Hand Gesture Recognition and Text Conversion Using Convolutional Neural Networks

PROF. JYOTSNA NANAJKAR[1], HARSH KSHIRSAGAR[2], ANIKET SHELKE[3], NIKITA SHINDE[4], MANOJ THOMBRE[5]

[1] *Assistant Professor, Department Information Technology, Zeal College of Engineering and Research Pune*

[2, 3, 4, 5] *Student, Department of Information Technology, Zeal College of Engineering and Research Pune*

Abstract— Sign language is an important means of communication for people with speech disabilities, but it presents significant challenges for non-signers due to a widespread lack of interpreters and awareness. This paper explores the development of a hand sign understanding and translation system using convolutional neural networks (CNN) to bridge the communication gap between hearing and deaf communities. Our research focuses on a three-step methodology: data collection, model training and extensive evaluation. Using a custom CNN architecture, our system can detect and convert hand gestures into real-time text, providing a complete communication solution. The methodology includes a dataset specially curated for this purpose, and the training phase uses the MNIST dataset to initially calibrate the model. Our system demonstrates a remarkable 95.7% accuracy in recognizing the 26 letters of the American Sign Language (ASL) alphabet, demonstrating its potential to facilitate seamless communication between signers and non-signers. This advance highlight the promising application of deep learning methods to improve accessibility and inclusion in deaf and hearing communities.

Index Terms- Sign Language, ASL, Hearing disability, Convolutional Neural Network (CNN), Computer Vision, Machine Learning, Gesture recognition, Sign language recognition, Hue Saturation Value algorithm.

I. INTRODUCTION

In 2023, the World Wellbeing Association (WHO) announced that north of 430 million individuals universally experience the ill effects of hearing misfortune, addressing around 5% of the total populace. This number is projected to raise to in excess of 700 million by 2050. Sign Arrangements features that there are more than 300 different gesture-based communications being used around the world, further convoluting correspondence between the conference and the hard of hearing or nearly deaf. Successful correspondence is fundamental for human connection, yet the conference and hard of hearing networks face huge difficulties in such manner. To overcome this issue, we propose an original methodology utilizing Convolutional Brain Organizations (CNNs) for hand communication via gestures acknowledgment and text interpretation. This study examines the plausibility of precisely and dependably making an interpretation of hand sign developments into text utilizing profound learning advancements. Via mechanizing the acknowledgment and characterization of hand signs, our methodology expects to give continuous correspondence answers for gesture-based communication clients, accordingly improving comprehensive correspondence. This exploration features the significance of creating hearty profound learning models that can sum up across various populaces and imaging conditions, utilizing huge scope commented on datasets. The philosophy includes building a custom CNN model to perceive motions in gesture-based communication.

The model contains 11 layers, including convolutional layers, max-pooling layers, thick layers, a levelling layer, and a dropout layer. We utilize the American Gesture based communication Dataset from MNIST to prepare the model, zeroing in on highlight extraction and arrangement precision. The remainder of this paper is coordinated as follows: Segment 2 surveys related work; Area 3 depicts the datasets and their attributes; Area 4 subtleties the model design; Segment 5 presents trial results and perceptions; Segment 6 examines difficulties and limits; and Segment 7 frameworks future examination bearings. This examination highlights the capability of profound learning methods in changing clinical picture

examination and further developing availability, effectiveness, and exactness in the symptomatic cycle for different circumstances, especially with regards to hand communication via gestures acknowledgment.

