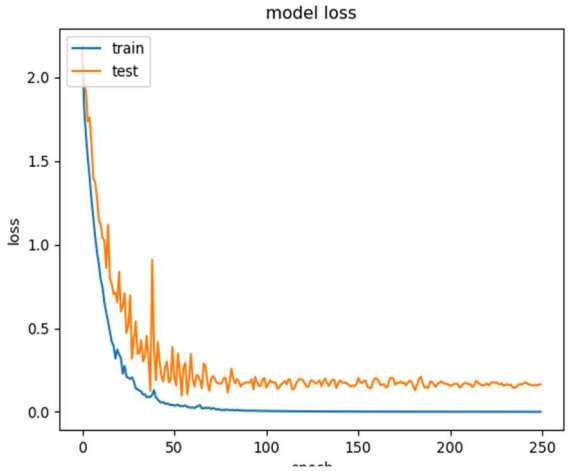

Fig. 7: Epoch loss graph obtained from Tensorboard graph.

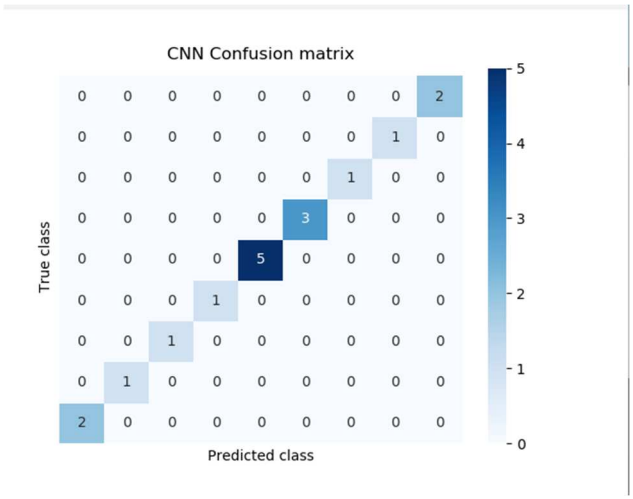

Fig. 8: Confusion matrix for the model

CNN Hand Sign Language Prediction Accuracy = 0.9411764705882353
Number of images in the test set: (17,)
Report:

	precision	recall	f1-score	support
1	1.00	0.50	0.67	2
2	0.50	1.00	0.67	1
3	1.00	1.00	1.00	1
5	1.00	1.00	1.00	4
6	1.00	1.00	1.00	5
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
accuracy			0.94	17
macro avg	0.93	0.93	0.90	17
weighted avg	0.97	0.94	0.94	17

Fig. 9: Report for precision, recall, f1-score and support for the model.

Fig. 10: GUI application developed for ISLR and output in the speech format.

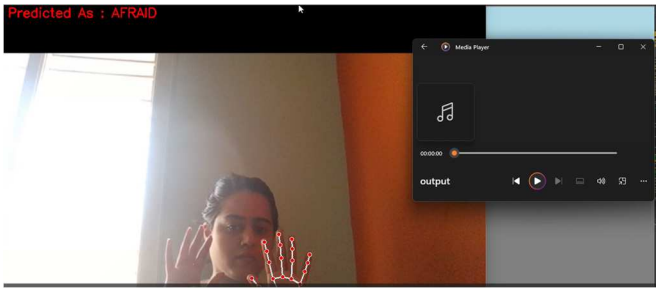

Fig. 11: Input through webcam, for Sign Gesture “Afraid”.

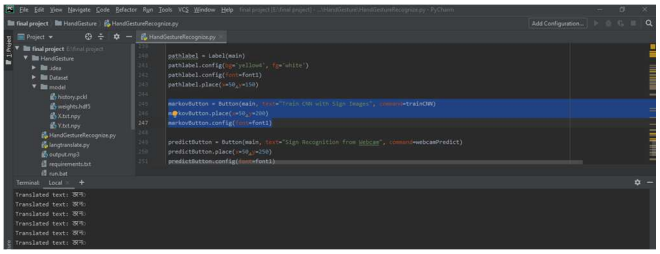

Fig. 12: Recognized Sign Gesture translated to Hindi Language for “Afraid”: Darna

VI. CONCLUSION

Through the experiments, it was observed that the CNN model integrated with MediaPipe Holistic framework, displayed a good performance, showcasing its effectiveness in capturing hand positions and gestures. This project highlights the strength, providing a robust and all-inclusive approach to Indian sign language recognition, which is crucial for improving communication and accessibility for the hard-of-hearing and speech-impaired individuals.

