International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Deep Learning Approach for Combined Indian Sign Language Recognition and Video Generation Model

**Prachi Pramod Waghmare*[1], Ashwini Mangesh Deshpande[1], Siddhi Dubewar [1] and Tanuja Dhaybar[1]**

**Submitted**:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

**Abstract:** To alleviate communication barriers experienced by the deaf population, this research offers a system that uses deep learning models to recognize hand positions for Indian Sign Language (ISL). Utilizing Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), VGG-16, and ResNet architectures on an exclusive dataset comprising 73 ISL videos, the suggested CNN model attains an exceptional 98% accuracy rate, signifying a noteworthy advancement in promoting inclusivity in communication for people with speech and hearing impairments. We investigate and propose a powerful combination of CNN and Generative Adversarial Networks (GANs) in artificial intelligence, with a focus on text-to-video streaming. The performance metrics as PSNR with 31.14 dB and SSIM value of 0.9916 indicate superior resolution and minimal distortion in the generated videos, affirming the GAN-CNN model's adept preservation of intricate video details.

***Keywords****: Indian Sign Language, Sign Language Recognition, Transfer learning, Video Generation, GAN*

## 1. Introduction

Sign language is an essential form of communication for deaf-mute people that allows them to convey their needs, wants, and opinions. Human body parts and gestures, including hands, faces, eyes, and gaze motions, are used in sign languages [1]. ISL recognition is the main focus of our study, and we have created word-level video datasets. Our goals are to create an accurate real-time system for identifying ISL gestures, map these gestures to text or speech, create animated avatars that mimic signed gestures, support regional variations, create an intuitive user interface, and carry out extensive testing. By using gesture detection and synthetic video creation to enable successful ISL-based communication, the study has the potential to significantly increase communication inclusiveness for the hearing-impaired community. This research aims to investigate the difficulties encountered in generating word-level video datasets linked to ISL and precisely identifying ISL gestures. We will also explore about how these gestures translate into writing and how important it is to take regional difference into account [2]. With an emphasis on hand gestures and the possibility of future extension to incorporate full-body pose gestures, the experimental results of our system's categorization of all 73 videos of ISL postures are presented [3, 4]. Before going over the methods for hand-pose translation and frame processing, the paper concludes with suggestions for future work that can further enhance communication inclusivity for the hearing-impaired community through the advancement of ISL translation. As we delve into the subsequent sections of this research, we unveil the intricacies of the YOLO architecture and its tailored application to text detection. The fusion of YOLO with the process of converting detected text regions into coherent textual content forms the core of our investigation. We evaluate the accuracy and efficiency of this approach, probing its potential in scenarios where the identification and interpretation of text within visual data are paramount. GANs have emerged as an important topic in the field of artificial intelligence, and are well known for their ability to generate reliable data, including images, text, and video. The main principle of GANs is the interaction between two neural networks (generative and discriminative) involved in the adversary learning process [5]. Designers try to produce content that is indistinguishable from real data, but judges try to distinguish between original models and created models. In this study, we specifically investigate the use of GANs in the context of text-to-video streaming using CNN. This new approach not only demonstrates the computational capabilities of GANs but also demonstrates the synergy between GANs and CNNs for complex data analysis tasks. The main idea is to convert textual descriptions into video sequences, effectively bridging the gap between verbal information and visual representations [6]. The importance of this study extends to the evaluation phase, where the videos are analyzed to ensure fidelity to the interpretation of the text. We aim to ensure the accuracy and reliability of the videos produced by examining patterns and predictive factors [7]. This involves a twofold process. The first is to convert the text into a video, and the second is to validate

[1] *Department of Electronics and Telecommunication*
*MKSSS's Cummins College of Engineering for Women, Pune, India*
*ashwini.deshpande@cumminscollege.in,*
*ORCID ID : https://orcid.org/0000-0002-9295-6602*
*siddhi.dubewar@cumminscollege.in, tanuja.dhaybar@cumminscollege.in*
** Corresponding Author: prachi.waghmare@cumminscollege.in*
*ORCID ID : https://orcid.org/0009-0002-9646-1022*

the created video against the real one so that the student can better understand the semantic content described in the text. As we explore the later parts of this research, we will discuss the complexities of GAN architecture and present the design choices and considerations that make GANs ideal for video analysis. The addition of CNNs to this framework adds complexity and sophistication, enabling students to identify complex and detailed patterns in text descriptions and video content. Through these studies, we enter not only the growing landscape of GAN applications but also the growing trend of multi-data processing, where written information is seamlessly transformed into dynamic visual representations. This research aims to overcome the limits of generation models, promote the development of multimedia content synthesis, and pave the way for various applications in fields such as entertainment, communicative communication, and education[9].The primary objective of this article is to explore an important question: how to efficiently use a deep neural network with a limited number of input videos to flawlessly combine both left-handed and right-handed signs without diminishing the accuracy of Indian Sign Language (ISL) recognition.

