International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Word recognition from Indian Sign Language using Transfer Learning Models and RNN Classifier

Naman Bansal[1] and Abhilasha Jain[2]

Submitted: 19/10/2023 **Revised**: 07/12/2023 **Accepted**: 16/12/2023

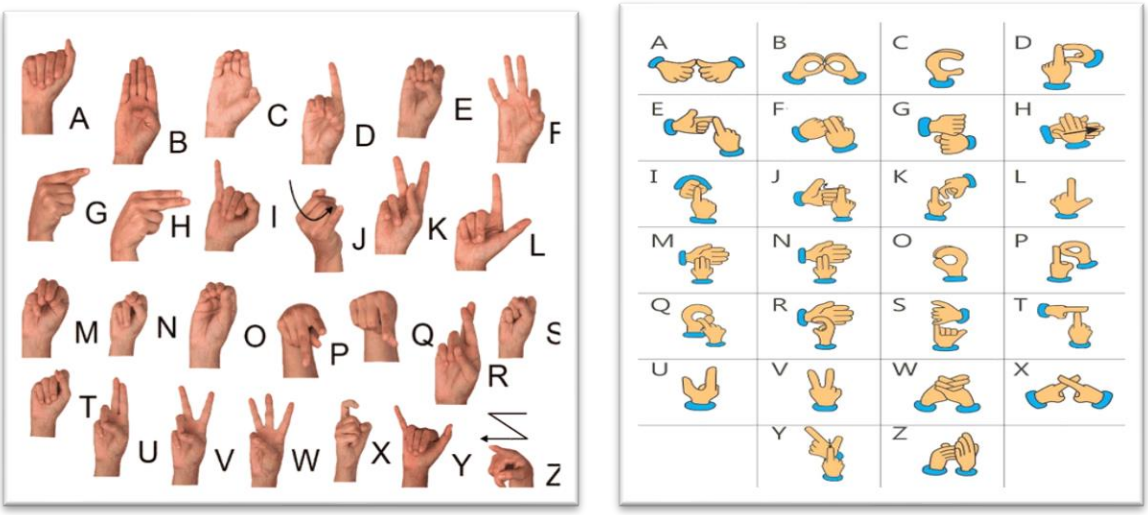
Abstract: Indian Sign Language (ISL) plays a crucial role in communication for the hearing-impaired community in India. Developing automated systems for words recognition in ISL can greatly enhance accessibility and inclusivity for individuals with hearing impairments. ISL consists of both static (Images) and dynamic (Videos) words. This paper presents a comprehensive approach to build an ISL recognition system for dynamic words. First, a representative dataset of 210 videos for three color is collected by collaborating with native sign language users, comprising a wide range of hand shapes, movements, and facial expressions. Next, Deep learning architectures, such as VGG19, InceptionV3, DenseNet121 and ResNet50 are used for feature extraction from the video frames to capture the salient characteristics of the hand movements and positions in ISL gestures. These features serve as input representations for classification model. Subsequently, a robust classifier model RNN is used for ISL recognition based on the extracted features. Accuracy of 99%, 97%, 98% and 98% is obtained for model VGG19, InceptionV3, DenseNet121 and ResNet50 respectively. The evaluation is performed on a separate test set, ensuring the generalization capability of the trained classifier. The results obtained highlight the potential of deep learning-based approaches for ISL recognition.

Keywords: *ISL, Transfer Learning, RNN classifier, video recognition*

1. Introduction

Humans are social creatures that engage with one another verbally or in writing to express their feelings and thoughts. The most effective form of communication is speech. Unfortunately, some people are not blessed with this gift. Globally, the WHO estimates that over 1 billion people have some form of disability, of which approximately 15% (or 150 million people) have difficulties with speech [1]. However, this estimate may not reflect the actual number of people with speech disabilities, as many individuals with disabilities may not have been officially diagnosed or reported. For this, an equivalent communication method has been established by human creativity and is known as "Sign Language (SL)". It is a visible language which is used by the community of deaf and hard-of-hearing,that plays a crucial role in facilitating communication and promoting inclusivity [2]. Sign Language is not Universal, similar to spoken languages, sign languages are also regionally specific around the world, with each having a vocabulary and grammar of its own. In the world, there are over 200 different SLs, such as American, Japanese, Chinese,British and Indian Sign Language (ISL)[3] which are different from each other.As shown in Figure 1, American sign language requires one hand whereas ISL requires to hand to represent English alphabets in sign language.

