A Sign Language Recognition System for Helping Disabled People

Hridoy Adhikari, Md. Sakib Bin Jahangir, Israt Jahan, Md. Solaiman Mia, Md. Riad Hassan
Department of Computer Science and Engineering, Green University of Bangladesh, Bangladesh
Email: niladhikari445@gmail.com, sakibdbz@gmail.com, isratjahan47474@gmail.com,
solaiman@cse.green.edu.bd, riad@cse.green.edu.bd

2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI) | 979-8-3503-9431-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/STI59863.2023.10465011

***Abstract*—People with disabilities have difficulty in communicating, social interaction, obsessions and repetitive behaviours. The situation gets risky when these disabled people left alone freely in the outside world. But they shouldn't be locked up for this reason. So we need a way to help and protect them. Sign language recognition is the field related to communication which is a visual language that uses body language and facial expressions to convey meaning. Recent technological advances have enabled the development of advanced sign language recognition systems that can interpret sign language and translate it into written and spoken language. These systems typically use computer vision techniques to analyse sign language gestures and movements and map them to written or spoken language. Sign language recognition technology have the potential to greatly improve the accessibility of communication for people with hearing and speech impairments and to improve communication between people who speak different languages. In this paper, our proposed system has achieved the accuracy of 91.67% which is better compared to the existing works in the literature.**

***Index Terms*—Speech Recognition, Sign Language, Image Classification, Machine Learning**

I. Introduction

The world population comprises 15% of the population with various forms of disabilities. More than 5% of the population is deaf, which is 466 million people [1]. According to the World Health Organization (WHO), by 2050 this could reach 500 million people, about 2.7 times the population in 2000. At least 70 million people are speech and hearing impaired [1]. It can greatly affect an individual's ability to communicate. So we can use Sing Language Recognition (SLR) system which can provide a means for disabled individuals to communicate with the world around them, promoting greater independence and quality of life. Sign language may not be understood by ordinary people, but it is used by the disabled people to express their feelings and thoughts towards ordinary people [2].

Industry 5.0, the latest evolution of industrial revolutions, represents a paradigm shift in manufacturing and production, intertwining cutting-edge technologies with human ingenuity [3]. While Industry 4.0 focused on automation and digitization, Industry 5.0 takes a step further by emphasizing the collaboration between humans and advanced technologies. In this context, our research in real-time SLR holds significant relevance. By developing a Convolutional Neural Network (CNN)-based model capable of detecting and interpreting sign language gestures in real-time, we have aligned with the core principles of Industry 5.0, where technology not only streamlines processes but also empowers inclusivity and human interaction. Our research bridges communication barriers for people with hearing and speech impairments, aligning perfectly with Industry 5.0's vision of integrating technology to enhance human capabilities and fostering a more inclusive, efficient and collaborative industrial landscape [4]. As Industry 5.0 envisions a future where technology serves as a facilitator rather than a replacement, our application stands as a tangible example of this cooperative and human-centered approach, enhancing communication accessibility and contributing to a more socially responsible and innovative industrial landscape. Industry 5.0 promotes a seamless collaboration between humans and machines. In the context of our proposed system, it highlights how the integration of voice recognition and sign language recognition technologies allows disabled individuals to interact with technology in a way that is most natural and convenient for them. This collaborative approach ensures that technology is a facilitator, not a barrier, to communication and control.

The contribution of this research lies in its ability to improve the quality of life of people with disabilities by providing them with the means to communicate with a wider audience, thereby promoting inclusion and equality. Additionally, the research paves the way for the development of assistive technologies that make everyday activities such as education, employment and social interactions more accessible and equitable for people with disabilities by using the sign language as the primary form of communication.

The proposed system utilises a large scale dataset (we have collected dataset from Kaggle [5] and our own surveys) of SLR to train a CNN. A CNN is a type of deep neural network which is used to process visual data [6]. The CNN extracts features from sign language and body language. This improves recognition accuracy for people with unique sign patterns. The system also includes a monitoring module that uses computer vision techniques to detect changes in the environment and respond accordingly [7]. We have evaluated the performance of the proposed system on a range of metrics including accuracy, response time and user satisfaction. We have compared our system to existing approaches and demonstrated its effectiveness in enabling disabled individuals to communicate with other people and environment. In this paper, we have also completed the sign-to-text conversation with the accuracy of over 90%.

