WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comprehensive Study on Deep Learning-Based Approaches for Hand Sign to Speech Conversion for Deaf and Nonverbal Individuals.

B Gnaneswar Rao

GMR institute of Technology

ABSTRACT

Deep Learning has evolved exponentially in various fields, and one of the most fascinating applications of deep learning is the conversion of hand signs to speech. This study focuses on the accuracy-based efficiency of numerous machines built based on different models to establish a proper sign-to-speech conversion. The most optimal and efficient technologies that can be used are Single Shot Detector (SSD), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) and You Only Look Once algorithm (YOLO). SSD is basically an architecture mostly used for object detection purposes using computer vision. Although the deep learning model SSD requires a considerable amount of data, it has some salient features such as parallel processing and reduction of temporal dependencies. On the other hand, LSTM, which could be considered a type of Recurrent Neural Network (RNN), involves sequential processing and maintains a memory for past data. Similarly, CNN is a type of artificial neural network that is widely used for image or object recognition and classification. Models can also be built based on YOLO, which is an advanced object detection algorithm. Conclusively, the more accurate the model, the more efficient it is for establishing a bridge of effective communication between people with hearing disabilities and well-abled people.

Keywords: Sign-to-speech conversion, Long Short-Term Memory (LSTM), RNN, Single Shot Detector (SSD), deep learning

INTRODUCTION

Communication is a fundamental aspect of human interaction, playing a major role in conveying thoughts, emotions, and ideas. For individuals who are deaf or nonverbal, traditional modes of communication may pose significant challenges. The emergence of deep learning technologies has opened up new possibilities for addressing these challenges, offering innovative solutions to bridge the communication gap. One such area of exploration is the development of deep learning-based approaches for converting hand signs into speech, providing a transformative means of communication for deaf and nonverbal individuals.

Sign language's role could be considered as a very important role in facilitating communication for differently abled people especially deaf and non-verbal people. Sign languages such as Indian Sign Language (ISL), American Sign Language (ASL) rely on a rich vocabulary of hand signs to convey messages. That being the reason, all the alphabets are included in the Indian Sign Language as shown in the above image, enabling effective and sufficient communication means for the deaf and non-verbal people.

This comprehensive study involves exploring and evaluating different deep learning models in understanding and interpreting hand gestures and initially translating them into text and finally into speech format. The major objective of the research being to identify the potentials and complexities in studied models and providing the in-depth exploration of the deep learning model so that the best of them can be decided.

The technical depths of the deep learning architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and their variants will be explored and discussed. Comparative analysis and case studies would also be provided so that the insights into the strengths and limitations of deep learning approaches would be provided.

Ultimately, the outcomes of this study is to inform the depths of various approaches involved in the Hand Gesture to Sign conversion leading to an establishment of an effective bridge of communication among the well-abled communities, and Deaf and nonverbal communities leading to effective communication without any limitations and with at most ease for the disabled communities.

LITERATURE SURVEY

Shashidhar, R et al.,(2022) study shows the use of CNN through pre-prepared data. A dataset was to be processed and for the pre-processing techniques, background subtraction was done. The main objective of the study was to develop an associate degree automatic language recognition system with the help of a convolution neural network and laptop vision techniques and to use natural visual sequences, without requiring the signer to wear coloured or informational gloves, and to be able to recognise a variety of signals. [1]

Advantages: Reported high accuracy with the help of the Convolutional Neural Network (CNN).

Limitations: No detailed explanation of the feature extraction process.

Agarwal, R et al.,(2021) study's main objective was to create a Human–Computer Interaction (HCI) that understands the gestures and signs used for communication by the deaf and mute community. In order to increase the efficiency and performance of the model, a dataset containing 58,000 images of English alphabets gestures with each gesture containing 2,000 images for each alphabet. [2]

Advantages: Additional features such as use of keys 'p' are used to make words out of the characters and 'w' key is used to make sentences.

Limitations: The system lacks automation and it was observed to be difficult for integration with different technologies.