[1] *Ph.D. Scholar, CSE Department, MRSPTU, Bathinda*
[2] *Professor, CSE Department, MRSPTU, Bathinda*
namanbansal07@gmail.com
abhilasha_cse@mrsptu.ac.in



a) American Sign Language (ASL) b) Indian Sign Language (ISL)

Fig 1: Representation of alphabets in ASL and ISL

(Image Source: https://www.startasl.com/american-sign-language-alphabet, https://www.sounderic.com/post/indian-sign-language)

According to the 2011 Census of India [4], there were 7.5 million people with speech disabilities in India, which includes mute, difficulty in speaking, deaf and hard-of-hearing people. The SL used by deaf community of India is named as Indian Sign Language (ISL). However, the linguistic and societal barriers faced by ISL users often hinder their ability to interact effectively with the hearing world. The development of an automatic ISL recognition system (ISLR) is motivated by the need to improve communication between the deaf and hearing community. An ISLRsystem can help bridge the communication gap by allowing real-time transformation of SL to spoken or written language [5]. This can improve the quality of life for hearing-impaired people by offering with increased access to education, employment, and social interaction.Deep learning models has brought about a significant transformation in the field of SL recognition, as it has enabled the development of mostly accurate and efficient systems. By utilizing neural networks, specifically convolutional and recurrent architectures, deep learning models are capable of processing static and dynamic data depicting sign language gestures. These models can extract meaningful features from data and can translate them into either speech or text [6].

The purpose of this research paper is to examine the efficiency of different transfer learning models such as VGG19, InceptionV3, DenseNet121 and ResNet50 for video data of Indian sign language with RNN classifier. The study aims to evaluate the accuracy of this combined approach and assess its potential for practical implementation in real-world ISL recognition systems.This paper is further divided into various sections:related work is discussed in section 2, section 3 and section 4 consists methodology and results resp., conclusion is discussed in section 5 and future scope in section 6.

2. Literature Review

In the past few years, there has been an increasing interest towards the advancement of automated systemsfor Sign Language recognition (SLR), leveraging advances in computer vision, machine and deep learning techniques. Mahyoub et al. [7] investigates the appropriateness of different deep learning models for recognition and classification of words in ISL, ASL and Turkish Sign Language. They discover that there exists a strong correlation between the complexity of the models and their respective performance, with the Inflated 3D model performing the best. Zaar et al. [8] presents a high-performance deep learning architecture based on CNN for recognizing multiple sign languages, achieving high accuracy rates for ASL, Irish Sign Alphabets, and Arabic Sign Language Alphabet. Kasukurthi et al. [9] proposes a model that recognizes ASL alphabet from RGB images using a squeezenet architecture, achieving an accuracy of 83.29%. Pandian et al. [10] presents the development of multiple systems using three different models: LSTM, CNN, and YOLOv5. The real-time test results of these models were compared, and YOLOv5 achieved the highest accuracy of 97%, followed by LSTM with 94% and CNN with 66.67%. Chu et al [11] proposed a SLRsystem depends on skeleton extraction and residual network to address communication challenges faced by the hearing-impaired. The authors build an RGB video dataset of Chinese sign language,

