Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766107

# EMOJI HAND GESTURE RECOGNITION WITH DEEP LEARNING

**[1]Vandan Bhatt,[2]Vipul Vekariya, [3]Ankita Gandhi, [4]Anil Patel**
[*] Computer Engineering Department, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India
*[1)] *Corresponding author: vandanbhatt007@gmail.com*
[2)] *vipul.vekariya18435@paruluniversity.ac.in*
[3)] *ankita.gandhi@paruluniversity.ac.in*
[4)] *anilkumar.patel2986@paruluniversity.ac.in*

## ABSTRACT

TheSign-Language is to help people with hearing or speaking disabilities who cannot communicate well with other people. The study proposed to help people with such disabilities using Emoji Sign-Language with the corresponding hand gesture. Deaf individuals will be able to communicate or interact with other people conveniently. The study proposed emoji hand gesture detection trained by using the YOLOv8 algorithm that aims to detect hand gestures that can recognize their equivalent hand emojis. The study tools such as LabelImg for annotating the data set, categorizing each image of hand gestures based on their equivalent Hand emojis.
**Keywords:** hand sign language, hand emojis, object detection, yolov8, CNN.

## I. INTRODUCTION

It [1] might be challenging to communicate with deaf individuals if you don't interpret hand gestures. The study that is being proposed uses ASL and its corresponding hand gestures as a means of resolving this problem. Deaf people can interact with others more easily if these HSL or hand gestures are detected. The suggested fix involves training an HSL or hand gesture detection to detect and recognize the corresponding letter alphabet for each hand gesture and use the YOLOv3 algorithm. This approach can provide a practical solution for people with hearing disabilities to interact with others who do not know HSL, enhancing their ability to participate in daily interactions and activities. In conclusion, the proposed system has the potential to bridge the interaction gap and provide a more inclusive environment for individuals with hearing disabilities.

As [2] the demand for human-computer interaction interfaces grows, the need for accurate visual hand-gesture recognition is becoming more critical. This research presents a solution to this problem by immediately classifying hand motions in photos using a deep convolutional neural network without any categorization or recognition stage that could remove unnecessary portions. With the suggested method, the network can immediately learn the characteristics of hand motions in the images rather than relying on manual segmentation or detection of the hand region. This method can boost the effectiveness and precision of hand-gesture identification in practical applications, as it avoids the need for additional preprocessing steps and reduces the chances of discarding relevant information. Finally, the suggested technique offers an effective response to the difficulties in visual hand-gesture recognition, which can enhance the development of human-computer interaction interfaces.

The [3] development of a system that recognizes sign language might greatly enhance the daily lives of those with hearing impairments. The main goal of this research is to create a method that, even in uncertain situations, can accurately categorize motions in Indian Sign Language (ISL) using static pictures. The article introduces a revolutionary method that divides motions into single-handed and double-handed movements. This strategy might enhance the effectiveness and precision of sign language identification by enabling the system to differentiate between various gestures and enhancing the recognition of intricate hand movements. By successfully classifying gestures in the ISL under ambiguous conditions, this approach can help to bridge the communication gap between people with hearing disabilities and those who do not know sign language. In conclusion, the proposed method provides a practical solution to the challenges of sign language recognition and has the potential to significantly enhance the daily lives of people with hearing disabilities.

SLR [4] has the potential to bridge the communication gap between people who are deaf and those who do not know sign language, thus enabling better communication between them. With the help of SLR technology, sign language users can communicate more effectively with the rest of the world, providing them with greater independence and opportunities in their personal and professional lives. Therefore, developing reliable and accurate SLR systems is crucial for facilitating better communication and inclusion of the deaf community.

Hand gestures offer a natural and intuitive interaction between humans and robots. Recently, vision-based dynamic HGR has gained popularity as a research topic due to its potential applications. The ability to recognize hand gestures is a highly sought-after feature in cv as it makes human-computer interaction more flexible and convenient. However, it is crucial to highlight that any research in this area has limits and will require more refinement before being implemented in commercial settings. Despite this, some study has identified sign language, which can be particularly helpful for the community of the deaf and speech loss. However, the high cost of commercialization of such technology remains a major hurdle to its widespread adoption. Sign language is a highly structured language system in which every gesture has a specific meaning assigned to it [5][6][7][8].

