Developing Convolutional Neural Networks for Recognition of American Sign Language

Farah Jawad Al-Ghanim[1], Salwa Shakir Baawi[2] and Nisreen Ryadh Hamza[3]

[1,3] *Department of Computer Science, College of Computer Science and Information Technology, University of Al-Qadisiyah.*

[2] *Department of Computer Information Systems, College of Computer Science and Information Technology, University of Al-Qadisiyah.*

[*]*Corresponding Author:* farah.jawad@qu.edu.iq, salwa.baawi@qu.edu.iq, nesreen.readh@qu.edu.iq

Received 12 Feb. 2025, Accepted 21 Mar. 2025, Published 30 June. 2025.

DOI: 10.52113/2/12.01.2025/1-12

Abstract: Rather than using speech to communicate with one another, the deaf and dumb use a set of signs known as "sign language". Yet, utilizing signs to interact with this society is too difficult for non-sign language speakers. To facilitate communication for the deaf public, an application that can identify sign language motions must be developed. Regarding its importance, there are approaches with differing degrees of accuracy for recognizing American Sign Language ASL. The study aims to enhance the accuracy of current ASL identification approaches by putting forward a deep-learning model. A CNN was developed and trained to correctly recognize hand gestures that describe the ASL letters (A-Z). The proposed model performs exceptionally well, attaining high accuracy on the dataset, with a test accuracy of 99.97%. The model is a possible tool for practical applications in assistive technology for the hearing impaired since the results show that it can distinguish between distinct ASL hand signs.

***Keywords*:** Hand gesture, American Sign Language, CNN, Sign Recognition, ASL letters.

1. Introduction

The social growth of a human being depends on communication. In a culture where oral languages are the norm, the deaf community has created many body-based communication systems that rely on facial expressions and motions of the hands and arms. These forms of communication are referred to as sign languages [1].

Global grammatical variances are frequently impacted by the vocal languages that coexist with them both historically and geographically. A signed representation of the alphabet spoken languages designed to enhance the capabilities of sign language is referred to as a manual or dactylological alphabet.

©Al-Ghanim, 2025. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International

The 27 signs that make up Mexico's dactylological alphabet are 21 static, and the remaining signs are dynamic [2].

The hand's orientation and position over a period of time without movement are known as static signs; if the hand moves during that period, it is referred to as dynamic. Other versions exist depending on the country or location, including French (Langue des Signes Française), Brazilian (Língua Brasileira de Sinais), and American (American Sign Language). Thus, a wide range of research has been done on creating software and technology that would let deaf and non-deaf individuals communicate [3].

Using a variety of finger alphabets and gestures, deaf people can express words and letters using sign languages. Facial expressions and head motions are significant components of sign language that improve expression in addition to finger alphabets and gestures. Many different countries have distinctive sign languages. The alphabet and numbers in American Sign Language (ASL) are displayed in Fig. 1, which was created by Massey University [4].

Research in the field of gesture recognition continues to grow due to the development of new technologies for smartphones, laptops, and other portable devices.

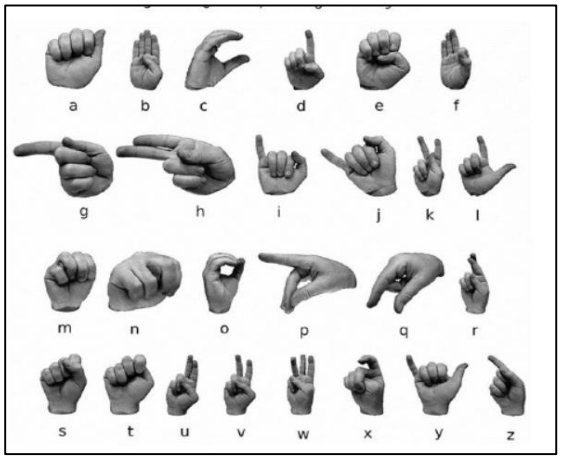

Fig. (1): American sign alphabet [4]

The gesture recognition system was first developed using a sensor based on wearing hand gloves, but the cost of production was exorbitant. These days, camera vision-based sensors are employed due to their ease of use and simplicity. Gestures can also be expressed with static hand configurations and positions that do not include any movement. While multiple-hand postures can be merged to portray a continuous motion, a single-hand position can represent a stance [5].

In artificial intelligence (AI), DL is a branch of machine learning. It is a collection of models and methods that use structures made up of several nonlinear transformations to provide high-level abstractions. In order to replicate the

human brain's ability for learning, analysis, observation, and inference—particularly when dealing with exceptionally challenging problems—DL algorithms use vast amounts of data to automatically extract features. DL architectures extract representations directly from data without the need for human interaction, generating learning patterns and establishing links beyond the data's immediate neighbors.

Numerous deep learning architectures exist, including convolutional neural networks, stacked autoencoders, deep belief networks, and others. CNN has employed a multi-layered artificial neural network (ANN) to achieve advanced accuracy in a variety of fields, including bioinformatics, medical image analysis, computer vision, speech recognition and etc. [6].

