Deep Science Publishing
https://doi.org/10.70593/978-81-983916-5-0_2

DeepScience
Open Access Books

Chapter 2

Indian sign language detection and translation using deep learning

Jishnu Plavinchottil Jayaraj [1], Rajesh Kanna R [2]

[1,2] *Department of Computer Science, CHRIST University, Karnataka, India.*

Abstract: Out of India's population, about 63 million use Indian Sign Language (ISL) as the natural means of communication. However, massive barriers in communication exist between hearing-impaired people and the general population, mainly in spheres like education, healthcare, and jurisprudence, which often require professional interpreters. This language gap brings before the community of hearing-impaired several social, academic, and professional issues. The recent progress on deep learning, especially the models and architectures based on Convolutional Neural Networks (CNNs) and Transformers, have demonstrated promising results in sign language recognition. These models can be employed for significant accuracy, robustness, and better use in communication gap bridging. The project aims to develop and optimize deep learning-based sign language recognition models using the INCLUDE dataset, the standardized resource for ISL gestures. A systematic comparison and evaluation on the performance of different models will be performed on exactly the same set of data. This research therefore contributes to work on sign language recognition, pointing toward possible future solutions for a real-time translation facility and communication systems for hearing-impaired people via accurate recognition of ISL gestures. In the end, it aims at improving accessibility and promoting inclusivity in a society where communication barriers still exist for the hearing-impaired.

Keywords: Deep Learning, Gesture Recognition, Indian Sign Language, Indian Sign Language

1.1 Introduction

There are major social, educational, and professional obstacles brought on by the communication gap between the hearing-impaired group and the public. According to WHO estimates, there are approximately 63 million people in India that have auditory impairment. These people with impaired hearing and speech depend upon Indian Sign Language (ISL) as their primary mode of communication (World Health Organization, 2023).

Conversely, those who are not familiar with sign language find it very challenging to communicate. As a result, highly qualified sign language interpreters are required for both educational training sessions and crucial appointments, such as legal or medical ones. The demand for interpreters has grown over the last couple of years. In recent years, there have been significant advances in the field of sign language recognition, primarily due to the application of deep learning techniques. These models have shown impressive accuracy and robustness across a variety of tasks, showing the potential of deep learning to bridge the communication gap for individuals with hearing and speech impairments. However, a critical challenge remains: the lack of standardization in datasets. Most of the existing works apply different datasets with different levels of complexity, size, and representation, which will make it hard to hold a fair and consistent comparison of model performances.

This project addresses this limitation by designing and implementing sign recognition models using the INCLUDE dataset, a standard resource specially curated for the task at hand. This method allows for systematic, in-depth comparison of performances that helps understand the effectiveness of each model and its potential practical applicability.

The important major objectives of the project are the following:

- To comprehend the benefits of deep learning as a technique for sign language recognition.
- To analyze relevant literature to investigate different deep learning models used for sign language recognition with a focus on the Indian Sign Language.
- Develop deep learning models for sign language recognition of Indian Sign Language gestures from the same dataset.
- Assess the accuracies of the models and perform adjustments for better performance.

1.2 Literature review

In 2019, Athira et al. presented a signer independent, vision-based recognition system of Indian Sign Language (ISL) that could classify static and dynamic gestures without coarticulation by employing a removal method of co-articulation. It classified images using the SVM technique along with feature extraction from Zernike moments and motion trajectories. A total of seven signers provided approximately 900 static images and 700 videos with singlehanded dynamic gestures, double-handed static gestures, and fingerspelling alphabets. This system obtained very high recognition rates of 91 percent for finger spelling and of 89 percent for single-handed dynamic gestures-this was much more efficient and accurate than other methods (Athira et al., 2019).

In 2022, Shagun Katoch et al. have provided a system on the recognition of the alphabets and numbers in the Indian Sign Language using SURF with SVM classifier as well as CNN classifier. They have used a custom data set consisting of about 36,000 ISL signs (A to Z and 0-9), which have been acquired and then been passed through techniques for feature extraction. Such techniques include but are not limited to skin-based segmentation as well as background subtraction through running average methods. Classification-wise, SVM scored 99.17% whereas CNN was marginally better at 99.64%. Besides this feature, it also sported real-time recognition, and reverse recognition features along with a user-friendly GUI that made the text-to-sign and sign-to-speech translation effective (Katoch et al., 2022).