In [9] Models for hand gesture identification based on deep learning are created to precisely forecast the emergency indicators of ISL. The videos for eight different emergencies are included in the collection. Three separate models are supplied with the number of frames taken from the videos. After annotating the structures, one object detection model is used, and two classification models are created. 3D CNN makes up the first model, and a pre-trained VGG-16 and RNN-LSTM make up the second. The final model is based on the cutting-edge object identification technique YOLOv5.

In [10] you can find a video dataset of terms used in emergencies in ISL. The films of 26 individuals (including 14 women and 12 men) at the ages of 22 and 26 contained eight ISL terms. Two samples were provided by each subject, which were collected indoors in environs with standard lighting. Such a video library is essential for the self-regulating identification of tension situations from sign language for the interest of the deaf.

In [11] binary representation of the categorical data, values are achieved using a one-hot encoding technique. The application of a CSA for choosing the best hyper-parameter for dataset train using CNN comes next. The extraneous factors are removed from reflection,

improving the hand motions' classification accuracy. The model produces Hundred % training and testing accuracy, supporting its advantage over other current-generation models.

In [12] explains the use of a CNN combined with spatial pyramid pooling (SPP) to recognize hand gestures using vision. By stacking multiple levels of pooling to increase the number of features fed into a fully connected layer, SPP is seen to mitigate the issue with conventional pooling. SPP also produces a fixed-length feature representation when given inputs of various sizes.

In [13] contrast, Convolutional neural networks (CNN), for example, use supervised learning to adapt to a variety of obstacles. However, It depends on the volume and collection of training data as well as the CNN's design to achieve acceptable generalization on unseen data. The EDenseNet is a particular network design that is recommended to recognize hand movements using vision. By deploying a bottleneck layer to transmit features that are reused to all convolution layers in a bottleneck fashion, the altered transitional layer in EDenseNet considerably enhances feature broadcasting. The subsequent Conv layer is then continuous out the undesirable features. The ablation study examines its performance gains and identifies differences between EDenseNet and DenseNet.

In [14] proposes an efficient HGR system for webcam-based, inexpensive color video. In the suggested model, Deep CNN(DCNN) is employed to extract effective hand characteristics for ASL recognition through hand gestures.

In [15] Convolutional neural network (CNN) developer platforms on deep learning are primarily intended for gesture-based SLR. When compared to other CNN designs currently in use, this model's compact representation delivers superior accuracy of classification with fewer model parameters. To assess the effectiveness of this approach, VGG-11 and VGG-16 were trained and assessed in this study.

In [16] implemented to transfer learning to combat the subject variation a current user brings. A DL model is pre-trained on a group of current users to categorize ISL indicators from time-series data. A PSLR system is matured for the current user by retraining the model using a small sample of cases from the more recent user to categorize the signs they execute. They were using a base model that has already been trained, cutting down on time and instances needed from a new user while keeping the model accurate.

In [17] Research focuses on the potential application of DL to address the problem of HGR in a library of emergency films. Several frames from the videos were taken and fed to the model. The model consists of an RNN-LSTM and a pre-trained VGG-16 (RNN-LSTM).

In [18] difficult problem is the acknowledgment of static hand gestures in complicated backgrounds with variable lighting. K-means clustering is employed to remove the background from depth images to solve this problem, and the segmented RGB-D data is fed into the suggested CNN classification network. The results show that segmented RGB-D can completely recognize a variety of hand gestures.

In [19] using both local hand form features and global body configuration features, this method is particularly successful at conveying complicated, organized hand movements in sign language. In this study, the open pose framework was used for hand region estimate and recognition. For the forecast and normalization of gesture space, the body parts ratios theory and a trustworthy face identification technique were both applied. Two distinct 3DCNN

instances were utilized to learn the coarse-grained characteristics of the overall body construct & the fine-grained features of the hand shape.
In [20] suggested a powerful deep CNN method for hand gesture identification. To overcome the lack of an extensive labelled hand gesture dataset, the suggested method used transfer learning. Additionally, the temporal dimension of hand gesture sample data was normalized using linear sampling. We employed the ratios between the lengths of the observed human body parts and faces for spatial dimension normalization. Then, we employed 3DCNN in two methods for feature learning. The initial approach trained a single 3DCNN instance to extract HG data from the full video. To remove the hand gesture features from the beginning, middle, and end of the video sample, three examples of the 3DCNN framework were incorporated in the second procedure.