A convolutional neural network (CNN) is regarded as one of the most commonly applied deep learning methods. It comprises convolutional layers that are followed by one or more fully connected layers. The CNN is a set of digital filters whose weights are calculated during the learning phase, based on computer science principles. The human brain, of course, is involved in more intricate processes. Each convolutional layer uses the analogy to extract features from training data. A CNN uses convolutional layers, combines learnt features and input data, and transforms this architecture into a form that is well-suited for data processing. CNNs use several hidden layers to learn how to identify various data features. The learnt data features get harder with each hidden layer [7].

The seeable language known as American Sign Language (ASL) is utilized by the Deaf and hard-of-hearing individuals. Individual sign language users and non-users may be capable of communicating more effectively through automated ASL recognition. Developing a model of a convolutional neural network (CNN) to automatically identify ASL letters from images is the core objective of this study. The technology has been designed to be strong and flexible, which will help develop sign language recognition systems that are applicable in everyday scenarios.

The remaining section of this article is organized as follows: Section 2 discusses the literature review. The suggested methodology, together with its dataset and system architecture and model architecture, is covered in Section 3.

Section 4 discusses comparative analysis and the final results of the proposed method. Section 5 concludes the article and recommends areas for future research.

2. Literature Review

Some studies have been done to automatically recognize ASL. While some of these studies employ the traditional neural network approach for classification, others have used manual feature identification, and pertinent feature selection is necessary for shallow neural networks. The performance of traditional neural networks has been much enhanced by the application of deep learning (DL) approaches to machine learning tasks, mainly those applying image recognition and computer vision.

In [8], Neto, Geovane, et al., a deep learning model was trained and tested for (LSA) Argentinian Sign Language in a dataset of 3000 videos and 64 signs of LSA. In their test, they obtained an accuracy of 93% using 3D-NN layers.

In [9], Cui, Runpeng et al., this study combined CNN and Bi-LSTM to create a system for continuous sign recognition in German sign language. They state that this method works better than HMMs in learning temporal dependencies; thus, they employ RNN as the sequence learning module.

In [10], Lee, Carman et al., they suggested a prototype for an ASL learning application. Since the classification approach depends on handling input sequences, the Long-Short Term Memory Recurrent Neural Network with the k-Nearest-Neighbor method is used. The model was trained using 2600 samples, 100 samples for each alphabet. In 5-fold cross-validation using a leap motion controller, the results of the experiment demonstrated that the recognition rate for 26 ASL alphabets provides an average accuracy rate of 99.44% and 91.82%.

In [11], Saiful, Md. Nafis et al., presented a new deep learning-based method for identifying sign language. They first created a dataset with 11 sign terms to detect real-time sign language. They then trained their customized CNN model using these sign words. According to the results, on this test dataset, the customized CNN model achieved the highest accuracy of 98.6%.

Li, Lanxi et al. in [12], conducted a study to identify ASL using the MNIST dataset. They evaluated how well different classification algorithms work with both raw data and data that has been preprocessed (noise reduction, hand

detection, etc.). Their test data validation scores for SVM, random forest classifier (RFC), KNN, SGD, and NBG machine learning and classification algorithms are 84.19%, 81.61%, 78.17%, 66.02%, and 38.90%, respectively, as the result of their research on raw data.

In [13], Ansar, Hira et al., developed a CNN architectural model for ASL recognition. They preprocess the raw dataset in some way. They use the MNIST dataset for two separate validations of their constructed model. A third of the dataset is employed for testing and the rest for training in the initial validation. While two-thirds of the dataset is applied for training and the rest is used for testing in the second validation. For the first evaluation, they obtained validation accuracies of 93.2%, and for the second evaluation, they got 91.6%.

In this research [14] Sreemathy, R. et al. presented a technique for automatically recognizing Indian Sign Language (ISL) two-handed signs. It uses Histogram Oriented Gradient (HOG) features to train a BPN, and the learnt model is used to test the motions in real time. They achieved (89.5%) total accuracy utilizing (50) hidden neurones and (5184) input features. In addition, AlexNet, GoogleNet, VGG-16, and VGG-19 were used in a deep learning technique, yielding accuracies of (99.1%), (95.8%), (98.4%), and (99.1%), respectively.

Bhaumik, Gopa et al. [15] suggested SpAtNet, a CNN-based network that learns spatial features for accurate hand gesture identification. To extract rich spatial information, they employ multiscale filters. Six benchmark datasets (ASL Finger Spelling, MUGD, NUS-II, Triesch, HGR-I, and ArASL) are used to validate the proposed approach. By using the LOSO method to evaluate their work on MUD, they achieve an 80.44% recognition rate.

Gu, Yutong et al. [16] introduced a new study that used wearable inertial motion capture technology to gather an American sign language dataset. Deep learning models will then be used to recognize and translate sign language statements from beginning to end. The model's accuracy ratings are 97.34% for sentence-level evaluation and 99.07% for word-level evaluation.