extract human skeleton using OpenPose technology, and construct a video sequence classifier on ResNet and GRU, achieving a testing accuracy of 98% in their small-scaleSL database. Sharma et al. [12] introduced a deep learning framework that includes a convolutional neural network (CNN), connectionist temporal classification (CTC) and two bidirectional long short-term memory (Bi-LSTM) layers. This model enables the recognition of complete sentences without the need for sign boundary knowledge. Al-Hammadi [13] put forward a novel framework for the recognition of dynamic hand gestures. This framework uses various deep learning architectures to segment the hand, represent local and global features, and globalize and recognize sequence features. The system outperforms existing state-of-the-art models, demonstrating its effectiveness. Gupta et al. [14] presented a transfer-learning based user-personalization approach ISL recognitionmodel, which achieves a significantly higher average accuracy of 95.6% compared to 3% without transfer learning, demonstrating the effectiveness of the proposed approach in handling subject variability. Khan et al. [15] built a system using transfer learning with VGG architecture was able to recognize hand gestures and translate them into corresponding letters with an accuracy of 98.89 percent. Kishore et al. [16] presents the development of an ISLR System that used a wavelet-based video segmentation technique in order to recognize head movement and hand signs within videos, and elliptical Fourier descriptions to extract shape features of hand gestures, reducing the feature vectors with a recognition rate of 96%. Das et al. [17] proposed an Expert System designed specifically for the purpose of recognizing ISL. This system, known as ESISLR, uses a combination of CNN and handcrafted features to effectively and efficiently predict isolated sign words. Furthermore, the model incorporates a stacked BiLSTM network to facilitate sequence learning. The proposed approach provides commendable results, achieving an average accuracy of 93.68% utilizing Hu Moments (HM) and 94.17% using Zernike Moments (ZM).Rokade et al. [18] proposed a method that involves segmentation, binary image transformation, Euclidean distance transformation, row and column projection, feature extraction using central moments and HU's moments, and classification using artificial neural networks and SVM. Gomathi et al. [19] introduces a hybrid framework comprising Inception V3-CNN and LSTM-RNN for the purpose of recognizing dynamic hand gestures from real-time videos, with a particular emphasis on ISLR. The presented framework achieved a 96% recognition rate, outperforming similar systems used for recognizing dynamic gesture.

There are several gaps in the existing literature on ISLR systems. Firstly, there is a lack of standardized datasets specifically designed for ISL recognition. Secondly, there is limited research on real-time ISL recognition, while many ISL recognition systems have shown promising results in controlled environments, there is a need to develop practical solutions that can be easily deployed and accessible in real-world settings, such as mobile applications or assistive devices. Additionally, the majority of work to recognize Indian Sign Language is done by using static hand gestures. However, it has been addressed that the representation of most of the words in ISL is dynamic.

In this paper, we attempt to address the above gaps from the literature. Firstly, Dataset collected is according to the Indian Sign Language Research and Training Center (ISLRTC). Also, video dataset collection is done using mobile phone in real-time environment for dynamic gestures.

3. Proposed Methodology

In this section, proposed methodology is discussed which comprises of manual dataset collection in the form of videos, pre-processing of frames, feature extraction using various transfer learning models and classification as shown in the Figure 2.

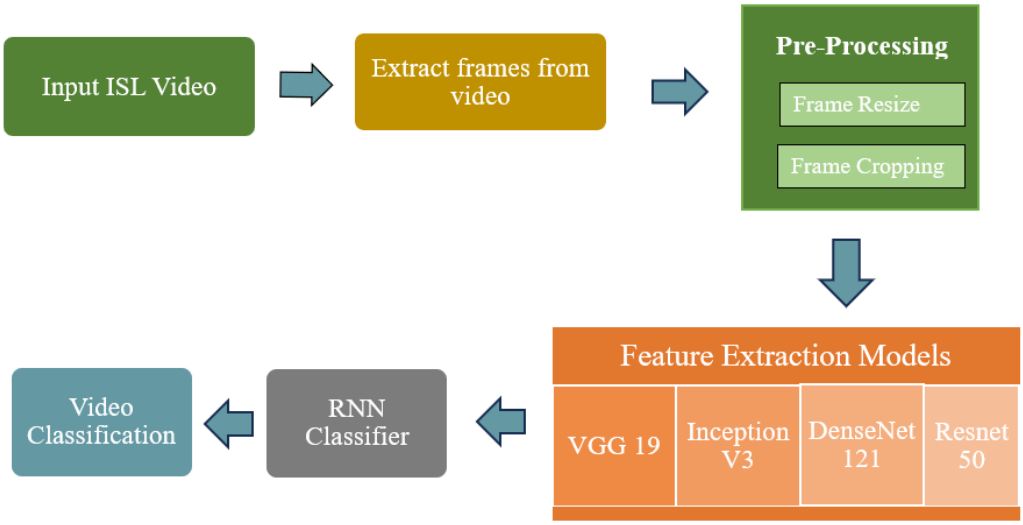


Fig 2: Proposed Methodology

Data Collection: In the absence of any benchmark dataset, a dataset of more than two hundred videos for three colors "green", "white" and "black" is collected from students of Mahant GurbantaDas School for Deaf & Dumb, Punjab under the supervision of their teachers, experts in ISL. Dynamic dataset is collected using mobile phone camera of 48MP for real time application. The video data serves as the foundation for training and evaluating the recognition system.