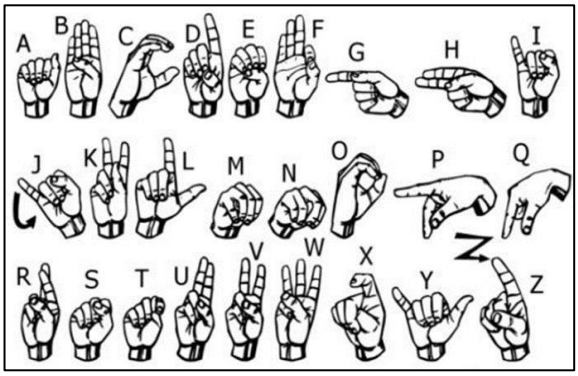

Fig1. American Sign Language Gestures

The most common sign language is American sign language. D&M people can only communicate through sign language because their only communication problem is connected to their inability to use spoken languages. The practice of exchanging ideas and messages through words, gestures, body, and visual aids is known as language communication. People who are deaf or dumb use their hands to communicate with others by making various gestures. Nonverbal cues are sent by gestures, which can only be perceived by eyes. Our study primarily focuses on building a model that can identify hand movements based on fingerspelling and combine each motion to produce a whole word.

II. RELATED WORK

[1] Wenjin Zhang, Jiacun Wang, Senior Member and Fangping Lan discuss in "Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks," how to use datasets such as MNIST, which provides an accuracy of 99.6% for image recognition using a CNN model, and Sports-1M, which provides an accuracy of 95.1% for vanishing gradient problem-solving using LSTM, for action recognition. The Jester dataset yields an accuracy of 94.85% and a change in loss. The average accuracy of the Nvidia dataset results is 88.4%.

Several studies have been carried out to address the sign identification in using a variety of methodologies and algorithms for images and videos, according to a review of the literature on our proposed approach.

[2]Hassan Mathkour, Mohamed Mekhtiche, Muneer Al-hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohammed the "Deep Learning-Based Approach for Sign Language Gesture Recognition with Efficient Hand Gesture Representation" dataset, which was recorded by 40 participants across five recording sessions with accuracy of 98.75%, is similar to the King Saud University Saudi Sign Language (ksuSSL) dataset. The fine-grained characteristics of the hand shape were learned using two distinct 3DCNN instances. The classification was done using the SoftMax function. The efficacy of the suggested system was demonstrated by the fact that it outperformed cutting-edge techniques in terms of recognition rate.

Indian sign language signs (A-Z) and (0-9) are recognized by Shagun Katoch, Varsha Singh, and Uma Shanker Tiwary's "Indian Sign Language Recognition System using SURF with SVM and CNN" using SVM with accuracy of 99.14% on the test data and CNN with accuracy of 94% on the training set. It was trained on all 36 ISL static alphabets and digits with an accuracy of 99%.

Kent and Yulius Obia Aditya Kurniawana, Said Achmada, Vetri Marvel Budimana, and Samuel Claudioa with the ASL Hand Sign Dataset (Gray scaled Thresholded), which comprises 24 classes with a gaussian blur filter added, a "sign language recognition system for communicating to people with disabilities" is created. Three dataset classes—classes J, Z, and 0 (blank)—were obtained from Nikhil Gupta; these datasets had training and validation accuracy of 89.1% and 98.6%, respectively. 96.3% accuracy is achieved by using a two-layer Convolutional Neural Network (CNN) with a confusion matrix.

The authors of "Indian Sign Language Communicator Using Convolutional Neural Network," Arvind Sreenivas, Mudit Maheshwari, Saiyam Jain, and Shalini Choudhari, employed their own dataset of two-handed motions utilizing the CNN model and the Relu activation function.

Sayali Gore, Swati Singh, and Namrata Salvi the CNN method is specifically utilized in "Conversion of Sign Language into Text Using Machine Learning Technique" to increase recognition accuracy in difficult situations including scale, rotation, and translation changes.

The accuracy of the output is increased when a large dataset is used.
Gesture segmentation and feature extraction play a crucial role in hand sign language recognition by facilitating the interpretation of meaningful signs. In order to overcome this difficulty, Chen et al. (2019) created a technique for dividing continuous sign language motions into discrete components. Their method greatly increased the accuracy of gesture detection by combining both temporal and spatial characteristics, which increased the effectiveness of sign language translation systems. Furthermore, training and assessing machine learning models has been made possible by the availability of extensive sign language datasets. A comprehensive database of sign language motions and expressions was assembled by Liang and Ren (2017), giving academics an invaluable resource for building reliable data-driven recognition algorithms.