This article is structured in the following way: In Section 2, a concise literature review of current Sign language recognition systems is presented. Section 3 outlines the proposed approach's steps and provides a detailed explanation. Section 4 discusses the implementation scenario and the results of ISL gesture recognition using the proposed method. Finally, Section 5 offers concluding remarks and suggestions for future research.

## 2. LITERATURE REVIEW

The recognition of sign languages involves the use of technology to interpret and understand the gestures and movements used in these languages. Several methods and approaches can be broadly categorized into two types: computer vision-based methods and sensor-based methods. Computer vision-based methods use image and video processing techniques to preprocess sign language images or video frames. Feature extraction methods such as edge detection or contour analysis help identify important information for recognition algorithms. Deep learning techniques, such as CNN and recurrent neural networks (RNNs), can also be used for feature learning and sequence modeling, respectively. LSTM networks are commonly used to recognize temporal dependencies in sign language sequences.

Sensor-based methods, on the other hand, use wearable devices such as data gloves equipped with sensors to capture the movement and orientation of the signer's hands. These sensors can include accelerometers, gyroscopes, and flex sensors to record gestures. However, glove-based sensors are not practical for signers to wear in their daily activities as they restrict the movement of signers. Additionally, this setup may not be available in emergencies. In contrast, the computer vision-based approach is cost-effective and flexible, while the touch-based approach is complex, costly, and difficult to deploy.

Research on sign language recognition has led to the development of various standard datasets that are available to the public. In a study conducted by Barbhuiya et al. in 2021 [10], the authors employed a combination of CNN and Support Vector Machine (SVM) to recognize individual alphabets and numbers in American Sign Language (ASL) by focusing on isolated gestures. Similarly, in a study conducted by the author in [11], the focus was on recognizing isolated words in Arabic Sign Language (Arabic SL) using the DeepLabv3+Bi-LSTM architecture.

In another study [12], the author used Microsoft Kinect to recognize both static and dynamic signs of International Sign Language (ISL) using Multi-class Support Vector Machine (SVM). The features were extracted from 20 skeleton joints of the human body, and the author achieved 86.16% accuracy on the test data.

The author in [13] proposed an approach to recognize ISL alphabets using double-handed signs, creating a dataset of 3 dynamic signs and 60 videos. The authors used YCbCr color segmentation, Principal Curvature Based Region (PCBR) detector, Wavelet Packet Decomposition (WPD-2) methods with Dynamic Time Warping (DTW) classifier to achieve 86.3% accuracy using Support Vector Machine (SVM).

In 2021, Elakkiya et al. conducted research [14] that utilized a combination of Generative Adversarial Network (GAN), Long Short-Term Memory (LSTM), and 3D Convolutional Neural Network (3D-CNN) to recognize dynamic sign language sentences, including both German Sign Language (GSL) and American Sign Language (ASL). The primary version of the CNN model introduced by authors Chen and Koltun [15] produces images from semantic layouts. The model investigates the different loss functions and produces photographic results. The model performance bottlenecks while handling the large scale of images and adds the various intrinsic challenges

## 3. METHOD

In the pursuit of developing an effective solution for ISL recognition, the implementation process encompasses a series of structured steps to harness the potential of diverse models and techniques. The dataset, consisting of 73 videos, serves as the foundation for this research, which was collected with expert signer from a regional firm with each video encapsulating actions representing different words in ISL.

**3.1 ISL Recognition**: The recognition process involves following steps:

a. Video Processing: The first crucial step involves converting the videos into frames, facilitating a granular analysis of the visual content. This frame-by-frame representation lays the groundwork for subsequent processing and model training. The extracted frames were then subjected to keyframe extraction using a meticulous sampling approach, ensuring the capture of essential moments that convey the distinctive gestures in ISL. The extracted frames are shown in Figure. 1.