In 2021, Dushyant Kumar Singh brought forward a proposed model-based approach on a 3D-CNN for dynamic Indian Sign Language (ISL) hand gestures. This model handles all the temporal-spatial attributes of gestures by utilizing a convolutional neural network in video frames. The dataset consisted of 20 ISL gestures recorded by ten participants under varying lighting and orientations which added to a total of 2,400 video samples. Singh's method, in comparison to the currently available models, achieved the accuracy of 88.24% and outperformed other benchmark methods. The proposed model has excellent precision, recall, and F1 score, which show robustness under different types of gestures and conditions (Singh, 2021).

In 2023, Zhenchao Cui et al. have introduced a new end-to-end model for continuous CSLR known as the Spatial-Temporal Transformer Network (STTN). The proposed end-to-end model was designed to efficiently align low-density video sequences with high density text sequences by employing an unusual method that chunk video frames into patches to minimize computational complexity. The STTN was experimented using two publicly available datasets, namely the Chinese Sign Language (CSL) dataset and the RWTHPHOENIX-Weather multi-signer 2014 dataset, PHOENIX-2014. The results obtained reveal that STTN outperforms several state-of-the-art methods in the CSLR task with superior recognition accuracy. The comparison presented in it revealed the model enhancement of its spatial as well as the temporal features relevant for right interpretation of sign language, so establishing effectiveness over convolutional neural network-based approach. (Cui et al., 2023).

In 2024, Razieh Rastgoo et al. proposed a new Transformer-based method for improving CSLR concerning the problem of recognizing isolated signs from continuous video. Their model takes hand key points features obtained from 3 isolated sign videos, which are then enhanced using the Transformer layer before classification. The experiment included two datasets: RKSPERSAINSIGN and ASLLVD, respectively, giving average Softmax outputs of 0.99 and 0.68. Higher recognition accuracy as compared to RKS-PERSAIN SIGH was the results in showing that they had larger sample sizes. The model can be viewed as a multi-task solution since it can perform both ISLR and CSLR at the same time though false recognition presents a problem. Investigating the model with regards to real-world continuous sign videos will form a good future work. This is going to prove precious data for the research community in the future (Rastgoo et al., 2024).

In 2024, Singla et al. came up with an innovative approach to Indian Sign Language Recognition called ISLR that uses a combination of Keras, Visual Transformers (ViT), and advanced data augmentation. Their Vision Transformer model is trained on the Indian Lexicon Sign Language Dataset (INCLUDE) dataset, consisting of 1,000 images across 108 classes of ISL gestures that include hand shape variations, facial expression, and lighting. Data augmentation strategies, including ImageDataGenerator, further enhanced the generalization of the model. Optimized hyperparameters resulted in loss in evaluation of 0.2941 and 97.52% in terms of accuracy. In comparison, it is significant that the proposed model offers considerable improvement over the baseline techniques, so it has established its efficacy for the recognition of different ISL gestures and alleviates the issues of overfitting. The research also acknowledges the inclusion of Punjabi sign language, demonstrating the adaptability of the model to linguistic diversity. The future work includes developing models based on ViT for dynamic datasets to achieve performance enhancement across different sign languages and push toward real-world applications in ISLR (Singla et al., 2024; Sridhar et al., 2020).

In 2024, Zhigang Huang et al. proposed a novel dual-stage temporal perception module (DTPM) for continuous sign language recognition. In the work, this paper focused on how complex temporal information of varying temporal scales can be extracted accurately from sign language videos. Unlike earlier approaches with a fixed-size temporal receptive field, DTPM combines the benefits of both temporal convolutions and transformers in a hierarchical architecture: a multi-scale local temporal module (MSLTM) followed by global-local temporal modules (GLTMs). This captures richer temporal features by first modeling local relations and then enhancing these through global relational modules. This effectiveness is confirmed by large experiments conducted on three CSLR benchmarks, PHOENIX14, PHOENIX14-T, and CSL, where DTPM showed its capability of being able to recognize accurately sequences of glosses from sign language videos. Despite using a common visual module in the form of ResNet18, the authors clearly state that future work shall be placed on designing more powerful visual modules that contribute to the further enhancement of performance while also emphasizing collaboration of the visual and temporal modelling components (Huang et al., 2024).