## II. METHODOLOGY

The goal of this research is to develop a system that recognizes hand emoji movements quickly & accurately using a pre-existing model, specifically YOLOv5, YOLOv6, YOLOv7, and YOLOv8. Figure 1 shows the researchers' actions to build the system and accomplish what they wanted. The initial stage was to generate the dataset, which was manually annotated and utilized for training and validation. A model was created and assessed for accuracy and precision during the training phase. Finally, the model that performed the best in terms of accuracy and precision was picked and put to the test using pictures of hand emoji movements.

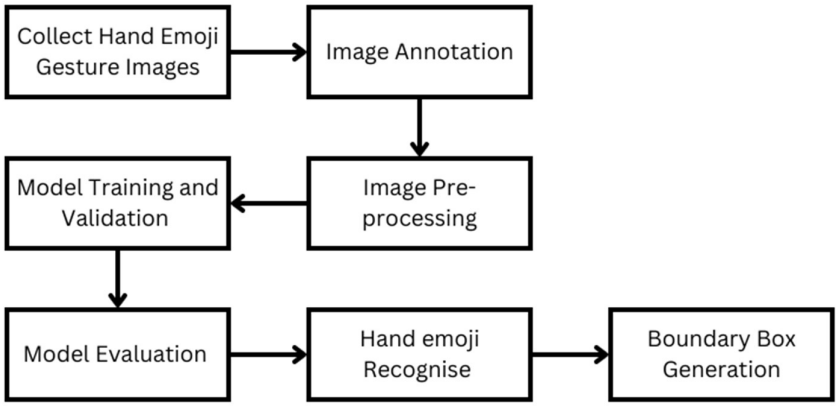


**Figure. 1:** Block Diagram

## III. IMAGE ANNOTATION & PRE-PROCESSING

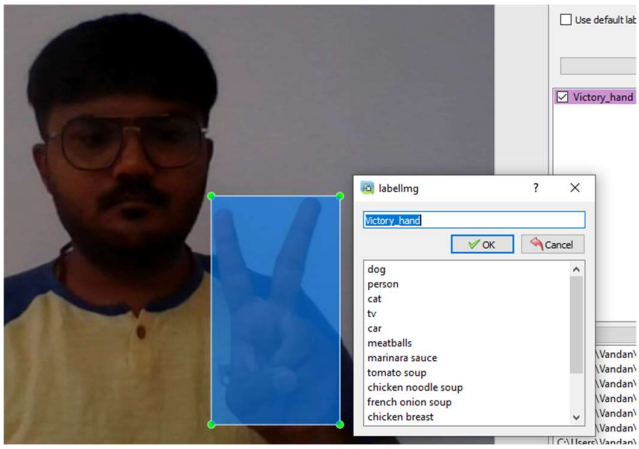


**Figure. 2:** Image Annotating using LabelImg

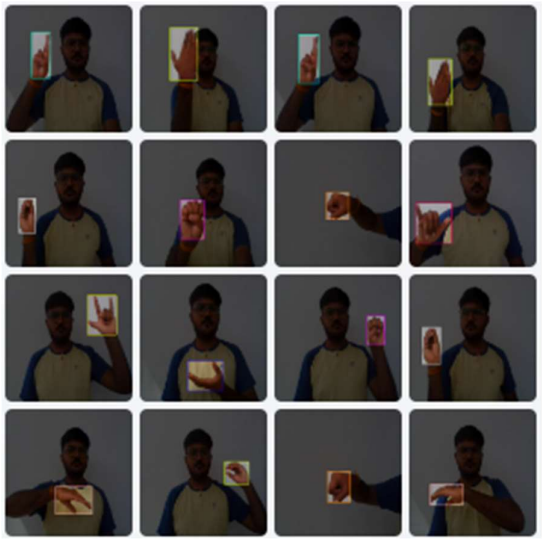

**Figure. 3:** Hand emoji Datasets

The bounding box coordinates (XMin, XMax, YMin, YMax) of the identified HG in the images are stored in the XML file that comes with each annotated image. After all, the pictures have been annotated, the labeled images of hand emojis are uploaded to the Roboflow website to obtain various label formats, such as JSON, TXT, and CSV, to ensure compatibility with the specific requirements of multiple models. This is necessary since each model requires the labels to be in a particular format.