Pre-Processing: It includes resizing and cropping of frames in the videos. Frames are extracted from each video until a max. frame count is obtained. If the frame count of video is less than max. count, padding with zeroes is performed. Frames are cropped from centre to obtain maximum information and are resized to dimensions 224 × 224.

Feature Extraction: Transfer learning approach is utilized for feature extraction. It leverages pre-trained models, such as VGG19, InceptionV3, DenseNet121 and ResNet50. These models are trained on large datasets and high-level features are extracted.

- **VGG19:**Introduced by Simonyan and Zisserman [20], it is a deep CNN architecture containing 16 convolutional layers followed by max pooling and three fully connected layers. It achieved significant success in large-scale image recognition tasks due to its simplicity and effectiveness.
- **InceptionV3:**The utilization of inception modules in the framework involves the incorporation of parallel convolutional filters of different sizes, which facilitates the extraction of distinct features across varying spatial scales. The outputs of these branches are concatenated to form the module's output [21]. InceptionV3 achieved excellent performance on various image classification tasks and demonstrated its effectiveness in capturing both local and global patterns in the input data.
- **DenseNet121:** It consists of densely connected blocks, in which each layer is connected to every other layer directly inside the block. This dense connectivity supports reusing of features and enables efficient flow of information through the network [22]. DenseNet121 has 121 layers and has shown strong performance in image classification tasks.
- **ResNet50:** It consists of 50 layers, including residual blocks that enable the training of very deep networks. ResNet50 [23] utilizes skip connections to mitigate the vanishing gradient problem and facilitate the flow of gradients through the network.

By utilizing the knowledge learned from these pre-trained models, the paper explores the efficiency and effectiveness of feature extraction for ISL recognition.

Classification: RNN classifier is used which captures the temporal dependencies and sequence information in ISL gestures. The classifier is trained on the extracted features from the video dataset. To reduce the calculated loss during training "Adam" optimizer is used after each epoch. An RNN classifier has the advantage of being able to model and identify sequential patterns and relationships in data, which makes it ideal to use in videos of ISL gestures.

The study then presents the experimental results, focusing on the accuracy achieved by the combined approach of transfer learning models, and the RNN classifier.

4. Results and Discussion

The proposed model is implemented in Python using Keras library. The whole dataset contains 210 videos for three colours: black, green, and white in ISL signs. Dataset is split into train and test data in 8:2 ratio. A

hybrid CNN and RNN model is applied for video classification. Meaningful features of video are extracted with various CNN transfer learning model like DenseNet121, ResNet50, InceptionV3 and VGG19 respectively. Later an RNN consisting of GRU layers is applied for data classification. The model is trained by setting various hyper-parameters such as number of epochs, frame and batch size, number of frames and features extracted from videoas discussed in Table 1.

Table 1: Hyper-parameters and values set for training the model

Hyper-parameter	Value
Epoch	100
Batch-size	64
Frame size	224*224
Number of frames extracted per video	20
Number of features extracted per video/frame	Varies from 512 to 2048 depending upon feature extraction model
Optimizer	Adam

The performance of different transfer learning models is compared and evaluated on the basis of accuracy. The accuracy of the model can be evaluated by determining the ratio of accurately classified instances to the overall number of instances, as described in Equation 1.

$$Accuracy = \frac{Tp+Tn}{Tp+Fp+Fn+Tn} \quad \text{Eq. 1}$$

Here, Tp is True Positive, Fp is False Positive, Tn is True Negative and Fn is False Negative.More accuracy of a model means better performance.