979-8-3503-9431-3/23/$31.00 © 2023 IEEE

II. Literature Review

Various techniques have been implemented by researchers in SLR systems. The authors of [2] used several techniques to solve the problem of sign language gesture recognition. Transfer learning techniques take pre-trained models from other domains such as object recognition, face recognition, natural language processing, etc. and adapt them to recognize sign language gestures. This approach helps reduce the amount of labelled data required to train a SLR model. Hyper-tuned CNNs rely on large dataset for optimal performance. However, SLR dataset are often limited in size and diversity, which can lead to model overfitting and poor generalisation. By using data augmentation techniques such as image mirroring and rotation, researchers can increase the size and diversity of their dataset. The system could use a camera to capture the signer hand and finger movements as well as a wearable device to capture hand and finger orientation and movement.

In [6], the authors used three models for the classification which are 3D CNN, combination of CNN and Long Short-Term Memory (LSTM) and object detection which is based on YOLO v5 algorithm for detecting hand gestures. These models express the impressive performance of the object detection model to identify dynamic gestures. But, they used less amount of dataset to complete their work and the accuracy of the CNN model is 82%. In addition, the models couldn't able to help blind people. If it is converted to voice recognition, then it will be able to help blind people.

The authors of [8] presented a method for recognizing letters and numbers of Indian sign language using the Bag of Visual Words (BOVW) model. They used segmentation based on skin color and background subtraction as well as histogram-based sign mapping. Finally, CNN and Support Vector Machine (SVM) were used for the classification. They also developed their Graphical User Interface (GUI) for easy access. A custom dataset of over 36,000 images were used in this research work to recognize Indian sign language. Binary edges and smart edges were generated on the dataset mask and features were extracted with Speeded Up Robust Features (SURF). By using SVM and CNN, they yielded the accuracy of 99.17% and 99.64%, respectively [8].

In [9], CNN model was used to recognize sign language for communicating with people. The authors would likely delve into the diverse landscape of assistive technologies and their impact on communication accessibility. They might discuss the evolution of SLR systems, exploring the strengths and limitations of various approaches employed by researchers in this field. By examining prior research endeavors, they could extract valuable insights on the efficacy of different techniques, potentially highlighting advancements like Deep Learning and Neural Network (NN) that have revolutionized the accuracy of SLR systems. They used 8,958 images for testing. The accuracy of their research is 89.1%.

The authors of [10] used the SVM model. The literature review may also delve into the realm of datasets used in gesture recognition research, highlighting the need for comprehensive and diverse datasets tailored for specific contexts like emergency situations. They might draw parallels with datasets from other sign languages, exploring how these datasets have been employed in studies related to gesture recognition, communication assistance and emergency response. By examining existing studies, the authors could elucidate the uniqueness of their contribution and emphasize how their video dataset fills a crucial void in the field. They used 824 images for testing. The accuracy of their model is 90%.

In [11], the authors introduced the SLR which tried to close the communication gap between hearing impaired people and non-hearing people. The ROBITA (Indian sign language gesture database) was used as the source of the input and pre-processing was carried out to remove extraneous artifacts. The literature review would likely draw connections between prior research endeavors that have harnessed CNNs for the SLR. This might involve discussing different model configurations, datasets used and notable outcomes from these studies. They employed 356 images in their research. They got the accuracy of 87.50% by using an ML-CNN with encoder.

The authors of [12] discussed how sign language helps hearing impaired people and communicate with the general public. They provided a modified LSTM model for continuous sequences of gestures, also known as continuous SLR. In their research, they could identify a series of interconnected gestures. In total 3,150 photos were used and they got the accuracy of 89.5% and 72.3%, respectively.