To prepare the data for the study, the researchers obtained a dataset of 1114 images featuring 21 different hand emojis gestures. Each gesture had between 50 to 70 images, as shown in Figure 3. The dataset was collected from ML and online data science communities, where users can publish and discover datasets and tools for data science. To ensure the data was well-suited for the study's needs, As shown in Figure 2, the pictures were conventionally labeled in Pascal VOC format but used the LabelImg program. The dataset was subsequently split into three

subsets: 70% (779 images) for train, 10% (112 images) for testing, & 20% (223 images) for validation.

## IV. DATASET TRAINING AND VALIDATION AND MODEL EVALUATION

To accomplish the desired result, the study thoroughly trained and validated the datasets using YOLOv5, YOLOv6, YOLOv7, and YOLOv8. This approach helped to ensure that the final model used in the study was the most accurate and efficient. During training, the study ensured that the models' weights were adjusted based on the predicted probabilities to generate the most accurate bounding boxes.

The study utilized Google Colaboratory, a free online platform that allows users to write and run code in a Jupyter Notebook-style interface to optimize the training process. Using this platform allowed the study to reduce the time required for training and validation, ultimately making the project more efficient.

The study kept track of training and validation losses throughout the training process to evaluate the models' performance. By doing so, the study could identify models that were overfitting the training data and select the best model to use in the final stage of the project.

The study aimed to improve the accuracy and efficiency of the system while reducing the time needed for training and validation. By training the datasets for 25 epochs, the study achieved an efficient and accurate model within just 12 minutes. The model generated after training was analyzed based on mean average precision, and the best model was selected for the study.

In conclusion, the study's use of pre-trained models, Google Colaboratory, and thorough training and validation helped ensure that the final model used was accurate, efficient, and capable of detecting hand emoji gestures with high precision.

## V. DETECTION MODEL TO TESTING

Testing the final detection by YOLOv8 involves evaluating its performance on a separate set of images or videos that were not used in the training process. This evaluation process is critical for determining the accuracy and effectiveness of the model in real-world scenarios.

The testing process typically involves the following steps:

- Preparing the test data: The test data is pre-processed in the same way as the training data to remove noise and enhance the features of the hand. This ensures that the test data is consistent with the training data and that the model can accurately detect and recognize hand gestures in the test data.
- Running the detection algorithm: The pre-processed test data is then fed into the YOLOv8 model, and the model performs object detection on the test data. The model outputs the location of the hand gestures in the test data and the corresponding class of the hand gesture.
- Analyzing the results: The output of the model is then analyzed to determine its accuracy and effectiveness. This analysis may involve comparing the model's output to ground truth annotations to calculate metrics such as precision, recall, and F1-score. It may also involve a visual inspection of the model's output to identify areas for improvement.
- Iterating and refining the model: Based on the results of the testing, the YOLOv8 model can be iterated and refined to improve its accuracy and effectiveness. This may involve

adjusting the parameters of the algorithm or collecting additional data to improve the model's performance on specific hand gestures or in specific environments.

- Deployment: Once the YOLOv8 model has been tested and refined, it can be deployed in a real-world application, as shown in fig 4. The deployment process will depend on the specific use case and may involve integrating the model with other software or hardware systems.


**Figure 4:** Detected Image Testing

| Method | Training/ box loss | Validation/ box loss | Precision | Recall | Metrics/mAP_0.5 |
|---|---|---|---|---|---|
| YOLOv5 | 0.019252 | 0.011688 | 0.9914 | 0.99792 | 0.993 |
| YOLOv6 | 0.3567 | 0.3321 | 0.900 | 0.918 | 0.993 |
| YOLOv7 | 0.01207 | 0.01051 | 0.992 | 0.997 | 0.993 |
| **YOLOv8** | **0.3551** | **0.33934** | **0.99624** | **0.99829** | **0.995** |

**Table1:**Comparison Table of Model

## VI. RESULTS AND DISCUSSIONS

The neural network YOLOv8 was used to train and test a dataset of 1114 emoji hand images for this study's validation and training results. The validation and training processes were repeated 25 times, each repetition of passing the dataset through the network referred to as an epoch. The results of the training and validation processes were graphically represented in Fig. 5, which displayed the box_loss, cls_loss, dfl_loss, precision(B), recall(B), and mAP50(B) metrics for each epoch. The graph evidenced that as the number of enhanced iterations, the train losses reduced, with train loss results ranging from 0.03 (3%) and validation loss results from 0.03 (3%). The precision was 0.99224, the recall was 0.99629, and the mAP was 0.995.