Rahim, M. A et al.,(2019) study's objective was to improve communication with the deaf community and the common people. For this, a hand gesture based human-computer interface was proposed in this paper. Feature fusion was done in order to learn image features fully for description of their internal information for increasing the accuracy. The features were extracted from two separate channels, one channel containing the grayscale image for extraction of intensity of the image and the other channel containing the HSV format image for extraction of colour of the image. [7]

Advantages: Two way channelled feature extraction process increasing the model's efficiency and resulting in obtaining of more information from the image.

Limitations: The model developed is only capable of recognizing certain words in hand gestures dataset not the individual characters.

METHODOLOGY

Reference 1:

[1] Shashidhar, R., Hegde, S. R., Chinmaya, K., Priyesh, A., Manjunath, A. S., & Arunakumari, B. N. (2022, October). Indian Sign Language to Speech Conversion Using Convolutional Neural Network. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)* (pp. 1-5). IEEE.

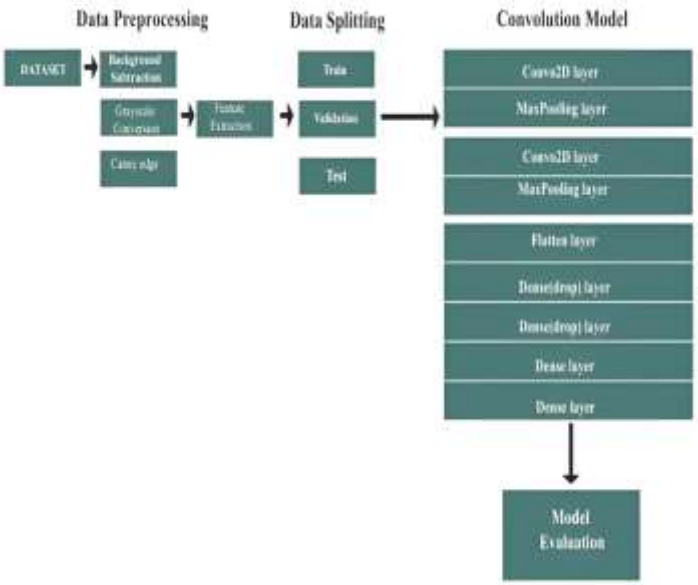


Fig 3.1

Dataset:

- Indian Sign Language dataset containing 35 classes (1-9 numbers and 25 Alphabets from A-Z) each class containing 1200 images.
- Total images in the dataset are around 42000.

Pre-processing:

- The pre-processing techniques used in this paper were background subtraction, grayscale conversion of the image and removing canny edges.
- The background subtraction involves removing the background of the image and only concentrating on the hand gestures and marking them with high frequency.

- After the backgrounds are removed from the images, the images ate then converted into grayscale images for extracting of intensity information of the image.
- Finally the canny edges were removed around the hand gestures which enable in more accurate feature maps.

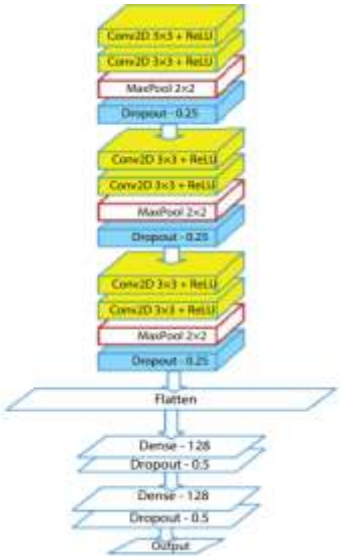
Implementation:

- **CNN** neural network containing two hidden layers, flatten layer, 3 dense layers with activation function ReLu and 1 dense layer with softmax function for feature extraction.
- For the preprocessing, the images have been initially converted into grayscale images and then Gaussian filters were applied to remove high-frequency components such as components in the image that contain rapid change in the intensity or color.
- The features are extracted from the images through some convolutional operations such as using small sized filters to produce feature maps by element wise summation and multiplication.
- The pooling layers used here reduce the dimensions of the image while retaining the important features.
- The CNN layers performed the following functions:

1. **Conv2D layers:** These layers take the two-dimensional images as the input and certain filters are applied to the images and hence the feature maps are obtained in this layer through the filters (kernels).
2. **MaxPooling Layer:** The function of the max pooling layers is to decrease the spatial size of the image by selecting the maximum representative element of the selected region.
3. **Flatten Layers:** The major function of the flatten layers is to reduce the dimensions of the feature maps obtained from the convolutional layers. These essentially reduce the features maps into one-dimensional feature vectors.
4. **Dense Layers:** These layers generally contain the activation non-linearity function such as ReLu function which helps in maintaining the non-linearity of the model useful for reading the complex information.
5. **Dense layer with softmax:** These layers generally convert the inputs into range of [0,1] through probability distribution functions which are helpful for the classification of the input. The input having the highest probability distribution value gets mapped to the particular neuron and this way classification is done.

Reference 2:

[2] Agarwal, R., Bansal, S., Aggarwal, A., Garg, N., & Kochhar, A. (2021). Study of Gesture-Based Communication Translator by Deep Learning Technique. *Smart and Sustainable Intelligent Systems*, 139-150.



- The neural network used is Convolutional Neural Network (CNN) which consisted of layers such as Conv2D (used for 2 dimensional data), MaxPooling layers, Dense layers, Dropout layers and flattening layers.
- For every single alphabet there has been a single neuron allocated which makes a total of 26 neurons in the last layer of CNN.

Dataset:

- The dataset used contains over 58000 images of alphabet gestures with each gesture containing 2000 different images.
- The images are pre-processed and then fed in to the CNN model and the pre-processing is done as follows.

Pre-processing:

- The images were preprocessed that is the images were converted from RGB format to HSV(Hue, Saturation, Value) format.
- Background subtraction techniques were also used in order to only concentrate on the features but not the background of the image.

Implementation:

- Initially the images are preprocessed that is converted from RGB pattern to HSV pattern (Hue representing the color, Saturation representing the intensity of the color 0%-gray and 100%-fully saturated and Value represents the brightness of the color).
- Then the preprocessed data is fed into the CNN model which initially extracts the features with the help of the filters and classification is done accordingly with the establishment of the neurons.
- Internally, after the flattening of the feature maps is done through "Flatten Layers", the flattened data is fed into dense layers (128 neurons established for each layer) and the output of the dense layers is then given to softmax layer which in turn produces a ranged output using probability distribution function helpful for mapping the output to the classes.
- The CNN model mentioned in this paper, consists of different layers namely:

1) **Conv2D Layers and ReLu activation function**: The major function of the convolutional layer is to apply some filters (small matrix of weights related to the image useful to produce a feature map). The 3*3 indicates the size of the filters which is to be applied in the kernels of the convolutional layers.

2) **Maxpooling Layers:** The function of the max pooling layers is to decrease the spatial size of the image. The 2*2 indicates that a 2*2 box with a fixed ridge is traversed through the image's regions extracting the maximum values in each region.

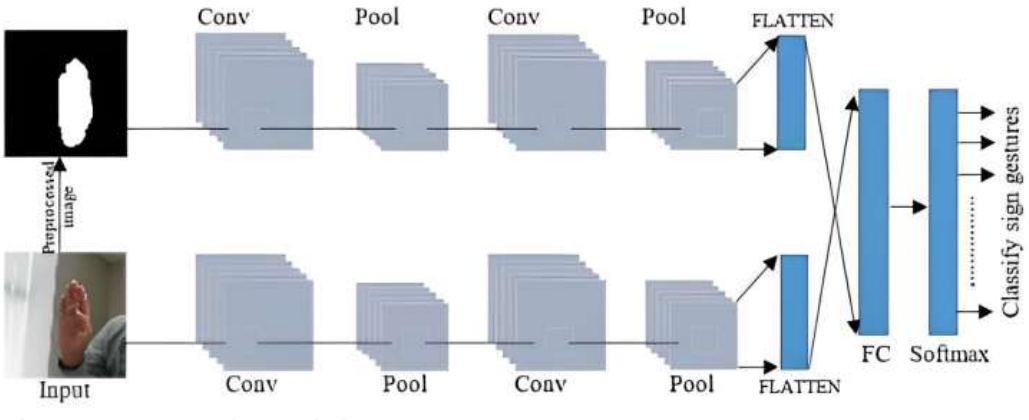
3) **Dropout Layers:** Dropout Layers were used in order to prevent overfitting (accurate and pinpoint learning of the training data. (The mentioned dropout rate is 25%)

4) **Flatten Layer**: The flatten layer as usual converts the feature maps into one-dimensional vectors only then which can be used to fed them to the dense layers.