III. Research Methodology

In this section, we have demonstrated our proposed research methodologies and procedures. The main goal of this research work is to detect different sign actions. We are going to design a CNN model which will detect sign actions properly and while doing this, maintain good accuracy and low parameters value. As a final output, the system recognizes and translates hand movements into sign language. The details about data collection, data preprocessing, data training, etc. are given in the following subsections.

A. Data Collection

We have collected data of 9 different classes. At first we have collected hand gestures data for training and then we resized all of the images. The original images were in large pixels, but we converted them to 190×190 pixels before using them for training. The testing data is selected from [6] and [9]. Those images were used for 9 different classes which are Washroom, Call, Food, Happiness, Helping, Medicine, Pain, Sadness and Thief. Some sample images of different classes in the dataset are shown in Fig. 1. The images were taken in the different lighting, weather and time of the day.

B. Data Description

At first, we had to understand what kind of sign languages are used to communicate with disabled people. After understanding, we have collected and stored the images. We have collected the images from different people. The size of our

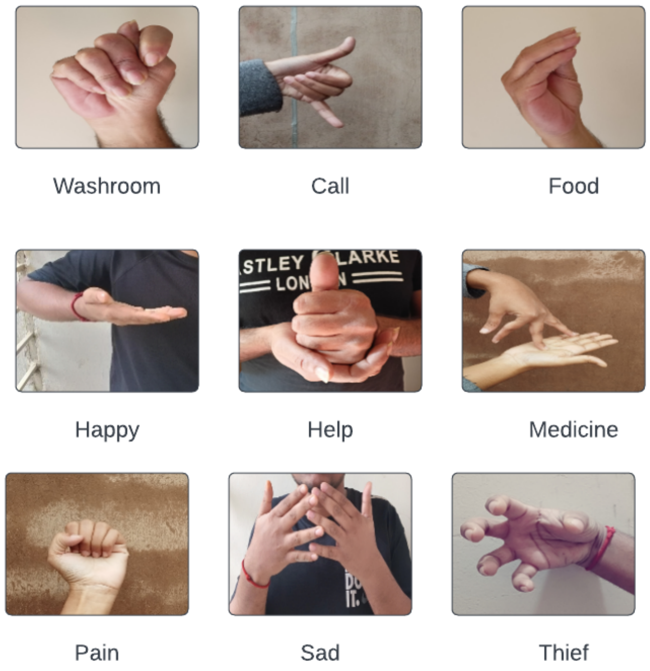


Fig. 1: Different Classes of Images from the Dataset

dataset is 6,182 which is depicted in Table I. After collecting the images, we resized them. We have used sign languages to help the disabled people who are deaf.

TABLE I: Training and Testing Images Count for Each Class

Class Name	Total Images	For Training	For Validation
Call	704	560	112
Food	945	655	159
Happiness	602	389	153
Helping	679	514	105
Washroom	584	395	129
Medicine	807	608	157
Pain	624	460	165
Sadness	510	387	113
Thief	727	639	133
Total	6,182	4,607	1,226

C. Models

For the investigation, we have primarily employed supervised Machine Learning (ML) approaches. We have used CNN model in our research.

1) Data Preprocessing: In this paper, we have used Keras and TensorFlow for the back-end development. We have converted our input images (RGB) into gray-scale. We have applied adaptive threshold to extract our hand from the background and resized the images into 190×190. We have used two sets of data named training data and validation data. Data preprocessing is a way of converting unstructured data into a more understandable format. To prevent negatively altering the outcomes, data processing should be done appropriately.

2) Data Normalization: In the context of ML, data normalization is the process of scaling and changing the input data to a standard range or distribution with the goal of improving the consistency of the learning algorithm's performance. The training set of data should first undergo through normalization and then the validation and test sets should also undergo the same modification. This guarantees that the normalization is the same for all data subsets and the learning method is not skewed by the data distribution. The formula which we have used for the normalization is given in Equation 1 [13].

$$normalization = layers.Rescaling(1/255) \tag{1}$$

3) Data Training and Validation: After performing all the steps listed above, we have passed the preprocessed input images to our proposed model for the training and testing. The prediction layer calculates the likelihood that the picture will fall into one of the classifications. As a result, the output is normalized and the total of each value in each class equals to 1. We have accomplished this by using the $Sigmoid$ and $Relu$ function.