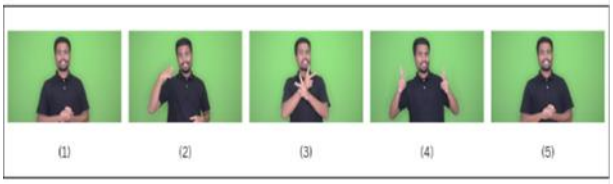


**Fig1**. Sample ISL Frames for word: clothes

*b. Data Augmentation*: Recognizing the significance of a robust dataset, data augmentation was employed to enhance the diversity and richness of the training data. By creating duplicate copies of frames, the dataset's resilience to variations in sign language gestures was significantly improved. This process contributes to the model's ability to generalize effectively across a spectrum of sign language expressions

c. Dataset Splitting: The dataset was judiciously split into training, testing, and validation sets to facilitate comprehensive model evaluation. This partitioning ensures that the models are trained on a diverse set of actions, validated on a separate subset, and tested on entirely new instances. This meticulous division lays the foundation for assessing the models' generalization capabilities.

d. Model Exploration: Four distinct models—CNN, LSTM, VGG-16, and ResNet—were systematically evaluated to ascertain their efficacy in ISL recognition. Additionally, object detection using the YOLO (You Only Look Once) architecture was explored to discern its suitability for identifying and localizing signs within the videos. After a comprehensive evaluation, CNN emerged as the most promising classification model, demonstrating superior performance across various metrics.

e. Model Training and Evaluation: The selected CNN model underwent extensive training using the augmented dataset. Hyperparameter tuning and optimization were applied to enhance the model's ability to capture intricate patterns in ISL gestures. The evaluation phase involved testing the model on the reserved test set, revealing an impressive accuracy of 98% across the 73 distinct signs. This outcome positions the CNN model as the primary solution for ISL recognition in this research endeavor. Figure 2 shows the proposed architecture to recognize the correct word.

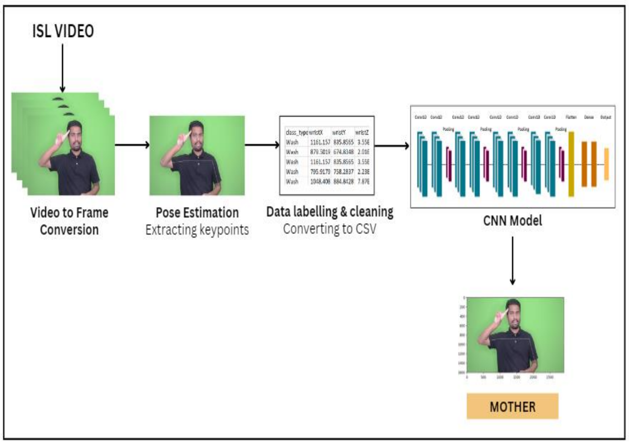


**Fig 2**. Overview of the proposed architecture

f. Object Detection with YOLO:

Parallelly, the YOLO architecture was employed for object detection within the videos, aiming to identify and localize ISL signs. This approach provides an alternative perspective on gesture recognition, with a focus on spatial localization. The results of this exploration were valuable in understanding the potential of object detection paradigms for ISL applications.

### 3.2 Video Generation

This section presents a Generative Adversarial Network (GAN) using TensorFlow and Keras, consisting of a generator and discriminator for image generation. The GAN architecture is designed for simplicity and clarity in the code, with a focus on image realism. Figure 3 shows the GAN architecture for the generation of ISL video for a particular word. The GAN network has two components of Generator and a Discriminator.

1. Generator Model: The generator's purpose is to transform a 100-dimensional random noise vector into realistic images.

Architecture: A dense layer processes the initial noise vector, followed by a series of up-sampling and causal convolutional layers, forming a reversed CNN structure.

The output layer utilizes a convolutional layer with a *tanh* activation function, generating images within the normalized range [-1, 1].

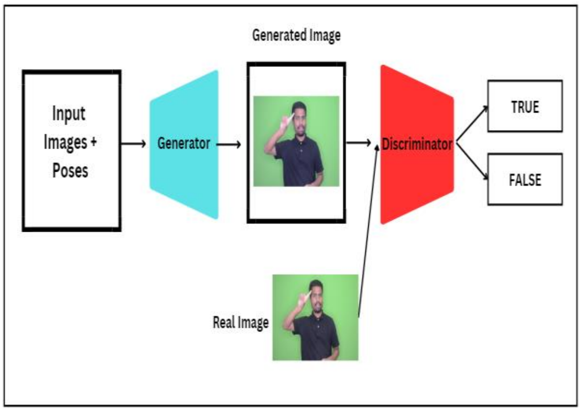


**Fig 3**. Overview of our GAN architecture

2. Discriminator Model:

The discriminator's objective is to differentiate between real and generated images.