In 2023, Anudyuti Ghorai et al. developed an Indian Sign Language (ISL) recognition system to fill the gap in communication between deaf and hearing individuals. The approach here utilizes a network deconvolution technique that minimizes both pixel-wise and channel-wise correlations in images, thus avoiding the problem of redundant data learning in traditional CNNs. To further improve the resilience of the model against spatial transformations, a spatial transformer network was introduced, and its integration improved the performance of the model on the ISL datasets: VUCS_ISL_I and new datasets, VUCS_ISL_II constructed besides general datasets of other sign languages like American, Arabic, and Spanish sign language. VUCS_ISL_I: comprises 26 static signs of the English alphabet, and the dataset has 2,400images in it; VUCS_ISL_II has 35 signs (1-9 & A-Z), which amounts to 4,000 images in a single sign with signs presented in multiple non-canonical poses. The proposed deep network, STN-ND-Net, presented more accurate results than current systems. Currently, it supports only static sign recognition, but authors are extending the same into video recognition of words and sentences (Ghorai et al., 2023; Nandi et al., 2022).

In 2023, G Khartheesvar et al. proposed a word recognition technique for isolated words in Indian Sign Language (ISL) using MediaPipe holistic pipeline for feature extraction with a Long Short-Term Memory (LSTM) network. The approach was tested on the INCLUDE dataset with 4,292 videos representing 263 classes, and its subset, INCLUDE-50, with 958 videos for 50 classes. The proposed approach achieved impressive accuracy rates of 94.8% and 87.4% on INCLUDE-50 and INCLUDE, respectively, along with macro averaged F1-scores of 93.5% and 86.6%, surpassing the state-of-the-art performance for both datasets. Data augmentation methods, which include cropping and generation of new key points, have the effective improvement of model robustness, particularly in countering inter-class similarities. So far, the approach remains challenging in doing proper classification of words represented by the same sign, but differing signs or performed differently by an individual signer. Future work will leverage larger ISL datasets and more advanced deep learning models, such as transformers, to further improve performance and extend capabilities into recognizing words from continuous sign language videos (Khartheesvar et al., 2023).

In 2020, Ankita Wadhawan and Parteek Kumar presented a CNN-based system that detects static signs in the Indian Sign Language (ISL) using a data set of 35,000 images representing 100 different signs. The approach obtained impressive training accuracies of 99.72% on colored images and 99.90% on gray-scale images, which surpassed the existing works focused only on fewer signs. The architecture was done by using multiple CNNs with different filter sizes that were aimed at enhancing recognition performance. Assessments that demonstrated robust precision, recall, and F-score metrics have been shown in those evaluations. Future work involves attempts to extend the scope of recognition capabilities into video datasets of dynamic signs while also developing a mobile application for real-time ISL recognition (Wadhawan & Kumar, 2020).

In 2020, T. Raghuveera et al. proposed a system that could translate Indian Sign Language (ISL) hand gestures into meaningful English text and speech, using Microsoft Kinect. The system used a dataset of 4,600 depth and RGB images of 140 unique gestures performed by 21 subjects who are involved in single-handed signs, doublehanded signs, and fingerspelling. Gestures of hands were indicated. Their accuracy could be seen with a system achieved average recognition accuracy to reach 71.85%, since an ensemble of three different Support Vector Machine classifiers which used Speeded Up Robust Features, Histograms of Oriented Gradients, and Local Binary Patterns segmented the hand region in total. For some of them, such as the ninth one or A, F, G, H, N, P, the accuracy reaches 100%. Although the system was effective, it sometimes produced wrong translations since it did not consider contexts. Future improvements would include an enlarged dataset, faster response time, dynamic updating of language dictionary, and dynamic gestures for more applications (Kothadiya et al., 2023).