D. Proposed Methodology

The investigation began by collecting data for training the CNN models. Then the data cleansing process involved using hashing to identify and remove duplicate images from the dataset which results to a clean dataset. This dataset then split into a training set and a test set. Data preprocessing techniques are applied to enhance data quality and suitability for training. Data normalization has been performed to prevent biases due to data distribution variations. The CNN model was developed using the Adam optimizer, categorical cross entropy loss function and accuracy as the assessment criteria. The model was trained with a batch size of 64 and 40 epochs, with 20% of the data used for the validation during training. After training, the model's performance has evaluated on a separate test dataset to assess its predictive capabilities.

E. Model Building

We have constructed a sequential CNN model. Our model has single input and output for each layer. All of these layers are placed on the top of one another to form the entire network. The system is a vision based approach. All the signs are represented with bare hands and so it eliminates the problem of using any artificial devices for interaction. In Fig. 2, the proposed CNN model is shown.

In the proposed model, a sort of feed forward NN called deep CNN model is used to alter the network's parameters in order to reduce the cost function's value [14]. The implemented model has 199,799 parameters. In Fig. 3, the proposed architecture of SLR is shown. The convolution layer has been utilized in this research and has a varied number of features with a kernel size of 3. $Relu$ has been selected as the activation function. The $SoftMax$ function is also used. In the proposed model, we have used two dropout layers. As

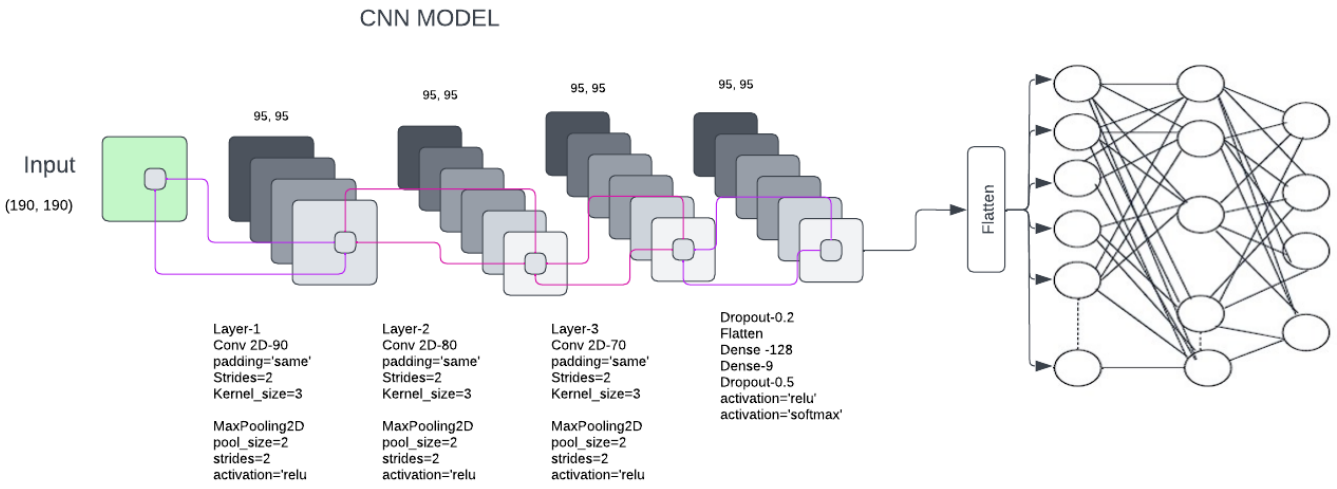


Fig. 2: Proposed CNN Model

data travels forward over the network, it is utilized to filter the data. For optimization, we have compressed the feature map using the $MaxPooling$ method. It establishes the highest or maximum value that each feature map patch contains. Here, multidimensional input is reduced to one dimension using a flattening layer. The fully connected layers are then created using the dense layer.