Architecture: The discriminator follows a basic CNN structure, beginning with a convolutional layer

Additional layers can be added for increased complexity. The output layer utilizes a sigmoid activation function, providing a probability score indicating the likelihood that the input image is real. The process of training a GAN involves an adversarial approach, where the generator and discriminator continually adjust their parameters to improve their performance. The generator strives to create realistic images to mislead the discriminator, while the discriminator aims to accurately distinguish between genuine and generated images. The training procedure relies on backpropagation and optimization, allowing for weight updates for both the generator and discriminator. This foundation provides a starting point for customizing and fine-tuning the architecture and training dynamics to attain optimal results for a specific image generation task, through experimentation with hyperparameters and model configurations.

$$min_{Gen}\ max_{Dis}\ V(Gen, Dis) = \mathbb{E}_{x} \sim p_{data}[log\ Dis(x)] + \mathbb{E}_{z} \sim p_{z}(z)[log\ (1 - Dis(Gen(z)))] \quad (1)$$

In equation (1), $p_{data}$ represents the real images and $p_z$ denotes the noise vector values. The basic GAN network models in our work for generating videos are employed. The generator and discriminator networks are fine-tuned to produce photo-realistic high-quality videos [16].

## 4. RESULT AND DISCUSSION

In this section, we are presenting the results for recognition and generation in two parts.

### 4.1 Recognition evaluation with the metrics

The performance of the CNN model is also compared with three popular models used for object recognition: ResNet50, VGG16, and LSTM. Here, it is important to note that even after 25 epochs, VGG16 could not achieve an accuracy of more than 50%. The ResNet50 model achieves a higher training accuracy but could get 89% validation accuracy

**Table 1**. Parameters for Training Models

| **Parameters** | **Values** |
|---|---|
| No. of classes (ISL words) | 73 |
| Average video length (seconds) | 3 |
| Total samples after data augmentation | 1600 |
| Training samples after data augmentation | 1307 |
| Validation samples after data augmentation | 291 |

**Table 2**. Hyperparameters configuration for performance comparison between models

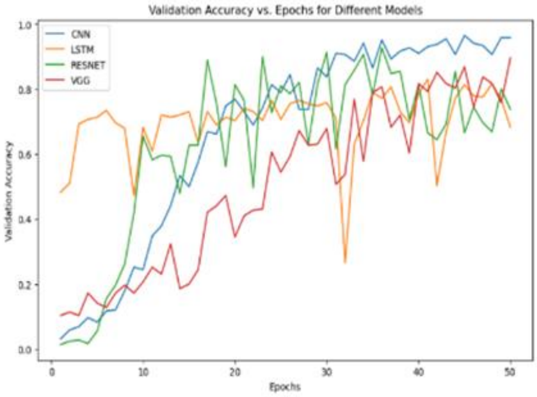
| **Model configuration** | **Parameters** |
|---|---|
| Trainable layers | 16 |
| Learning rate | 0.001 |
| Dropout | 0.2 |
| Batch size | 32 |
| Activation function | ReLU |
| Optimization algorithm | Adam |

**Table 3.** Performance comparison of classification models

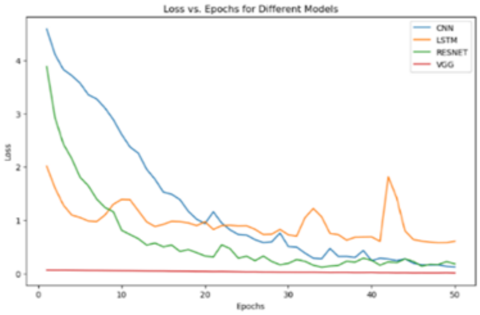
| **Parameters** | **CNN** | **RESNET -50** | **VGG-16** | **LSTM** |
|---|---|---|---|---|
| Total no. of Layers | 16 | 26 | 14 | 4 |
| Training Accuracy (%) | 98.01 | 94.26 | 86.84 | 79.04 |

| | | | | |
|---|---|---|---|---|
| Validation Accuracy(%) | 95.86 | 73.79 | 89.66 | 68.28 |
| Precision | 0.97 | 0.77 | 0.91 | 0.88 |
| Recall | 0.96 | 0.72 | 0.87 | 0.60 |
| F1-score | 0.97 | 0.73 | 0.89 | 0.71 |
| Training loss | 0.1724 | 0.1788 | 0.0099 | 0.6053 |
| Validation loss | 0.6065 | 1.5837 | 0.0099 | 1.3663 |