The model's performance was indicated by its loss, representing the total number of errors. A train loss of 3% indicated a training accuracy of 97%, which was also true for the validation accuracy. In conclusion, this study demonstrated the effective use of YOLOv8 for training and validating a dataset of emoji hand images.

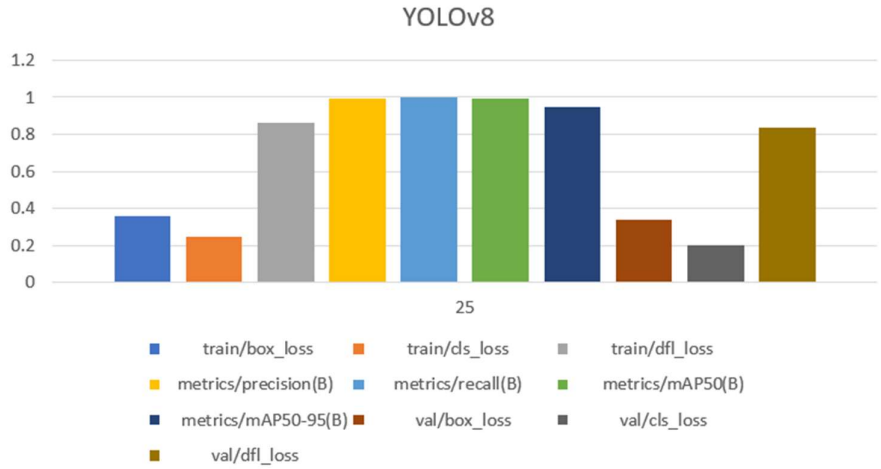


**Figure. 6:** Results

During each epoch, new models are created and analyzed based on their mAP, which the system calculates using Darknet. The neural network's mAP is determined by the tests conducted in the connectivity. Figure 7 captures the tested capable in every epoch, with mAP ranging from 0.31596 or 31.596% (model 1) to 0.995 or 99.5% (model 25). (Model 18). Model 18 had an mAP of 99.5%, an accuracy training is 97%, & a validation accuracy of 93% during the epoch. Therefore, it was selected as the model to be used in the system. In conclusion, evaluating each model based on its mAP allowed for selecting the most effective model, which in this case was model 18, with its high accuracy and precision.

The studies aim was to ensure that the program's method could detect emoji hand gestures precisely, so images of hand emojis were presented for testing. The desired output was obtained, as the identified hand emoji gestures in the pictures generated a high precision accuracy of 99.224%, as shown in Fig. 7. This high precision accuracy indicates that the model was successful in accurately identifying and classifying the hand emoji gestures in the images. In conclusion, the testing results confirmed the model's effectiveness in accurately detecting and classifying hand emoji gestures.