Simultaneously, attempts have been made to create real-time systems for translating sign language, which will let deaf people communicate with non-signers. A CNN-based model was incorporated by Wang et al. (2021) into a wearable gadget that could translate American Sign Language (ASL) movements into text instantly. This creative approach has great potential to promote inclusive communication and enable deaf people to interact more easily with the larger community. Even with these developments, problems such guaranteeing robustness in changing environmental circumstances and model scalability remain, necessitating more study and development in the area of hand sign language recognition and text conversion.

III. SIGN LANGUAGE DATASET

To guarantee reliable results for sign language identification, a large dataset of hand sign photos was gathered. The collection is made up of pictures from the MNIST platform's American Sign Language (ASL) collection, which is openly accessible and contains thousands of pictures of hand signs. These pictures depict a broad variety of ASL capturing diverse hand shapes and postures against varied backdrops and lighting conditions. To improve the dataset, bespoke photos were also gathered, guaranteeing a varied depiction of hand signs for precise and trustworthy model training. This combined dataset offers a strong basis for creating a convolutional neural network (CNN) model that can identify and translate hand signals into text.

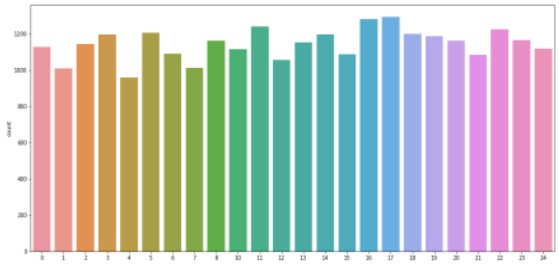

Fig 2. Count plot of the Dataset motions

IV. METHODOLOGY

The following describes the process for utilizing Convolutional Neural Networks (CNNs) to recognize hand sign language and convert it into text, including every stage from data collection to ongoing refinement.

1. Information Gathering:

By means of data gathering and dataset acquisition, an extensive collection of sign language gestures was obtained. In order to ensure variety and representation of varied hand forms, orientations, lighting situations, and sign languages, this collection comprises pictures and video clips of hands performing various signs.

Data Annotation: The correct label identifying the particular sign being done was marked on each picture or video frame. This labelled dataset, which provides the ground truth needed to train the model, is crucial for supervised learning.

2. Preprocessing Data:

Data preprocessing was done to improve and standardize the raw data once it was collected. Image

Resizing: All of the photographs were resized to a common size, such as 64x64 or 128x128 pixels, so order to standardize the input data.

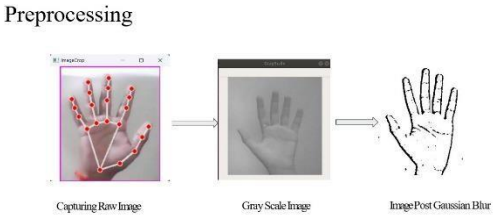

Fig 3: Image Processing

This stage lowers computing cost and guarantees consistency throughout the dataset. Normalization: By dividing the pixel values by 255, the maximum pixel value, the picture pixel values were normalized to a range of [0, 1]. By doing this, a common scale for the data is guaranteed. Data Augmentation: To improve the dataset, many transformations were used, including rotation, flipping, zooming, and shifting. By adding more diversity to the training set, data augmentation improves the model's ability to generalize to new, untested data.