5) **Dense Layers**: The dense layers are the fully connected layers which take the one-dimensional vectors obtained from the flatten layers and provide the classification task. (The number 128 indicates the neurons that is to be established)

Reference 3:

[7] Rahim, M. A., Shin, J., & Islam, M. R. (2019, July). Dynamic hand gesture based sign word recognition using convolutional neural network with feature fusion.



Dataset:

- The dataset used was collected through a webcam and the data collection process was done by three volunteers.
- Each volunteer's isolated gestures were considered (300 images for each gesture) making 900 images for each gesture with a single image size of 200*200 pixels.
- The dataset totally consists of 13500 images for a total of 15 isolated gestures.

Pre-processing:

- The dataset was split into 80% and 20% for training and testing sets respectively before feeding to the CNN model.
- The preprocessing involved the following techniques:

1) **YCbCr conversion:** The initial step of preprocessing was to convert the grayscale image into YCbCr format. YCbCr format consists of Y (Luminance - representing the intensity of the image) and Chrominance (Cb and Cr – representing the color components of the image).

2) **Binarization**: The image obtained is converted into binary format. This is done using a threshold value that is pixels with value below threshold are set to 0 (zero) and pixels with values above threshold values are set to 255 (white).

3) **Erosion**: This is mainly used to remove the boundaries of the foreground pixels (contributing pixels such as gestures) reducing the size of the foreground pixels and enhancing the feature extraction.

4) **Hole Filling**: There may be gaps occurring in the process of erosion. So to fill the gaps, hole filling is done which in turn helps in efficient feature extraction.

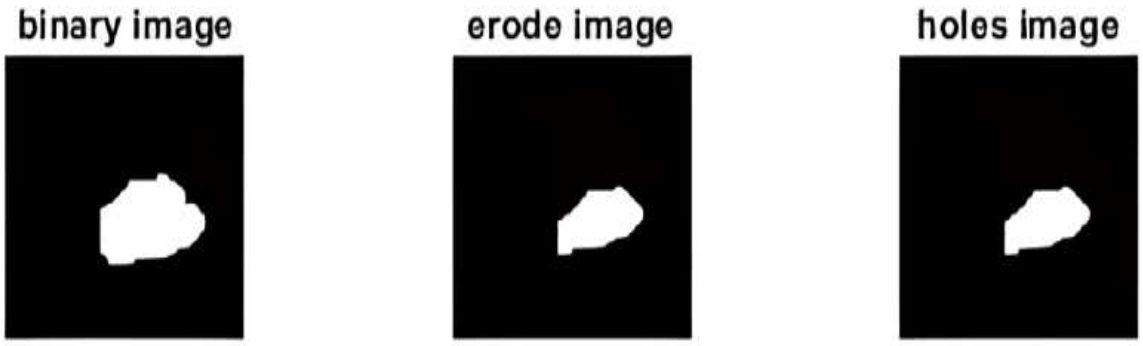


Fig 3.4

Implementation:

- The CNN model used in this paper consists of different layers such as convolutional layers, pooling layers, and fully connected layers.
- The input to the model is provided through two channels, which are the gesture image and segmented image. Two channel inputs are given in order to increase the detection of local features so that more information can be gained through the data.
- One of the channel take grayscale images as input which records the intensity of the image whereas the other channel takes YCbCr formatted images which records the color information.
- The features are then combined at the end for improving accuracy of the predictions.
- The provided CNN model consists of the following layers which perform the functions:

1) **Convolutional Layer:** This layer's major function is to apply the filters (kernels – matrices) which slide over the image performing the dot product between the kernel weights and input feature of the region. Generation of feature maps is done this way by performing the dot product between the kernel pair and local feature.