With the exception of batch size (64), drop rate (0.2, 0.5), epochs (40), validation percentage (20%) and testing percentage (80%), most of the training options were left at their default settings. We have increased the training batch size from the normal 132 to 32 because there weren't many training images. The number of images utilized in a batch which determines how many iterations were made. Since more training rounds will eventually result in more reliable results, we have included the entire validation dataset in the validation batch size.

F. Model Summary

The total summary of the proposed model is given in Fig. 4. Details of three Convolution layers, three Max-pooling layers, three Dropout layers, one Flatten layer, two Dense layers, shape and parameters are presented in this figure. We have used total 11 layers in this model, which is less than the other popular models. To train our proposed CNN model with the images took about 4 minutes per epoch in the Jupyter Notebook.

IV. Experimental Results and Analysis

In this section, the outcomes of the proposed method are described together with a comparison to earlier models. We have evaluated the proposed models using the selected features from the feature selection process since they somewhat enhanced performance.

A. Evaluation

We have evaluated the implemented model after training it. Fig. 5 depicts that the training accuracy is 0.9917 and the validation accuracy is 0.9167. The training and validation loss graphs are displayed in Fig. 6 which represents that both of the training and validation loss are decreasing.

B. Comparative Results

Using the CNN approach, the classification model's efficacy has been assessed. Table II provides a comparison between the proposed model with the existing models. We have contrasted the accuracy and size of the datasets, detection, etc. Table II shows how the proposed model outperforms the existing ones. The 91.67% accuracy in the proposed system is truly remarkable, especially in the context of systems intended to assist people with disabilities. For many people with disabilities, technology can be a way to regain independence. High precision in a system means it can understand and respond to their commands or needs effectively, thereby improving their quality of life. This can include tasks such as controlling smart home devices, accessing information on the Internet or even communication. In some cases, people with disabilities rely on technology for their safety and well-being, e.g., voice-activated emergency response systems. High accuracy is crucial in such situations to ensure a quick and reliable response when needed. High-precision assistive technologies can promote inclusion and improve social interactions for people with disabilities. This can facilitate communication with friends and family, access to educational and employment opportunities and participation in various aspects of the society.

C. Real-Time Sign Language Detection

By implementing the proposed method, the real-time detection has been able to detect the sign actions correctly. Fig. 7 shows some demos from which we can conclude that we are