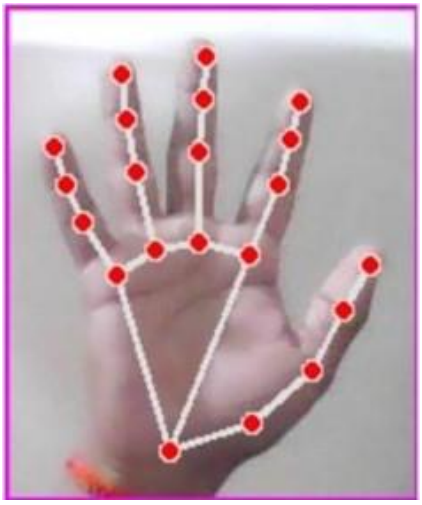

Fig 4

3.Architecture Model:

To recognize hand signs, a bespoke Convolutional Neural Network (CNN) was created. Among the architectural features are:
Convolutional Layers: To extract spatial information from the photos, four convolutional layers with different filter sizes (e.g., 7x7 filters for the first layer and 3x3 filters for subsequent layers) were employed.
Maximum Pooling Layers: To lower the dimensionality and down sample the feature maps, three max-pooling layers with 2x2 pooling were added. This lowers computational complexity and avoids overfitting.
Flatten Layer: The 2D feature maps were transformed into a 1D vector so that fully connected layers may use them as input. Fully Connected Layers: From the flattened vector, two thick layers were utilized to extract complicated patterns.
Dropout Layer: To avoid overfitting during training, a 20% dropout rate dropout layer was implemented. This layer randomly drops neurons during training.

Output Layer: To generate probabilities for each of the 24 sign classes, a multi-class classification layer using a SoftMax activation function was included.

4. Training the Model Loss Function:

The loss function for multi-class classification was categorical cross-entropy. This function compares the predicted class probabilities to the true labels to assess the model's performance.
Optimizer: The Adam optimizer, which dynamically modifies the learning rate to enhance convergence and model performance, was used.
Training Process: Using a validation set to keep an eye on performance and avoid overfitting, the model was trained on the pre-processed dataset. To prevent overfitting, early halting was applied if the validation loss did not improve after a predetermined number of epochs.

5. Assessment:

Several indicators were used to assess the model's performance once it had been trained.
Accuracy and Loss measures: On both the training and validation datasets, the model's performance was evaluated through the use of accuracy and loss measures. These measures shed light on the model's learning and generalization performance.
Confusion Matrix: To see how well the model performs in various classes and to spot indications that are commonly misclassified, a confusion matrix was created.

Cross-validation: To make sure the model is reliable and broadly applicable, cross validation was carried out. To validate the model's performance, this included splitting the dataset into many folds and training the model on various subsets.

6. Implementation:
Real-Time Recognition: Using the trained model as a basis, an application was created to recognize sign language in real-time. This required employing a webcam or other video device to record and capture hand motions during live broadcasting, then processing the frames in real-time.
Post-Processing: To increase the stability and accuracy of the recognition system, post-processing techniques were used, such as smoothing the predicted labels throughout a sequence of frames.
7. Continuous Improvement:
Feedback Loop: In order to gather user opinions on the functionality of the system, a feedback loop was put in place. To increase the model's accuracy, misclassifications were found and corrected, and fresh data was constantly supplied.
Model Fine-Tuning: In order to improve overall performance and adjust to variations in sign language gestures, the model was periodically adjusted using new data. This required using an updated dataset for retraining the model and modifying the hyperparameters as necessary.
By employing this technology, it is hoped to create a reliable system for identifying and converting hand sign language into text, helping those who are hard of hearing to communicate effectively.

V. NETWORK ARCHITECTURE (CNN

The CNN model developed has 11 layers: Four Convolutional Layers: The initial convolutional layers use a filter (or kernel) size of 7x7.These are followed by convolutional layers with a filter size of 3x3.Three Max Pooling Layers: Each max pooling layer uses a 2x2 pool size, which helps reduce the spatial dimensions efficiently Other Layers. Two Fully Connected (Dense) Layers: These layers are used after the data has been flattened.

One Flatten Layer: This layer converts the 2D data from the convolutional and pooling layers into a 1D vector, which can be fed into the dense layers.