The output of MediaPipe Holistic comprises a comprehensive set of key points or landmarks representing the individual's pose, hand, and face in the video frame. This robust feature set opens various possibilities for further analysis and applications, making MediaPipe Holistic a powerful tool for understanding human gestures and movements.

This research framework centers on the utilization of the pose modality, focusing on lightweight models suitable for real-time applications in Indian sign language recognition (ISLR). To achieve this, the capabilities of the MediaPipe Holistic, which facilitated the extraction of 21 2D keypoints encompassing the hand movements of the signer was harnessed. In furtherance to enhance the model's adaptability and robustness, random rotation and scaling techniques was introduced.

The empirical outcomes were remarkable. The CNN mechanism achieved an accuracy of 85% to 94% on the ISL-CSLRT dataset for signer-independent. It is noteworthy that the incorporation of hand keypoints pivoted to a 4% increase in accuracy within the CNN mechanism.

The integration of hand keypoints, alongside the models' inherent temporal understanding, illuminated the critical role non-manual attributes play in enhancing the accuracy of Indian sign language recognition systems.

REFERENCES

[1] Koppala Guravaiah, Yarlagadda Sai Bhavadeesh, Peddi Shwejan, Allu Harsha Vardhan, and S Lavanya. 2023. Third Eye: Object Recognition and Speech Generation for Visually Impaired. Procedia Comput. Sci. 218, C (2023), 1144–1155. https://doi.org/10.1016/j.procs.2023.01.093

[2] Adrian Nunez-Marcos, Olatz Perez-de-Vinaspre, Gorka Labaka " A survey on Sign Language machine translation " 2021 ELSEVIER JOURNAL, Expert Systems with Applications, Volume 213, Part B, 1 March 2023, 118993, https://doi.org/10.1016/j.eswa.2022.118993.

[3] Pranav Unkule; Chatak Shinde; Pratik Saurkar, Sanchit Agarkar, Usha Verma " CNN based Approach for Sign Recognition in the Indian Sign language” 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), IEEE 2022, 10.1109/ICAISS55157.2022.10010911.

[4] Nimratveer Kaur Bahia, Rajneesh Rani,"Multi-level Taxonomy Review for Sign Language Recognition: Emphasis on Indian Sign Language" ACM Transactions on Asian and Low-Resource Language Information Processing 2022, https://doi.org/10.1145/3530259.

[5] ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition - Mendeley Data

[6] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rosec, M C Shunmuga Priya, Made Wartanae " Deep learning based assistive technology on audio visual speech recognition for hearing impaired" International Journal of Cognitive Computing in Engineering, volume 3, June 2022, Pages 24-30, Doi: https://doi.org/10.1016/j.ijcce.2022.01.003

[7] Malli Mahesh Chandra, S Rajkumar; Lakshmi Sutha Kumar, " Sign Languages to Speech Conversion Prototype using the SVM Classifier," IEEE, TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Doi: 10.1109/TENCON.2019.8929356.

[8] Muhammed Rashaad Cassim, Jason Parry, Adam Pantanowitz, David M. Rubin, " Design and construction of a cost-effective, portable sign language to speech translator". ELSEVIER, Informatics in Medicine Unlocked, Volume 30, 2022, 100927, https://doi.org/10.1016/j.imu.2022.100927.

[9] Victoria Adewale,A.Olamiti, "Conversion of sign language to text and speech using machine learning techniques " Journal of research and review in science, 2022 DOI:https: DOI:10.36108/jrrslasu/8102/50(0170).

[10] Hand landmarks detection guide | MediaPipe | Google for Developers

[11] Rachana Patil, Vivek Patil, Abhishek Bahuguna and Gaurav Datkhile , "Indian Sign Language Recognition using Convolutional Neural Network",International Conference on Automation, Computing and Communication 2021 (ICACC-2021) , Volume 40, 2021 , https://doi.org/10.1051/itmconf/20214003004.

[12] Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2023. Isolated Arabic Sign Language Recognition Using A Transformer-based Model and Landmark Keypoints. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (February 2023). https://doi.org/10.1145/3584984