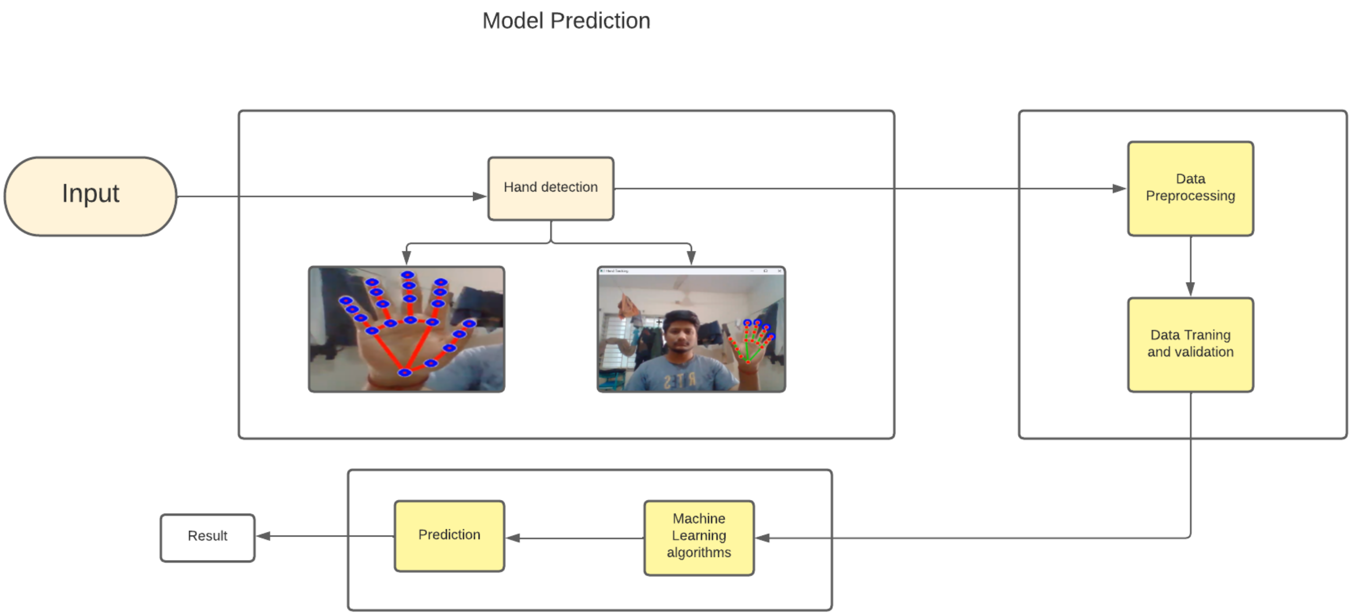
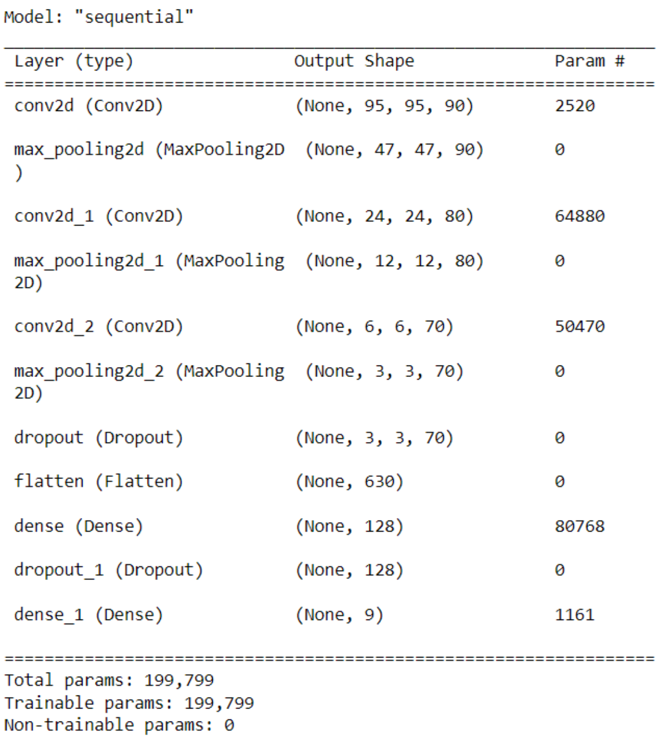


Fig. 3: Proposed Architecture for the Sign Language Recognition



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 95, 95, 90)        2520

 max_pooling2d (MaxPooling2D  (None, 47, 47, 90)       0
 )

 conv2d_1 (Conv2D)           (None, 24, 24, 80)        64880

 max_pooling2d_1 (MaxPooling  (None, 12, 12, 80)       0
 2D)

 conv2d_2 (Conv2D)           (None, 6, 6, 70)          50470

 max_pooling2d_2 (MaxPooling  (None, 3, 3, 70)         0
 2D)