2) **Pooling Layer:** The basic function of the pooling layer is to perform transformation function on the feature maps obtained from the convolutional layer and reducing the dimensions of the image along with retaining the important features of the image.

3) **Fully Connected Layer:** After passing through the flattening layers, obtained one dimensional vectors are fed to fully connected layer which integrates all the extracted features to feed to softmax classifier.

4) **Softmax Layer:** The function of the softmax layer is known which converts the value into a range of [0,1] by applying a probability distribution function which in turn is helpful for the process of classifying the values to the nearest neuron having the nearest

RESULTS

4.1 Indian Sign Language to Speech Conversion Using Convolutional Neural Network

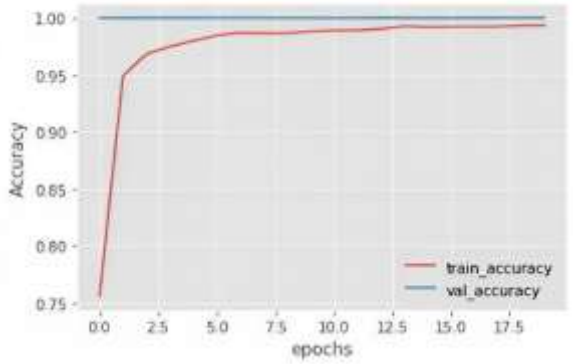


Fig 5 Accuracy Graph

- The training process has been carried out and the accuracy has been found out to be 0.99 after 13 epochs and the model steadily showed the steady increase in the accuracy before the 13th epoch.
- After the 13th epoch, the model showed a continuous and constant accuracy.

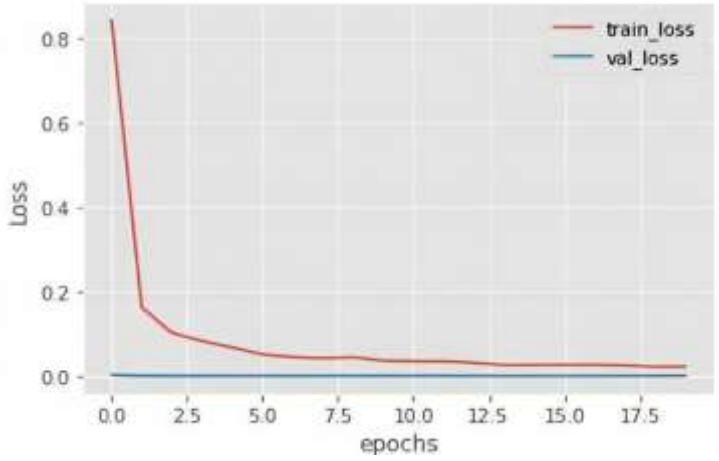


Fig 6 Losses Graph

1 The training loss during the first epoch was recorded to be 0.80 and the loss gradually decreased until it reaches the 17th epoch where the validation loss touches 0.0000000001.

2 Later and constantly the validation loss is found out to be at constant zero.

Work	Proposed Method	Accuracy Obtained (%)	Dataset
Sabeenian [8]	KNN	97%	ArSL
Yogeshwar et al.[1]	SVM	92.2%	ISL
Likhar[2]	LSTM	97.71%	ISL
Shashidhar et al[15]	2D CNN	95%	ISL
Proposed Method	CNN	99%	ISL

Fig 7 Comparison Table

4.2 Study of Gesture-Based Communication Translator by Deep Learning Technique.

```
369ms/step - loss: 0.7301 - acc: 0.7683 - val_loss: 0.1626 - val_acc: 0.9663
362ms/step - loss: 0.1913 - acc: 0.9353 - val_loss: 0.1488 - val_acc: 0.9701
361ms/step - loss: 0.1299 - acc: 0.9547 - val_loss: 0.0938 - val_acc: 0.9811
361ms/step - loss: 0.0956 - acc: 0.9671 - val_loss: 0.0817 - val_acc: 0.9905
358ms/step - loss: 0.0816 - acc: 0.9718 - val_loss: 0.1114 - val_acc: 0.9879
360ms/step - loss: 0.0653 - acc: 0.9784 - val_loss: 0.0958 - val_acc: 0.9877
360ms/step - loss: 0.0611 - acc: 0.9800 - val_loss: 0.0333 - val_acc: 0.9931
358ms/step - loss: 0.0550 - acc: 0.9815 - val_loss: 0.0403 - val_acc: 0.9917
```