One Dropout Layer: This layer randomly drops 20% of the neurons during training to prevent overfitting.

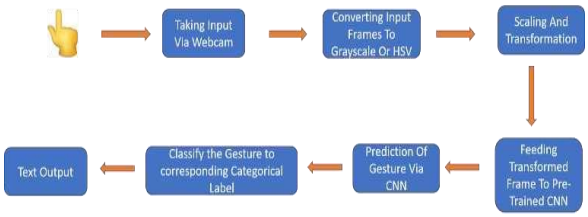

Fig. 5: High Level System Architecture

System flowchart:

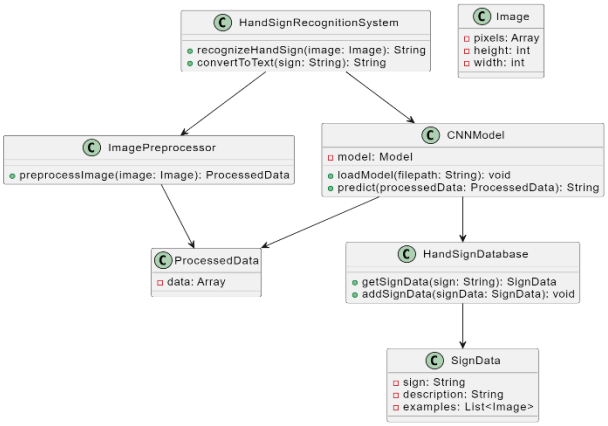

Fig 6: System Flowchart

The model uses ReLU (Rectified Linear Unit) activation in the fully connected layers to keep the positive values and discard the negative ones. The final layer uses a SoftMax activation function to classify the input into one of 24 classes.

After training the model from scratch, it achieves an accuracy of 99% on the training set, indicating a very high level of performance.

VI. PROPOSED WORK

Individuals who are deaf or hard of hearing try to communicate with the general public in American Sign Language, yet they are unaware of this sentence construction in the language. In the end, crucial to developing an intelligent and programmable arbitrator to obtain them. The suggested system operates by adhering to the specified methods for identifying and translating the hand motions into written communication and vice versa.

A. Convert ASL into Speech and Text

Take a picture or a video and feed it into the system.

1. The most basic stage of processing digital images is image acquisition and enhancement.
2. Every pertinent feature is retrieved, and every redundant and unnecessary detail is disregarded.
3. The input is analysed to identify the correct character.
4. Text and speech are produced from the identified ASL characters.

B. Transcribing Text or Speech into ASL

1. Record the user's voice or speech and feed it into the system as an input.
2. The input that is received will be examined and transformed into its corresponding text
3. The corresponding text is found
4. The text is there after transformed into the matching ASL character.

Setup for an experiment:

1. There are 16,120 entries in our data set overall, and 26 labels are used for training (30%).
2. The dataset is being divided at random by divided into 20% testing and 80% training.
3. We need processors with Intel I3 or above.
4. Memory: 80 GB, RAM: 4 GB
5. It will result in improved GPU performance.
6. The camera is positioned appropriately for taking a picture of the input.
7. The Eclipse/NetBeans IDE/ Visual Studio Code
8. Python for the dataset's testing and training with CNN

Sequence diagram:

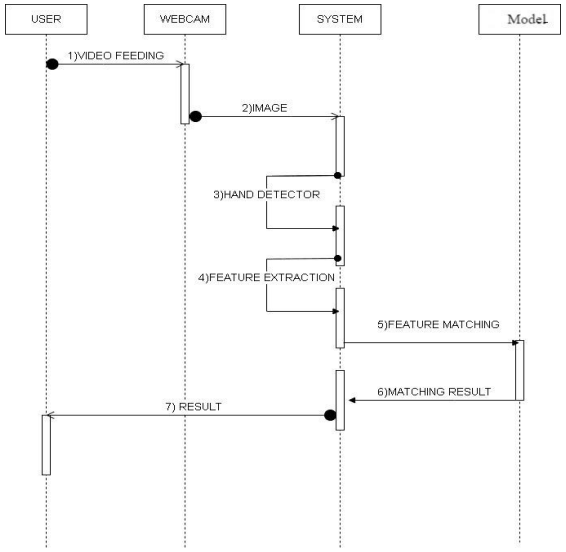

Fig 7: Sequence Diagram

VII. RESULT AND DISCUSSION

Using a customized CNN model, our research to recognize sign language produced encouraging results. The following are the main results:

Great Accuracy: When classifying and identifying sign language motions from webcam video input, the customized 11-layer CNN model showed great accuracy. The preprocessing and segmentation stages made a substantial contribution to performance improvement and error reduction.

Low False Positives: We reduced background noise and extraneous information by concentrating on a particular area of interest and grayscale-converting photos. This method greatly reduced the false positive rate, which increased the system's dependability.

Effective Preprocessing: The model's input was consistent with the data it was trained on thanks to the preprocessing procedures, which included scaling and grayscale image conversion.

Accurate Recognition in Real Time: Real-time video input was handled by the system efficiently, allowing for fast gesture prediction and frame segmentation. After the gestures were identified, text was presented to give instant feedback.

OUTPUT:

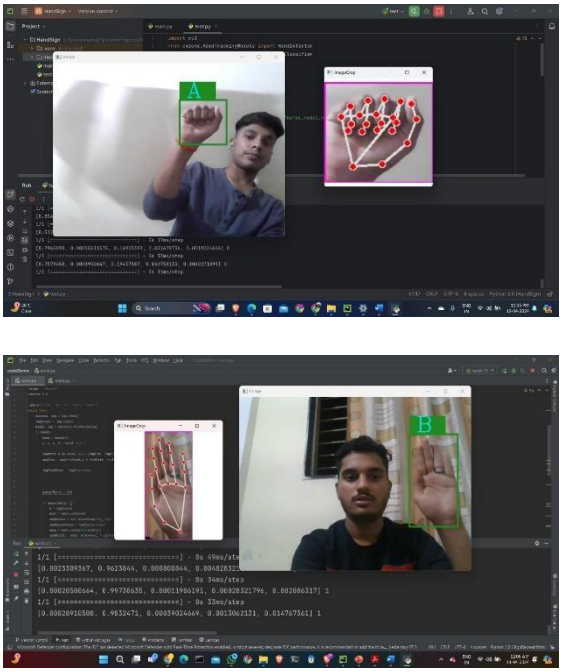

Fig 8: System Output

Sturdy Performance: The model demonstrated resilience and generalizability by exhibiting strong performance in a variety of user and environment contexts. It is therefore a workable solution for practical uses.

In conclusion, our project was successful in creating a real-time, accurate, and dependable sign language recognition system. These findings show how our method may be used to enhance communication resources for those who use sign language.

CONCLUSION

This paper introduces a Convolutional Neural Network (CNN) based approach for the recognition and classification of sign language using computer vision. The proposed method demonstrates superior accuracy and a significantly lower rate of false positives compared to traditional techniques. This advancement is crucial for improving the reliability and effectiveness of sign language recognition systems. Sign language recognition has been approached in various ways, including vision-based and glove-based methods. Vision-based methods, which utilize video and image processing, offer the advantage of being non-intrusive, unlike glove-based methods that require the user to wear special devices. However, vision-based methods have traditionally faced challenges in terms of accuracy and the rate of false positives.