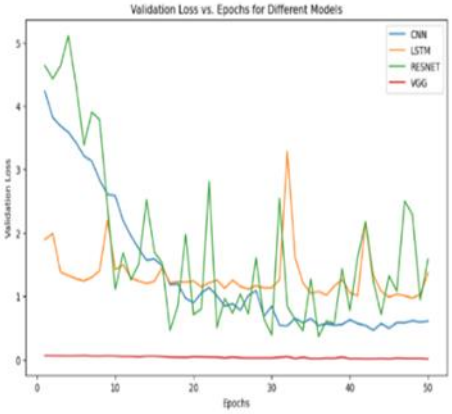
The overall performance of the proposed approach on our dataset is summarized in Figure 2 and Figure 3.



(a)



(b)



(c)

**Fig 4.** (a) Validation accuracy, (b) Training loss, (c) Validation loss

For the calculation of precision, recall, and F1-score, the following formulas are used [16]. The comparison, from Table 3 and Figure 4 reflects that CNN model outperforms the recognition task as compared with all other models for our dataset. Table 1 and Table 2 indicate the hyperparameters used for training and validation of the models. Figure 5 shows the performance comparison for CNN method with other methods.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

**Table 4:** Video Generation with GAN evaluation metrics

| **Parameters** | **Words generation** |
|---|---|
| PSNR (dB) | 31.1323 |
| SSIM | 0.9966 |
| Loss after 20 epochs | 0.164 |

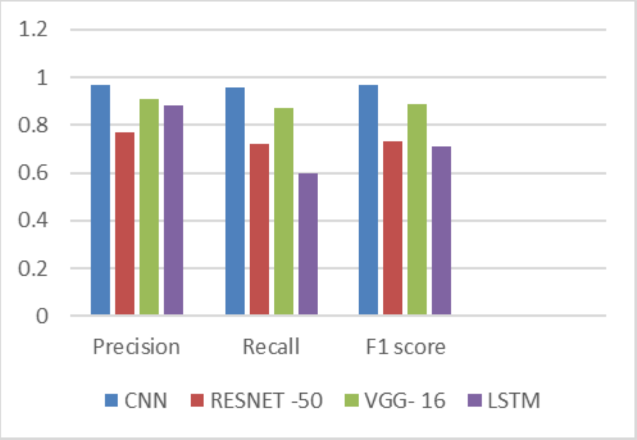


**Fig. 5** Performance comparison of proposed approach with other models

### 4.2 Video Generation Using GAN

The generated video quality is evaluated using the PSNR quality metric. It compares the quality of generated results using ground truth images and provides the score. The higher PSNR value indicates improved quality in generated results. The Structural Similarity Index Measure (SSIM) metric [17] is used for assessing the image quality. We use the SSIM metric for comparing the model performance with existing approaches. This metric assesses the structural information degradation of generated video frames. The results for generation are presented in Table4 with evaluation metrics.

$$PSNR(gt, gen) = 10log_{10}\left(\frac{255^2}{MSE(gt,gen)}\right)$$

(5)

$$MSE(gt, gen) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(gt_{ij} - gen_{ij})^2$$

(6)

Structural Similarity Index Measure (SSIM) (7)

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

### Conclusion

The system presented in this paper can accurately track hand gestures using techniques such as object detection and keyframe extraction. After training four individual models on the data set, an evaluation of their predictive ability was performed. A comprehensive comparison of performance measures showed that the CNN model showed the best accuracy, reaching 98%. These results confirm CNN as the most effective model among the tested alternatives, demonstrating its ability to capture complex patterns in datasets and generate perfect predictions. This finding highlights the importance of model selection for optimal performance in any application and demonstrates the importance of CNN for accurate and reliable predictions in any situation. This system provides more accurate and faster sign language recognition than other methods discussed in the literature. The system presented in this paper can be extended to other sign languages if the data set is suitable.