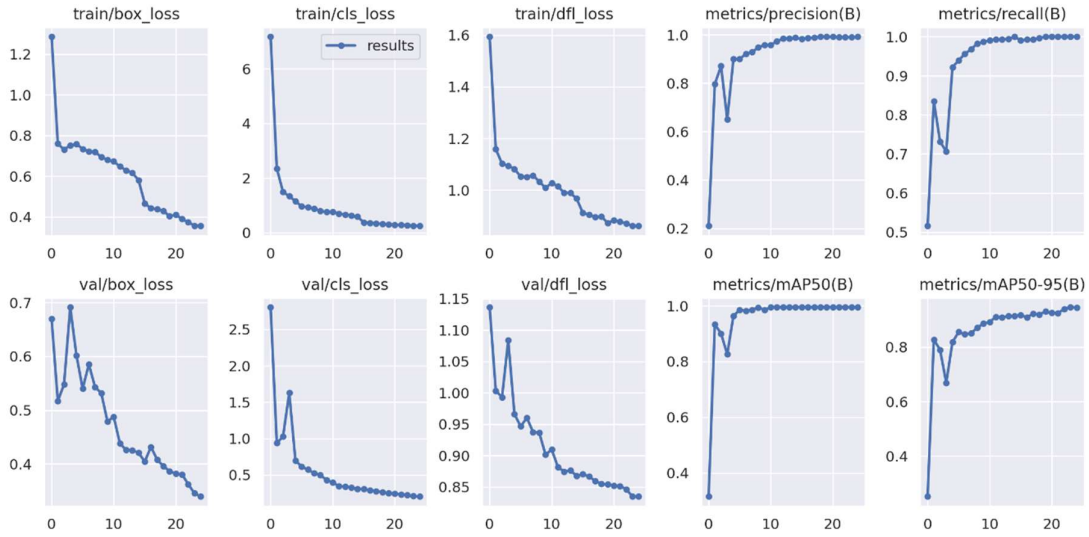


**Figure. 7:**Results

Fig. 7 displays the model's precision accuracy % after real-time testing using various hand emoji movements for detection. The real-time detection process was conducted at 2 frames per second. The results in Fig. 8 indicate that the hand emoji gestures were detected with high accuracy ranging from 92% to 98%. In summary, the model had a recognition rate of 100%, as it could detect all hand emoji gestures on every frame in real time. These results demonstrate the effectiveness and accuracy of the model in real-time hand gesture detection, which is essential for the practical application of the system. In conclusion, the high accuracy achieved in the real-time detection of hand emoji gestures proves the model's effectiveness in accurately identifying and classifying different hand gestures.


**Figure. 8:** Detection in Real-time

## VII. MODEL EVALUATION METRICS

Numerous indicators are available to assess machine learning models in diverse applications. Let's examine the evaluation metrics to evaluate a machine learning model's performance. This is essential in any data science project because it seeks to estimate a model's generalization accuracy on future data.

### A. Precision

Accuracy can become an unreliable criterion for gauging our success when there is a class imbalance. As a result, we also need to consider class-specific performance indicators. One of these measurements, called precision, is defined as positive predicted values.

$$PRICISION = \frac{TRUEPOSITIVE}{(TRUEPOSITIVE\ +\ FALSEPOSITIVE)}$$

### B. Recall

A recall is another crucial indicator; it measures the percentage of real positive cases that were accurately identified.

$$RECALL = \frac{TRUEPOSITIVE}{(TRUEPOSITIVE\ +\ FALSENEGATIVE)}$$

### C. F1-score

Precision and Recall are two significant error metrics that make up the F1 score together. In light of binary data categorization, it can be seen as the Harmonic mean of Precision and Recall error metrics for an unbalanced dataset.

$$F1\ SCORE = \frac{2 * PRECISION * RECALL}{PRECISION\ +\ RECALL}$$

### D. Accuracy

An indicator of the model's performance across all classes is accuracy. When all types are equally important, it is helpful. It is calculated by ÷the total number of % by the number of accurate forecasts.

$$ACCURACY = \frac{TP + TN}{(TP\ + TN +\ FP +\ FN)}$$

### E. MeanAveragePrecision

It is a well-liked evaluation metric for object detection in computer vision. An instance's localization pinpoints its precise location, whereas its classification identifies its nature.

$$mAP = 1 \sum_{i=1} AP_i$$

## VIII. CONCLUSION

The study's goal is to create a faster and more accurate emoji hand detection algorithm using the emoji hand dataset and Dl, the YOLOv8 algorithm. The established system has an mAP of 99.5% with 0.3551 train-loss box & 0.33934 Val-loss box, & the model generates over 95% precision accuracy and 100% detection accuracy during testing. This system's development was hampered by numerous issues, including model identification precision and accuracy, and

Google Colab's GPU consumption, among others. Several datasets were used in the study, which made going to trained time-consuming. Upgrading Google Colab's built-in GPU can be advantageous because it can speed up the training process. Creating a model with a high mAP& achieving a low Val-loss & train-loss became difficult. To fulfill this, more datasets must be training & much more time must be allocated. The system that we create has room for improvement. Future work on this study may include gesture tracking, in which the action of the hand can be understood, as well as adding more datasets of face emoji gestures that can also be identified by facial emotions