In 2021, Sakshi Sharma and Sukhwinder Singh proposed a deep learning-based Sign Language Recognition System (SLRS) that aimed to improve the communication of the non-signer community by recognizing Indian Sign Language (ISL) gestures. The contributions of this study are threefold: Firstly, a new dataset of ISL with 26 static alphabet signs from 65 users of uncontrolled environments; enhancement of intraclass variance; augmentation via affine transformations. Three more copies were copied for every image in the training set. This kind of method boosts the training data and helps in making things robust. Thirdly, a Convolutional Neural Network using Depthwise Separable Convolution, referred to as CNN-DSC, has been developed for feature extraction and classification of samples. Highly accurate recognition rates of 92.43%, 88.01%, and 99.52% were realized for three different datasets using this model, which consist of a self-collected ISL dataset and the publicly available ASL dataset. From the performance measurement in terms of precision, recall, and F-score, one can see that the model has high robustness and generalizability. Furthermore, CNN-DSC shows remarkable effectiveness in handling variations of scale and size. Future work would comprise improving the real-time accuracy and the recognition capacity of static and dynamic ISL gestures (Sharma & Singh, 2021).

In 2023, Kothadiya et al. proposed a Transformer Encoder-based approach for static ISL sign recognition, which was better than traditional convolutional architectures. The vision-based recognition system uses positional embedding to split up each sign into patches. These patches are then passed through a transformer block of four self-attention layers and a multilayer perceptron to achieve accuracy of 99.29% in fewer epochs than the training. The dataset consists of RGB images over 36 classes with more than 1,000 images per class and was augmented to improve generalization. The model has been shown to be robust against various augmentations such as different angular positions and brightness. Future work in this direction would be the extension of the transformer-based approach to isolated and continuous sign recognition in video-based ISL recognition, thus advancing applications in human-computer interaction for the hearing impaired (Kothadiya et al., 2023).

In 2020, Kayo Yin and Jesse Read proposed the STMC-Transformer model for SLT. It presented significant improvement over previous RNN-based approaches on RWTH-PHOENIXWeather2014T and ASLG-PC12 datasets. The STMC-Transformer model outperformed the previous state-of-the-art systems with more than 5 and 7 BLEU points of gloss-to-text and video-to-text translations on PHOENIX dataset with improvements of more than 16 BLEU on ASLGPC12. Their results indicate that glosses are not the most optimum representations of sign language since translation of video directly to text is more efficient than from gloss to text and require reviewing the present gloss supervision regime. Further, it advised future research to include only end-to-end training on SLT without relying on gloss supervision or seek different schemes of annotation on sign language to minimize lost information (Yin & Read, 2020).

In 2024, Shetty et al. proposed a real-time sign language detection system which uses PoseNet algorithms in extracting key pose points that can be used to perform gesture recognition using LSTM models. Motivated by enhancing communication for those suffering with speech impairment, in relation to the increasing use of video conferencing through the period of the COVID-19 pandemic, the system realized an accuracy of 98%. The research used the INCLUDE dataset, which contains 4,292 videos, allowing the model to work with variations in signers' clothing and significantly reduce computation time, thus not requiring specialized hardware. This application will show text in the signer's frame during video conferences, making communication real-time. Future visions are expansions of the system to accommodate more sign languages, improved hardware usage, and importation of efficient procedures for sentence paraphrasing for improved real-time performance (Shetty et al., 2024).

The literature shows great improvements in sign language recognition with deep learning, where each paper is addressing specific challenges and datasets. Athira et al. used a custom dataset of static and dynamic gestures and showed high accuracy using SVM and feature extraction techniques. Katoch et al. used a larger custom dataset of 36,000 signs and showed superior performance with CNNs and SVMs. Likewise, Singh exploited a much smaller dataset with 3D-CNN-based gesture recognition, underlining temporal-spatial features. The others, Rastgoo and Cui et al., exploited continuous sign language recognition by means of the transformer-based models with publicly accessible datasets, namely CSL and RWTHPHOENIX, while Ghorai et al., Sharma et al., and Wadhawan et al. developed specific datasets that catered only to ISL and achieved high recognition rates for the static signs. However, the heterogeneity of the datasets used in these studies further makes it difficult to compare model performance. To overcome this limitation, our project will work towards building sign recognition models based on the standardized ISL resource that is the INCLUDE dataset. All the models are trained and evaluated on the same dataset to make it systematic and fair for comparison of performance; it gives useful insights into how effective they are and contributes to furthering standardization efforts in sign language recognition research.

1.3 Methods and materials

Figure 1.1 below describes the methodology to be used for this project. The methodology defines the different phases of the project. After the collecting the data and performing the preprocessing operations, the different models are trained, and their accuracies are evaluated before completing the comparative analysis.