Fig 8 Efficiency of the training model

- The training model was found out to have an accuracy of 99.17% and the values of the losses and accuracy values have been recorded.
- The model depicted a validation loss value of 4.03% and an overall increase if the performance of the model was observed as the size of the dataset was constantly being increased.

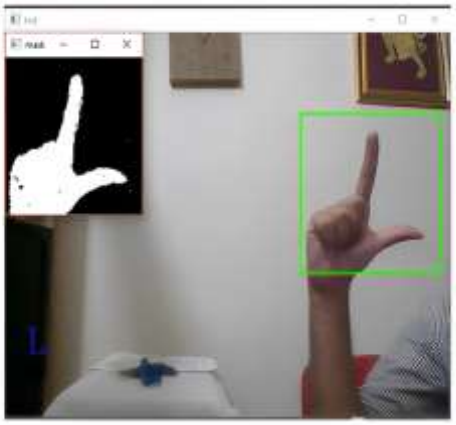


Fig 9 Recognition of letter 'L'

- With the help of the model, individual letters in the form of gestures were recognized with a very high efficiency and with the process of combining the letters recognized, several sentences were formed by the model.
- The gesture of each alphabet in the phrase was provided, recognized, and eventually stacked up to a mold of word resulting in the formation of a meaningful sentences.

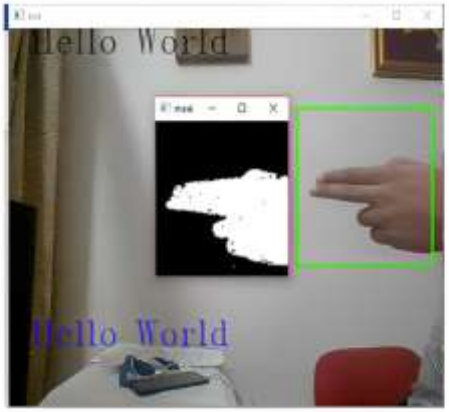


Fig 10 Formation of sentence "Hello World"

4.3 Dynamic hand gesture based sign word recognition using convolutional neural network with feature fusion.

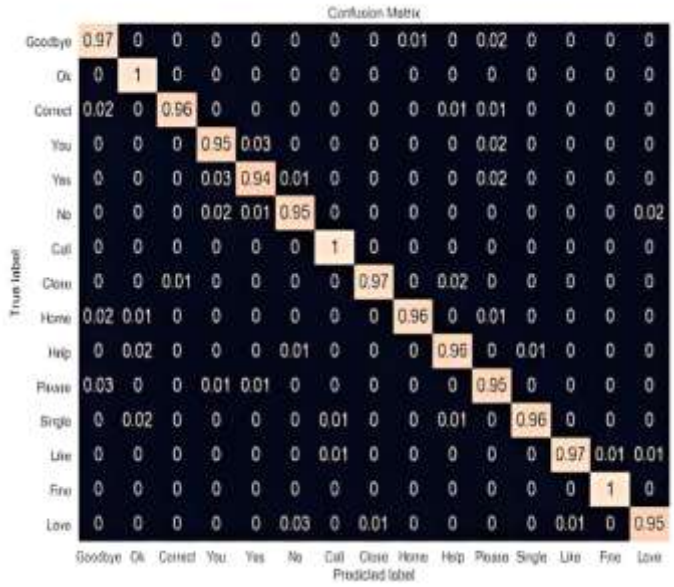


Fig 5 Confusion Matrix

- The proposed CNN model along with segmentation was claimed to be accurate on the test gestures and the true positive rate is shown as high as of in the above depicted confusion matrix.
- The True negative rates for the remaining gesture predictions was also good representing the high accuracy of the model predicting the correct gestures.