Performance evaluation metrics, including PSNR and SSIM provide a comprehensive evaluation of model performance. A PSNR value of up to 31.14 means that the generated videos have good resolution and less distortion than the original images. This measure is important to determine the quality of the generated videos, and the GAN CNN model showed the correct preservation of video details. In conclusion, our GAN CNN model shows excellent performance in video reconstruction and lays a solid foundation for a variety of applications

### Author contributions

**PW:**Writing, Data curation, Preparation, Validation; **AD:** Conceptualization, Methodology, Software, Investigation, Writing-Reviewing and Editing. **SD:** Writing-Original draft, Visualization, Investigation, Software; **TD:** Writing-Original draft, Visualization, Investigation, Software.

### Conflicts of interest

The authors declare no conflicts of interest.

### References

[1] K. Mistree, D. Thakorand B. Bhatt, “Towards Indian Sign Language Sentence Recognition using INSIGNVID: Indian Sign Language Video Dataset”, vol. 12, no. 8, Jan. 2021, doi: 10.14569/IJACSA.2021.0120881

[2] K. Shenoy, T. Dastane, V. Raoand D. Vyavaharkar, “Real-time Indian Sign Language (ISL) Recognition,” IEEE, Jul. 2018. doi: 10.1109/ICCCNT.2018.8493808.

[3] R. A. Muhamad and M. Husni, “The Recognition of American Sign Language Using CNN with Hand Keypoint”, vol. 9, no. 2, pp. 86–95, Dec. 2023.

[4] D. Kothadiya, C. Bhatt, K. Sapariya, K. R. Patel, A.-B. Gil-Gonzálezand J. M. Corchado, “Deepsign: Sign Language Detection and Recognition Using Deep Learning”, vol. 11, no. 11, Jun. 2022, doi: 10.3390/electronics11111780.

[5] B. Natarajan and R. Elakkiya, “Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks”, pp. 13153–13175, Dec. 2022.

[6] S. Krishna and J. Ukey, “GAN based Indian Sign Language synthesis”, Computer Vision, Graphics and Image Processing, Dec. 2021.

[7] Zhang, Fan, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. "Mediapipe hands: On-device real-time hand tracking." *arXiv preprint arXiv:2006.10214* 2020.

[8] Mekala, P.; Gao, Y.; Fan, J.; Davari, “A. Real-time sign language recognition based on neural network architecture”e. In Proceedings of the IEEE 43rd Southeastern Symposium on System Theory, Auburn, AL, USA, 14–16 March 2011.

[9] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu,"Generative adversarial networks." arXiv preprint arXiv:1406.2661 2014.

[10] Abbas, R. K. Karshand R. Jain, “CNN based feature extraction and classification for sign language”, vol.

80, no. 2, Jan. 2021, doi: 10.1007/S11042-020-09829-Y.

[11] Aly, Salehand W. Aly, “A "DeepArSLR: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition pp. 83199–83212 2020, doi: 10.1109/ACCESS.2020.2990699.

[12] K. Mehrotra, A. Godboleand S. Belhe, “Indian sign language recognition using kinect sensor.”, in Image Analysis and Recognition, Springer International Publishing, Jan2015[https://doi.org/10.1007/978-3-319-20801-5_59](https://doi.org/10.1007/978-3-319-20801-5_59).

*[13]*J. Rekha, J. Bhattacharya and S. Majumder, “Shape, texture and local movement hand gesture features for Indian Sign *Language recognition”, IEEE, Dec. 2011. doi: 10.1109/TISC.2011.6169079.*

[14] R. Elakkiya, P. Vijayakumar, N. Kumarand N. Kumar, “An optimized Generative Adversarial Network based continuous sign language classification”, vol. 182, Nov. 2021, doi: 10.1016/J.ESWA.2021.115276.

[15] Q. Chen and V. Koltun, “Photographic Image Synthesis with Cascaded Refinement Networks”, IEEE Computer Society, Jul. 2017. doi: 10.1109/ICCV.2017.168.

[16] Powers, David. (2008) “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”, ArXiv, abs/2010.16061.

[17] Z. Wang, A. C. Bovik, H. R. Sheikhand E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity”, IEEE transactions on image processing vol. 13, no. 4, Apr. 2004, doi: 10.1109/TIP.2003.819861.