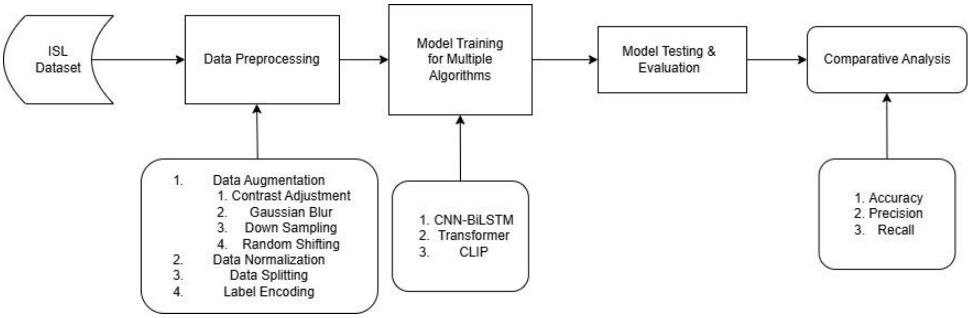

Fig. 1.1 Project Methodology

Table 1.1 Functional requirements of the project

Sr. No.	Requirement Description	MoSCoW Priority
FR1	Finding suitable data to train the deep learning models	Must
FR2	Data pre-processing before training the models	Must
FR3	Create and train the sign recognition models	Must
FR4	Evaluation and comparison the accuracies of the models	Must
FR5	Adjust the model hyperparameters to achieve higher accuracy scores	Must

Table 1.2 Functional requirements of the project

Sr. No.	Requirement Description	MoSCoW Priority
NFR1	Accuracies of the developed models are evaluated.	Must
NFR2	Code should have relevant comments and be readable.	Must
NFR3	Code needs to be error-free and consistently generate results when executed.	Must
NFR4	Code should be backed up and stored using version control.	Should
NFR5	Technology utilized in the project must be free and open source.	Must

Table 1.1 and Table 1.2 above summarize the functional and non-functional requirements of the project along with their priorities.

1.4 Results and discussions

1.4.1 Dataset & Data Augmentation

Due to low computing resources, the dataset was limited to the below signs from the INCLUDE dataset:

1. Loud
2. Quiet
3. Happy
4. Sad
5. Beautiful
6. Ugly
7. Deaf
8. Blind

As the dataset was limited to 8 signs, the dataset was much smaller with less 10 videos for some signs.

Count of Videos before Data Augmentation:

- Number of beautiful videos: 9
- Number of blind videos: 8
- Number of deaf videos: 8
- Number of happy videos: 21
- Number of loud videos: 21
- Number of quiet videos: 21
- Number of sad videos: 8
- Number of ugly videos: 8

Total number of videos: 104

Data augmentation was done synthetically increase the size of the dataset to improve model generalization.

Augmentations done to video frames:

- Contrast Adjustment – brightening frames
- Gaussian Blur – blurring frames
- Down Sampling – deleting some frames
- Random Shifting – randomly shifting frames

128 videos were generated for each sign to prevent class imbalance, which gives a total of 1024 videos.

The augmented dataset is then split into training, testing, and validation sets for model training.

1.4.2 CNN-BiLSTM Model

The EfficientNetV2S model is used as a feature extractor to process individual frames of videos into feature vectors.

These feature vectors are produced for each frame in a video and then combined to be passed to the Bidirectional LSTM sequence model for prediction.

The model is trained for 50 epochs with validation using the validation at each epoch using the validation dataset.

Figure 1.2 below shows the accuracy and loss of the training over the 50 epochs.

The accuracy graph shows that the training and validation accuracy increase consistently over the epochs and level out after the 30th epoch, which denotes the model has learned the features in the data. It can also be seen that the model is not overfitting as the validation accuracy is closely following the training accuracy.

The loss graph shows that the training and validation loss decrease steadily over the epochs, indicating that the model is improving and reducing errors in prediction. It is also observed that the model is generalizing well as the gap between the validation loss and training loss is minimal.