## REFERENCES

1. Alon, H. D., Ligayo, M. A. D., Melegrito, M. P., Cunanan, C. F., & Uy II, E. E. (2021, March). Deep-hand: a deep inference vision approach of recognizing a hand sign language using american alphabet. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 373-377). IEEE.
2. Bao, P., Maqueda, A. I., del-Blanco, C. R., & García, N. (2017). Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Transactions on Consumer Electronics, 63(3), 251-257.
3. Singh, A., Arora, S., Shukla, P., & Mittal, A. (2015, December). Indian Sign Language gesture classification as single or double handed gestures. In 2015 Third International Conference on Image Information Processing (ICIIP) (pp. 378-381). IEEE.
4. Wang, Z., Zhao, T., Ma, J., Chen, H., Liu, K., Shao, H., ... & Ren, J. (2020). Hear sign language: A real-time end-to-end sign language recognition system. IEEE Transactions on Mobile Computing, 21(7), 2398-2410.
5. Zhang, W., Wang, J., & Lan, F. (2020). Dynamic hand gesture recognition based on short-term sampling neural networks. IEEE/CAA Journal of AutomaticaSinica, 8(1), 110-120.
6. Zhu, C., Yang, J., Shao, Z., & Liu, C. (2019). Vision based hand gesture recognition using 3D shape context. IEEE/CAA Journal of AutomaticaSinica, 8(9), 1600-1613.
7. Sahana, T., Paul, S., Basu, S., &Mollah, A. F. (2020). Hand sign recognition from depth images with multi-scale density features for deaf mute persons. Procedia Computer Science, 167, 2043-2050.
8. Anderson, R., Wiryana, F., Ariesta, M. C., & Kusuma, G. P. (2017). Sign language recognition application systems for deaf-mute people: a review based on input-process-output. Procedia computer science, 116, 441-448.
9. Areeb, Q. M., Nadeem, M., Alroobaea, R., &Anwer, F. (2022). Helping hearing-impaired in emergency situations: A deep learning-based approach. IEEE Access, 10, 8502-8517.
10. Adithya, R. Rajesh,Hand gestures for emergency situations: A video dataset based on words from Indian sign language, Data in Brief , Volume 31,2020,106016, ISSN 2352-3409.
11. Gadekallu, T.R., Alazab, M., Kaluri, R. et al. Hand gesture classification using a novel CNN-crow search algorithm. Complex Intell. Syst. 7, 1855–1868 (2021).
12. Tan, Y.S., Lim, K.M., Tee, C. et al. Convolutional neural network with spatial pyramid pooling for hand gesture recognition. Neural Comput&Applic 33, 5339–5351 (2021).

13. Yong Soon Tan, Kian Ming Lim, Chin Poo Lee , Hand gesture recognition via enhanced densely connected convolutional neural network, Expert Systems with Applications, Volume 175,2021,114797, ISSN 0957-4174.
14. Islam, M. R., Mitu, U. K., Bhuiyan, R. A., & Shin, J. (2018, September). Hand gesture feature extraction using deep convolutional neural network for recognizing American sign language. In 2018 4th International Conference on Frontiers of Signal Processing (ICFSP) (pp. 115-119). IEEE.
15. Sharma, S., & Singh, S. (2021). Vision-based hand gesture recognition using deep learning for the interpretation of sign language. Expert Systems with Applications, 182, 115657.
16. 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
17. Areeb, Q. M., & Nadeem, M. (2021, October). Deep Learning Based Hand Gesture Recognition for Emergency Situation: A Study on Indian Sign Language. In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 33-36). IEEE.
18. Kumar, N. D., Suresh, K. V., & Dinesh, R. (2022, February). CNN-based Static Hand Gesture Recognition using RGB-D Data. In 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) (pp. 1-6). IEEE.
19. 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.
20. Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., &Mekhtiche, M. A. (2020). Hand gesture recognition for sign language using 3DCNN. IEEE Access, 8, 79491-79509.
21. M. Yasen and S. Jusoh, ‘‘A systematic review on hand gesture recognition techniques, challenges, and applications,’’ PeerJComput. Sci., vol. 5, p. e218, Sep. 2019.
22. Mohamed, N., Mustafa, M. B., &Jomhari, N. (2021). A review of the hand gesture recognition system: Current progress and future directions. IEEE Access.