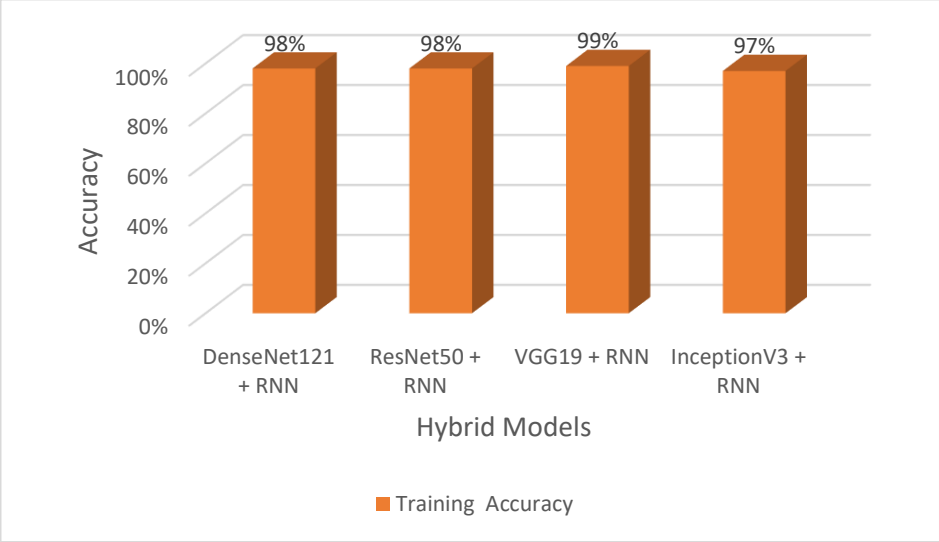


Fig 3: Training Accuracy comparison of different models

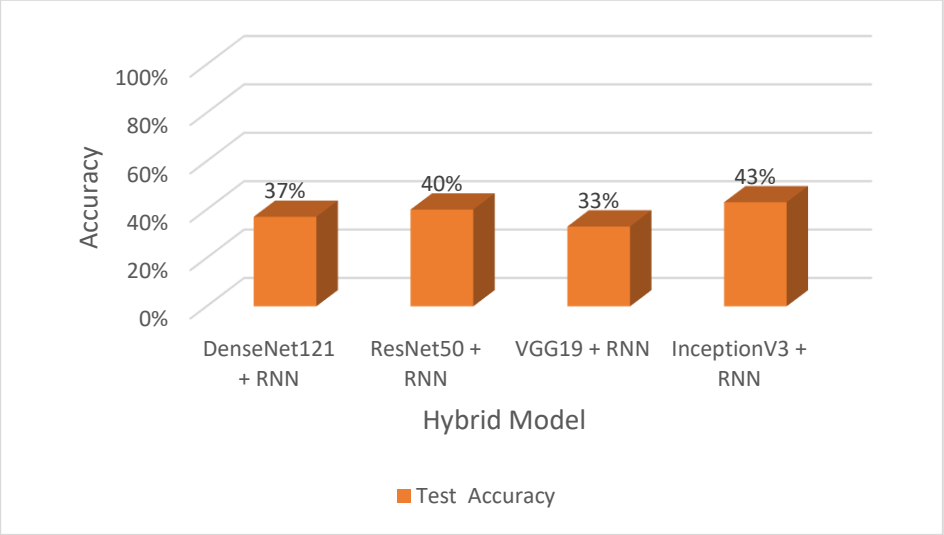


Fig 4: Test Accuracy comparison of different models

Table 2 provides a comparison of different deep learning models combined with a recurrent neural network (RNN) for classification. FromTable 2 it is observed that the number of trainable parameters for InceptionV3+RNN and ResNet50+RNN are 99,891 whereas DenseNet121+RNN and VGG19+RNN have 50,739 and 26,163 respectively, a higher number of trainable parameters might allow a model to capture more complex patterns in the data.The training accuracy represents how well each model performed on the given training dataset whereas the test accuracy is an indicator of how well the different models generalize to new, unseen data.Figure 3 and Figure 4 shows the comparison of different models on the basis accuracy. The training accuracy between different learning models is shown in Figure 3 whereas Figure 4 shows the comparison of test accuracy of models on the given ISL dataset, where InceptionV3 + RNN outperforms all other hybrid models.Selection of best model depends on the specific task, dataset, and trade-off of model complexity (number of parameters) and generalization (test accuracy). From comparison in Table2, InceptionV3 + RNN seems to strike a good balance in this context, offering a highest accuracy with comparable trainable parameters as compared to ResNet50 + RNN.