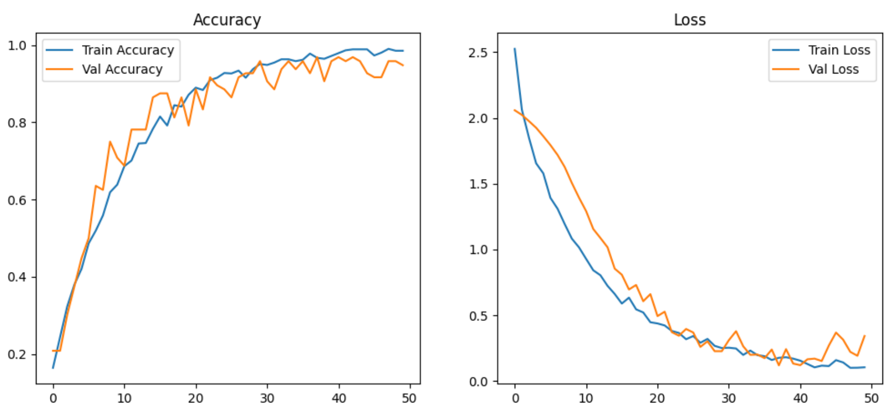

Fig. 1.2 CNN-BiLSTM model training history

The results of the model evaluation can be seen from the confusion matrix in Figure 3 and classification report in Table 1.3 below.

The performance analysis, based on Table 3 below, shows that the CNN-BiLSTM model has an accuracy of 91%, with a macro-average precision of 92% and recall of 91%. These results show that the model is highly effective in classifying sign language videos, where it can predict with high accuracy and balanced precision and recall across all classes.

The confusion matrix in Figure 3 below displays the model is performing well for classification with most labels predicted correctly. Some misclassifications have occurred, with the model confusing the signs for "beautiful" with "ugly" and "blind" with "beautiful", indicating the model has some trouble differentiating between these classes.