 dropout (Dropout)           (None, 3, 3, 70)          0

 flatten (Flatten)           (None, 630)               0

 dense (Dense)               (None, 128)               80768

 dropout_1 (Dropout)         (None, 128)               0

 dense_1 (Dense)             (None, 9)                 1161

=================================================================
Total params: 199,799
Trainable params: 199,799
Non-trainable params: 0
```

Fig. 4: Summary of the Proposed Model

able to clearly detect the signs. In the context of sign language recognition, it may mean that the model overfits the specific data in the training set and does not generalize very well for the new signs.

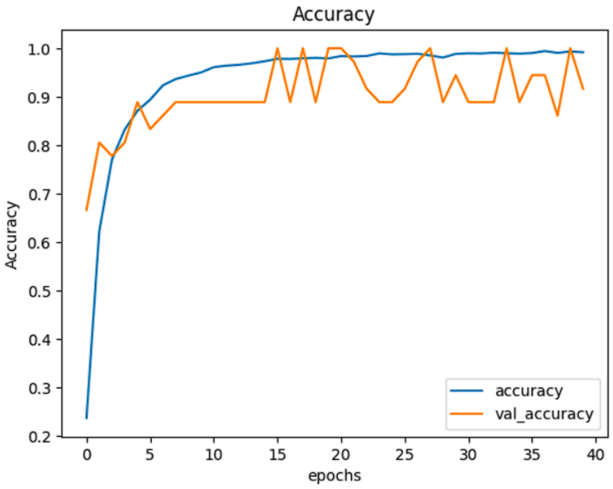


Fig. 5: Evaluation of the Training and Validation Accuracy

TABLE II: Comparison with the Existing Models

Existing Works	Model Name	Accuracy	Image	Detection
[6]	CNN	82.00%	2,060	YES
[9]	CNN	89.1%	8958	NO
[10]	SVM	90.00%	824	YES
[11]	CNN	87.50%	356	NO
[12]	LSTM	89.5%	3150	NO
Proposed Method	CNN	91.67%	6,182	YES

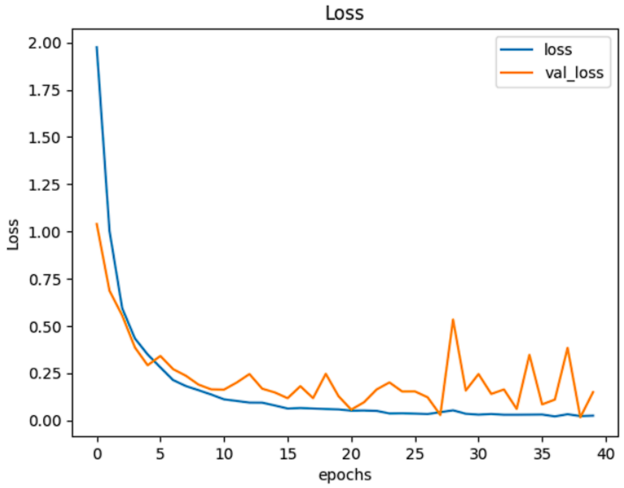
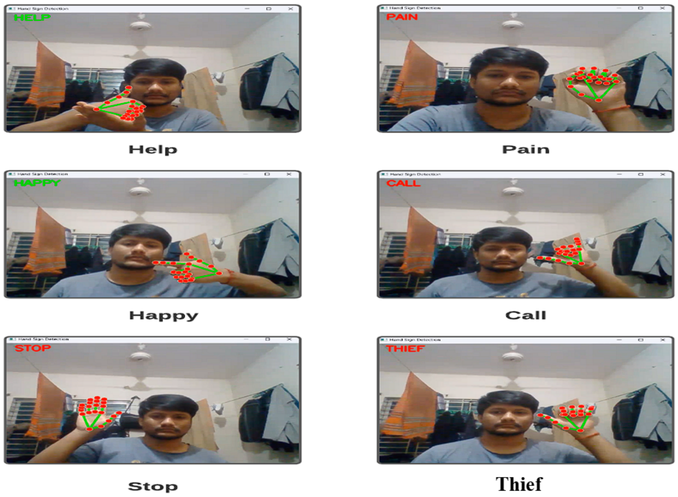


Fig. 6: Evaluation of the Training and Validation Loss



Help Pain

Happy Call

Stop Thief

Fig. 7: Real-Time Sign Language Detection

V. Conclusion

In this paper, we have focused on sign language to help the disabled people. We have used CNN model and found better accuracy in sign language recognition. Our accuracy stands at 91.67%, with real-time detection of sign language. The proposed technique has contributed to an enhanced accuracy in sign identification. Yet, there are still a lot of problems that require further study. In future, weâll try to construct utterance level representations using a variety of statistical moments. Moreover, as technology continues to evolve, our next critical objective revolves around voice-to-text conversion. This expansion into voice recognition will bridge another communication gap, enabling seamless interactions between sign language users and those relying on spoken language. Our dedication to serving the needs of individuals with disabilities remains unwavering and we are poised to continue our journey toward creating innovative solutions that foster inclusion, understanding and empowerment. Our proposed system may have difficulties in interpreting accurately the speech of individuals with different communication styles, accents or language disorders. Accents, dialects and variations in speech patterns can pose challenges for speech recognition systems. The system may raise privacy and security concerns, especially if the system continuously listens or records user interactions.

Acknowledgment

This work was supported in part by the Center for Research, Innovation, and Transformation (CRIT) of Green University of Bangladesh (GUB).