Table 2: Training Accuracy for various Hybrid Models implemented on ISL color classification

	Number of trainable parameters	Training Accuracy	Test Accuracy
DenseNet121+RNN	50,739	98%	37%
ResNet50+RNN	99,891	98%	40%
VGG19+RNN	26,163	99%	33%
InceptionV3+RNN	99,891	97%	43%

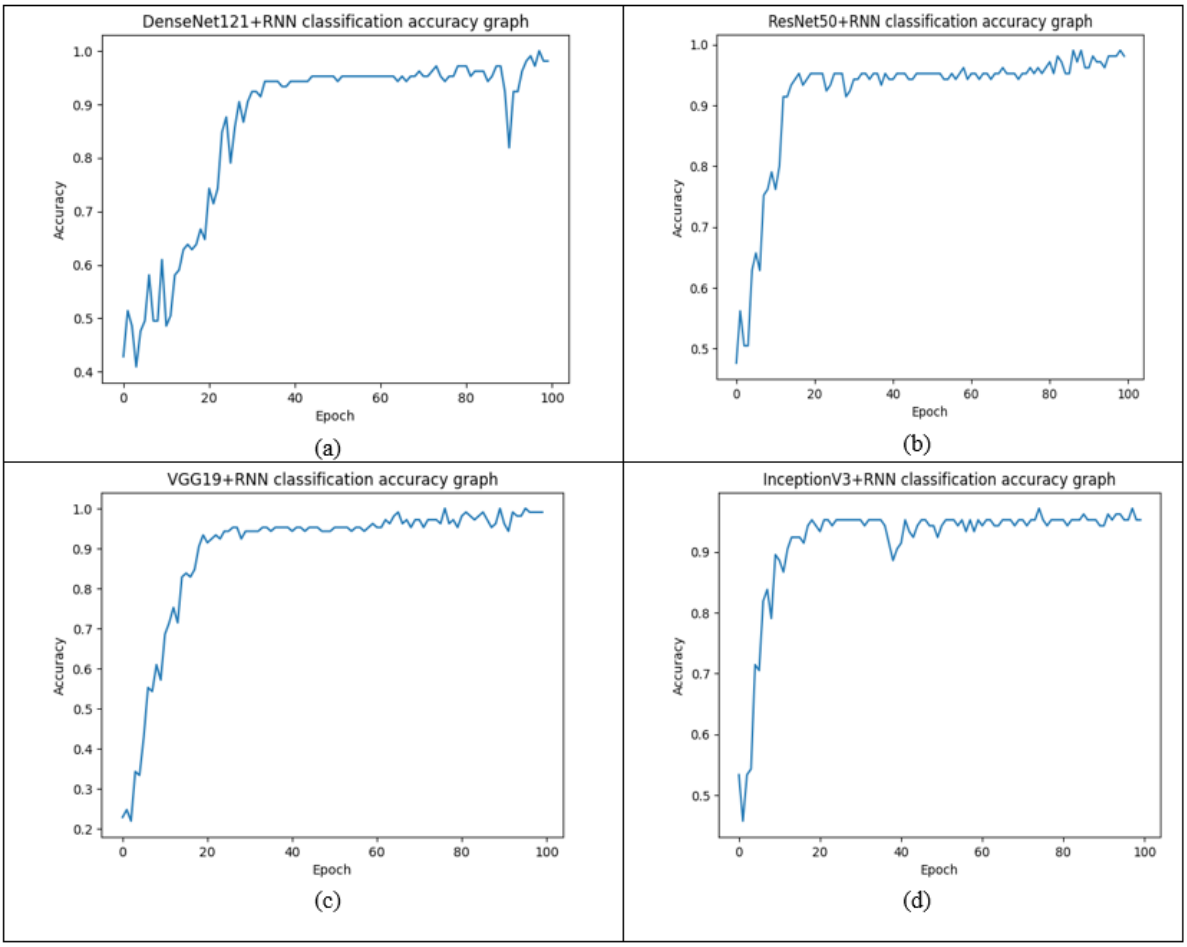


Fig 5: Accuracy graphs for training data for all the four hybrid models

Accuracy graphs obtained for training data for all the four hybrid models have presented in Figure5(a-d).It has been observed from the Figure 5, that various models when trained on same data set, VGG19 gives maximum accuracy of 99%, but model gives lowest accuracy when tested for unseen data. Therefore, VGG19 does not work well for colour classification in ISL. ResNet50 and Inception V3 models gives better accuracy when tested for unseen data as compared to other two models. But it is even needed to be improved for real time implementation of ISL recognition system.