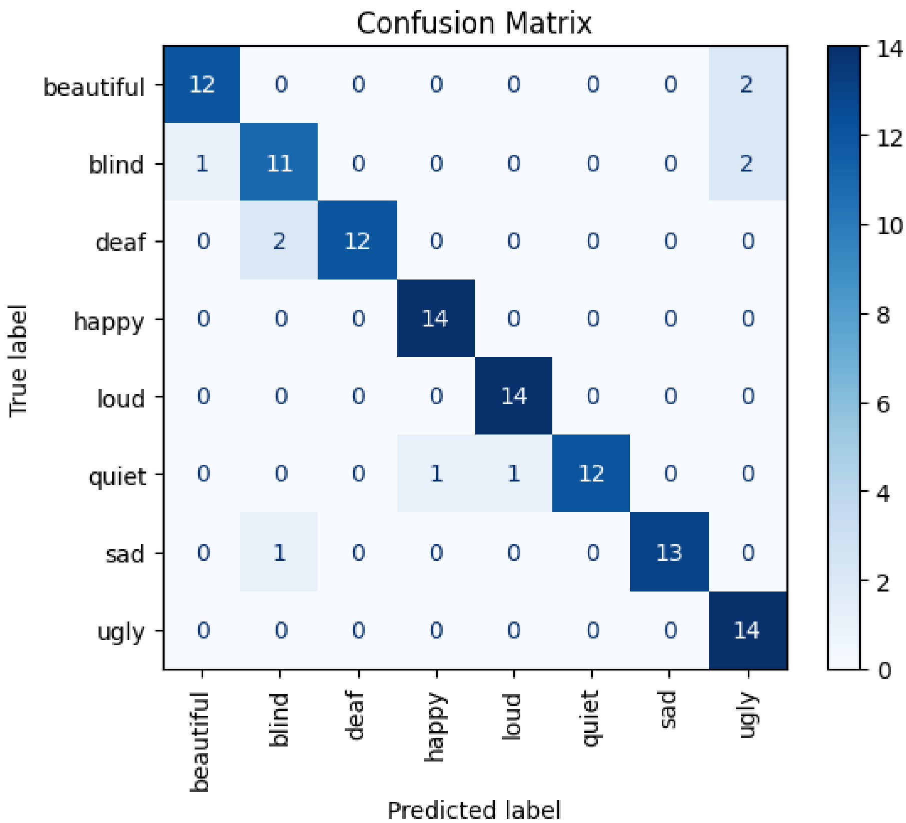

Fig. 1.2 CNN-BiLSTM Test Prediction Confusion Matrix

Table 1.3 CNN-BiLSTM Classification Report

	Precision	Recall	F1-Score	Support
Beautiful	0.92	0.86	0.89	14
Blind	0.79	0.79	0.79	14
Deaf	1.00	0.86	0.92	14
Happy	0.93	1.00	0.97	14
Loud	0.93	1.00	0.97	14
Quiet	1.00	0.86	0.92	14
Sad	1.00	0.93	0.96	14
Ugly	0.78	1.00	0.88	14
Accuracy			0.91	112
Macro Average	0.92	0.91	0.91	112
Weighted Average	0.92	0.91	0.91	112

Conclusions

In conclusion, the objective of the project was the bridging of the communication gap for the deaf by utilizing the deep learning-based models of recognition that are applied for ISL. The standardized dataset INCLUDE ensured systematically fair comparisons made between the various model performances negated the problems otherwise faced with inconsistent datasets across previous work.

This paper effectively fulfils its tasks by developing and perfecting models of accurate recognition for ISL gestures, thereby being part of greater efforts toward making research in the field standardize. The comparison-based results set a firm base for further research into ISL recognition, notably in dynamic gesture recognition and in real-time translation systems. The outcome of this research will finally pave the way for enhanced accessibility and inclusivity of hearing-impaired individuals in various societal settings.

References

Athira, P., Sruthi, C. J., & Lijiya, A. (2019). A signer independent sign language recognition with coarticulation elimination from live videos: An Indian scenario. Journal of King Saud University – Computer and Information Sciences, 34(3), 771–781.

Cui, Z., Zhang, W., Li, Z., & Wang, Z. (2023). Spatial–temporal transformer for end-to-end sign language recognition. Complex Intelligent Systems, 9(4), 4645–4656.

Ghorai, A., Nandi, U., Changdar, C., Si, T., Singh, M. M., & Mondal, J. K. (2023). Indian sign language recognition system using network deconvolution and spatial transformer network. Neural Computing and Applications, 35(28), 20889–20897.

Huang, Z., et al. (2024). Dual-stage temporal perception network for continuous sign language recognition. The Visual Computer.

Katoch, S., Singh, V., & Tiwary, U. S. (2022). Indian sign language recognition system using SURF with SVM and CNN. Array, 14, 100141.

Khartheesvar, G., Kumar, M., Yadav, A. K., & Yadav, D. (2023). Automatic Indian sign language recognition using MediaPipe holistic and LSTM network. Multimedia Tools and Applications.

Kothadiya, D. R., Bhatt, C. M., Saba, T., Rehman, A., & Bahaj, S. A. (2023). Signformer: Deep vision transformer for sign language recognition. IEEE Access, 11, 4730–4739.

Nandi, U., Ghorai, A., Singh, M. M., Changdar, C., Bhakta, S., & Pal, R. K. (2022). Indian sign language alphabet recognition system using CNN with DiffGrad optimizer and stochastic pooling. Multimedia Tools and Applications.

Rastgoo, R., Kiani, K., & Escalera, S. (2024). A transformer model for boundary detection in continuous sign language. Multimedia Tools and Applications.

Sharma S. & Singh, S. (2021). Recognition of Indian sign language (isl) using deep learning model, Wireless Personal Communications.

Shetty, S., Hirani, E., Singh, A., & Koshy, R. (2024). Gesture-to-text: A real-time Indian sign language translator with pose estimation and LSTMs. Procedia Computer Science, 235, 2684–2692.

Singh, D. K. (2021). 3D-CNN based dynamic gesture recognition for Indian sign language modeling. Procedia Computer Science, 189, 76–83.

Singla, V., Bawa, S., & Singh, J. (2024). Enhancing Indian sign language recognition through data augmentation and visual transformer. Neural Computing and Applications, 36(4), 10205–10218.

Sridhar, A., Ganesan, R. G., Kumar, P. R., & Khapra, M. M. (2020). Include. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1366–1375.

Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. Neural Computing and Applications.

World Health Organization. (2023). World hearing day 2023.

Yin, K., & Read, J. (2020). Better sign language translation with STMC-transformer. arXiv preprint arXiv:2004.00588.