The CNN-based approach outlined in this paper addresses these challenges effectively. By leveraging deep learning, the system can learn complex patterns and features from large datasets, which enhances its ability to accurately recognize and classify sign language gestures. This results in improved performance metrics, making the system more reliable for practical applications. One of the key advantages of the CNN-based approach is its ability to generalize across different users and environments. This generalization is essential for creating robust sign language recognition systems that can be used in diverse real-world scenarios. The system's lower false positive rate also ensures that the user experience is more seamless and less prone to errors, which is particularly important for applications in communication aids and assistive technologies. While the current implementation focuses on static gesture recognition, there are promising extensions to this work that are being explored. One such extension is the inclusion of dynamic gesture recognition, which involves recognizing gestures in motion over time. This would significantly expand the system's capabilities, allowing it to understand more complex sign language phrases and sentences, thereby making it more useful in practical applications. In conclusion, the CNN-based approach presented in this paper marks a significant step forward in the field of sign language recognition. Its high accuracy and low false positive rate, combined with its potential for future enhancements like dynamic gesture recognition, position it as a valuable tool for improving communication for individuals who rely on sign language

REFERENCES

[1] Wenjin Zhang, Jiacun Wang, Senior Member, IEEE, and Fangping Lan, “Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks,” IEEE/CAA JOURNAL OF AUTOMATICA SINICA, VOL. 8, NO. 1, JANUARY 2021. 4 3

[2] Muneer Al-Hammadi, Ghulam Muhammad, Senior Member, IEEE, Wadood Abdul, Member, IEEE, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche, "Hand Gesture Recognition for Sign Language Using 3DCNN," DOI 10.1109/ACCESS.2020.2990434, IEEE Access.

[3] Mehreen Hurroo, Mohammad Elham Walizad, "Sign Language Recognition System using Convolutional Neural Network and Computer Vision," International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS120029 Published by: www.ijert.org Vol. 9 Issue 12, December2020.

[4] R.S. Sabeenian, S. Sai Bharathwaj, M. Mohamed Aadhil "Sign Language Recognition Using Deep Learning and Computer Vision," Jour of Adv Research in Dynamical & Control Systems, Vol. 12, 05-Special Issue, 2020

[5] Arvind Sreenivas, Mudit Maheshwari, Saiyam Jain, Shalini Choudhary, Dr.G. Vadivu, "Indian Sign Language Communicator Using Convolutional Neural Network," International Journal of Advanced Science and Technology Vol. 29, No. 3, (2020), pp. 11015 - 11031.

[6] Ahmed Sultan, Walied Makram, Mohammed Kayed, Abdelmaged Amin Ali, "Sign language identification and recognition: A comparative study," https://doi.org/10.1515/comp-2022-0240 received April 29, 2021; accepted April 12, 2022.

[7] Sayali Gore, Namrata Salvi, Swati Singh, "Conversion of Sign Language into Text Using Machine Learning Technique," International Journal of Research in Engineering, Science and Management Volume 4, Issue 5, May 2021 https://www.ijresm.com | ISSN (Online): 2581-5792.

[8] Rachana Patil1, Vivek Patil1, Abhishek Bahuguna1, and Mr. Gaurav Datkhile, "Indian Sign Language Recognition using Convolutional Neural Network," ITM Web of Conferences 40, 03004 (2021) ICACC2021.

[9] G. Poon, K. C. Kwan, and W.-M. Pang, "Occlusion-robust bimanual gesture recognition by fusing multi-views," Multimedia. Tools Appl., vol. 78, no. 16, pp. 23469–23488, Aug. 2019.

[10] N. K. Bhagat, Y. Vishnusai and G. N. Rathna (2019). Indian Sign Language Gesture Recognition using Image Processing and Deep Learning. Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia. pp. 1-8, doi: 10.1109/DICTA47822.2019.8945850.

[11] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, and M. S. Hossain, ''Hand gesture recognition using 3D-CNN model,'' IEEE Consum. Electron. Mag., vol. 9, no. 1, pp. 95–101, Jan. 2020.

[12] G. Muhammad, M. F. Alhamid, and X. Long, ''Computing and processing on the edge: Smart pathology detection for connected healthcare,'' IEEE Netw., vol. 33, no. 6, pp. 44–49, Nov./Dec. 2019.