Comparison:

- The proposed methodology (Segmentation and CNN) was claimed to be more effective when compared to the other methods such as :
- DWT (Discrete Wavelet Transform) and SVM
- Pure CNN etc

Method	**Function**	**Gestures**	**Accuracy**
DWT and SVM	Hand Gestures	15 Gestures	89.67%
CNN	Hand Geatures	7 Gestures	95.96%
CNN	Hand Gestures	15 Gestures	94.22% (Proposed dataset)
Segmentation and CNN	Sign Word	15 Gestures	96.96%

CONCLUSION

The utilization of deep learning techniques such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) based technologies such as LSTM in the context of hand sign to speech conversion represents an important advancement in bridging the communication gap for deaf and non-verbal individuals.

The application of CNN for hand gestures recognition has proven effective in the terms of capturing the features of the image and then hence classifying the gestures. The layers of CNN namely Convolutional layers, Pooling layers, Flattening layers, Dense layers, fully connected layers and softmax layers have proven to a efficient and relevant set of technology for the Hand gesture recognition task.

On the other hand, Recurrent Neural Networks (RNN) have demonstrated their specialities in capturing sequential dependencies, particularly in the generation of relevant text generation from the recognized hand signs. The temporal dependencies are essentially done by the LSTM s, derived technology of Recurrent Neural Networks.

Furthermore, the integration of the text-to-speech synthesis through deep learning based methods ensure that the generated text is transformed into speech through TTS (Text-to-speech) enabling the flexible hand gesture conversion for any community regardless of the disability.

In conclusion, this study not only contributes to the academic understanding of deep learning applications in hand gesture recognition but also provides a practical and impactful solution for the deaf and non-verbal communities enabling and addressing the gap in communication among the normal and deaf non-verbal people.

REFERENCES

[1] Shashidhar, R., Hegde, S. R., Chinmaya, K., Priyesh, A., Manjunath, A. S., & Arunakumari, B. N. (2022, October). Indian Sign Language to Speech Conversion Using Convolutional Neural Network. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)* (pp. 1-5). IEEE.

[2] Agarwal, R., Bansal, S., Aggarwal, A., Garg, N., & Kochhar, A. (2021). Study of Gesture-Based Communication Translator by Deep Learning Technique. *Smart and Sustainable Intelligent Systems*, 139-150.

[3] Someshwar, D., Bhanushali, D., Chaudhari, V., & Nadkarni, S. (2020, July). Implementation of virtual assistant with sign language using deep learning and TensorFlow. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 595-600). IEEE.

[4] Pariselvam, S. (2020, July). An interaction system using speech and gesture based on CNN. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-5). IEEE.

[5] Fernandes, L., Dalvi, P., Junnarkar, A., & Bansode, M. (2020, August). Convolutional neural network based bidirectional sign language translation system. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 769-775). IEEE.

[6] Korkmaz, S. (2019). Integrated Deep Learning Structures for Hand Gesture Recognition. In *13th International Conference on Theory and Application of Fuzzy Systems and Soft Computing—ICAFS-2018 13* (pp. 129-136). Springer International Publishing.

[7] Rahim, M. A., Shin, J., & Islam, M. R. (2019, July). Dynamic hand gesture based sign word recognition using convolutional neural network with feature fusion. In *2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII)* (pp. 221-224). IEEE.

[8] Kagirov, I., Ryumin, D., & Axyonov, A. (2019, July). Method for multimodal recognition of one-handed sign language gestures through 3D convolution and LSTM neural networks. In *International Conference on Speech and Computer* (pp. 191-200). Cham: Springer International Publishing.

[9] Babu, C. G., Thungamani, M., Chandrashekhara, K. T., & Manjunath, T. N. (2018). Real Time Hand Gesture Recognition for Differently-Abled Using Deep Learning. In *RTIP2R (1)* (pp. 329-335).

[10] Roy, K., & Sahay, R. R. (2018, December). Dynamic Gesture Recognition with Pose-based CNN Features derived from videos using LSTM. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing* (pp. 1-9).