5. Conclusion

This research paper aimed to develop a comprehensive model for ISL recognition in videos. The methodology involved dataset collection, pre-processing, feature extraction, and classification to achieve accurate recognition. The dataset collection phase involved gathering a diverse set of ISL videos, encompassing various gestures and signers to capture real-world variations. The collected dataset was then pre-processed to resize and crop the frames. Feature extraction was performed using deep learning techniques, specifically leveraging transfer learning models such as VGG19, InceptionV3, DenseNet121 and ResNet50. This allowed for the extraction of discriminative features from the videos, enabling effective representation of sign gestures. For classification, Recurrent Neural Network (RNN), was trained on the extracted features. The model was fine-tuned using appropriate hyperparameters. The accuracy of the developed system was evaluated on a testing set and is obtained as 99%, 97%, 98% and 98% for model VGG19, InceptionV3, DenseNet121 and ResNet50 respectively. The achieved accuracy demonstrated the effectiveness of the proposed methodology in accurately recognizing Indian Sign Language gestures and facilitating communication between deaf and hearing individuals.

6. Future Scope

Further research can focus on expanding the dataset, handling continuous sign language recognition and developing user-friendly interfaces. These efforts will lead to more robust and practical ISL recognition systems with improved accuracy and usability.

References

[1] https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Diwakar, S., & Basu, A. (2008). A multilingual multimedia Indian sign language dictionary tool. *IJCNLP 2008*, *57*, 8-13.

[3] Adam, R. (2015). Standardization of sign languages. *Sign Language Studies*, *15*(4), 432-445.

[4] http://censusindia.gov.in/, 2018

[5] Rajam, P. S., & Balakrishnan, G. (2010, July). Indian sign language recognition system to aid deaf-dumb people. In *2010 Second International conference on Computing, Communication and Networking Technologies* (pp. 1-9). IEEE.

[6] Sharma, S., & Singh, S. (2022). Recognition of Indian sign language (ISL) using deep learning model. *Wireless personal communications*, 1-22.

[7] Mahyoub, M., Natalia, F., Sudirman, S., & Mustafina, J. (2023, January). Sign Language Recognition using Deep Learning. In *2023 15th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 184-189). IEEE.

[8] El Zaar, A., Benaya, N., & El Allati, A. (2022). Sign language recognition: High performance deep learning approach applyied to multiple sign languages. In *E3S Web of Conferences* (Vol. 351, p. 01065). EDP Sciences.

[9] Kasukurthi, N., Rokad, B., Bidani, S., & Dennisan, D. A. (2019). American sign language alphabet recognition using deep learning. *arXiv preprint arXiv:1905.05487*.

[10] Pandian, K., Razman, M. A. M., Khairuddin, I. M., Abdullah, M. A., Ab Nasir, A. F., & Isa, W. H. M. (2023). Sign Language Recognition using Deep Learning through LSTM and CNN. *MEKATRONIKA*, *5*(1), 67-71.

[11] Chu, C., Xiao, Q., Xiao, J., & Gao, C. (2021, September). Sign Language Action Recognition System Based on Deep Learning. In *2021 5th International Conference on Automation, Control and Robots (ICACR)* (pp. 24-28). IEEE.

[12] Sharma, S., Gupta, R., & Kumar, A. (2021). Continuous sign language recognition using isolated signs data and deep transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, 1-12.

[13] Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., Alrayes, T. S., ... & Mekhtiche, M. A. (2020). Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *Ieee Access*, *8*, 192527-192542.

[14] Gupta, R., Golaya, S., & Srinivasan, R. (2022, March). Transfer-Learning Based User-Personalization of Indian Sign Language Recognition System. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 615-620). IEEE.

[15] Khan, M. D., Patro, B. S., Ranjan, R., Behera, M. C., Kumar, R., & Raj, U. (2021, September). Real-Time American Sign Language Realization Using Transfer Learning With VGG Architecture. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1-5). IEEE.

[16] Kishore, P. V. V., & Kumar, P. R. (2012). A video based Indian sign language recognition system (INSLR) using wavelet transform and fuzzy logic. *International Journal of Engineering and Technology*, *4*(5), 537.

[17] Das, S., Biswas, S. K., & Purkayastha, B. (2023, February). Indian Sign Language Recognition System for Emergency Words by Using Shape and Deep Features. In *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)* (pp. 1-6). IEEE.

[18] Rokade, Y. I., & Jadav, P. M. (2017). Indian sign language recognition system. *International Journal of engineering and Technology*, *9*(3), 189-196.

[19] Gomathi, V. (2021). Indian Sign Language Recognition through Hybrid ConvNet-LSTM Networks. *EMITTER International Journal of Engineering Technology*, *9*(1), 182-203.

[20] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[21] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[22] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).