References

[1] K. K. Podder, M. E. H. Chowdhury, A. M. Tahir, Z. B. Mahbub, A. Khandakar, M. S. Hossain and M. A. Kadir, “Bangla Sign Language (BdSL) Alphabets and Numerals Classification Using a Deep Learning Model," Sensors, vol. 22, no. 2, pp. 574(1-18), 2022.

[2] A. Mannan, A. Abbasi, A. R. Javed, A. Ahsan, T. R. Gadekallu and Q. Xin, “Hypertuned Deep Convolutional Neural Network for Sign Language Recognition," Computational Intelligence and Neuroscience, vol. 2022, Article ID 1450822, pp. 1-10, 2022.

[3] F. Tanrisever and K. A. W. Voorbraak, “Crowdfunding for Financing Wearable Technologies," 2016 49th Hawaii Int. Conf. System Sciences (HICSS), pp. 1800-1807, 2016.

[4] S. C. Das, M. B. Alam, M. S. A. Moon and M. S. Mia, “An Application Programming Interface to Recognize Emotion using Speech Features," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1-6, 2022.

[5] A. Mavi, “27 Class Sign Language Dataset,â Kaggle, [Online]. Available: https://www.kaggle.com/datasets/ardamavi/27-class-sign-language-dataset. [Accessed: 26-Oct-2023].

[6] Q. M. Areeb, Maryam, M. Nadeem, R. Alroobaea and F. Anwer, “Helping Hearing-Impaired in Emergency Situations: A Deep Learning-Based Approach," IEEE Access, vol. 10, pp. 8502-8517, 2022.

[7] A. Halder and A. Tayade, “Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning," Int. J. Research Publication and Reviews, vol. 2, no. 5, pp. 9-17, 2021.

[8] N. Mohamed, M. B. Mustafa and N. Jomhari, “A Review of the Hand Gesture Recognition System: Current Progress and Future Directions," IEEE Access, vol. 9, pp. 157422-157436, 2021.

[9] Y. Obi, K. S. Claudio, V. M. Budiman, S. Achmad and A. Kurniawan, “Sign language recognition system for communicating to people with disabilities," Procedia Computer Science, vol. 216, pp. 13-20, 2023.

[10] V. Adithya and R. Rajesh, “Hand gestures for emergency situations: A video dataset based on words from Indian sign language," Data in Brief, vol. 31, pp. 1-7, 2020.

[11] G. A. Prasath and K. Annapurani, “Prediction of sign language recognition based on multi layered CNN," Multimedia Tools and Applications, vol. 82, pp. 29649-29669, 2023.

[12] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian and B. B. Chaudhuri, “A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion," IEEE Sensors Journal, vol. 19, no. 16, pp. 7056-7063, 2019.

[13] Z. H. Nayem, I. Jahan, A. A. Rakib and M. S. Mia, “Detection and Identification of Rice Pests Using Memory Efficient Convolutional Neural Network," 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE), pp. 1-6, 2023.

[14] A. T. Ali, H. S. Abdullah and M. N. Fadhil, “Voice recognition system using machine learning techniques," Proceedings Materials Today, pp. 1-7, 2021.