The suggested CNN approach balances accuracy and computational cost, unlike earlier research that depended on machine and deep learning models or traditional feature extraction.

3. Methodology

The study's approach is organized to recognize the American alphabet in sign language ASL. To achieve ideal classification performance, it combines the benefit of a CNN model, data augmentation, and a pre-processed dataset. A convolutional neural network (CNN) was developed in this study and trained to accurately recognize hand signs for the ASL letters (A-to-Z). These phases are depicted in Figure 2 below.

3.1 Dataset

High-quality images for American Sign Language for letters in the English alphabet are included in this collection dataset. It contains (26,331) images that represent (26) classes, individually of which is described by a letter in the American English alphabet in Figure 3. This Fingerspelling Dataset is perfect for translating sign language into voice or text. [17]

The variety of backgrounds is intentionally used to improve visual clearness. The dataset images are resized to (64*64) pixels as part of the pre-processing phase.

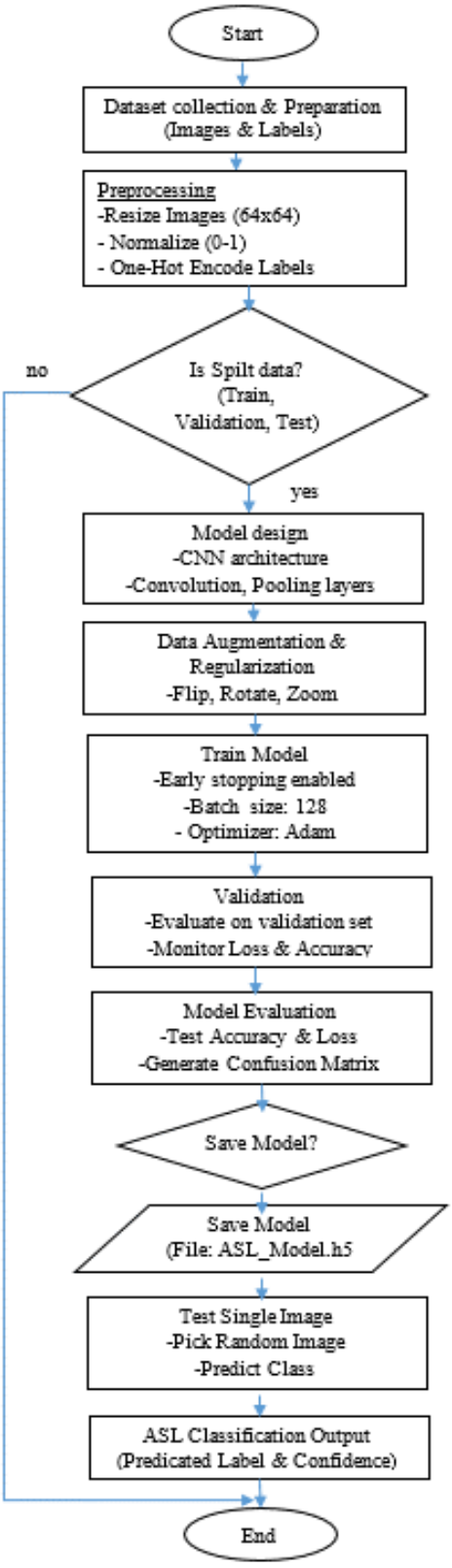

Fig. (2): The proposed system flow chart

It's may be employed by researchers to train and test models for computer vision to identifying sign language, offering communication systems, and assistive technology.

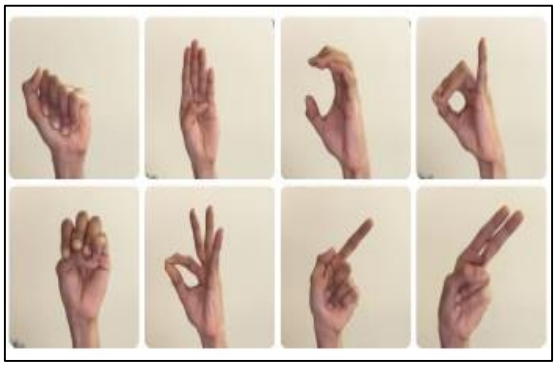

Fig. (3): Illustrates an example of a American Alphabet Sign Language Dataset

3.2 Model Architecture

When it comes to sign language detection, CNNs have shown impressive performance. CNNs are first trained to identify patterns in a dataset of sign language motions in order to be used for ASL identification. This dataset often includes pictures or videos of people displaying the signs together with the relevant signs identified on them. The CNN recognizes signs it has never seen before by using the distinctive features it has learnt to recognize during training. A deep CNN was designed to classify the ASL letters, model architecture consisted of:

- Three Convolutional layers with filter sizes (32, 64, 64) are followed by MaxPooling and ReLU activation.
- Dropout layers (0.2) to prevent overfitting are used.
- A fully connected layer.
- Two dense with 64 and 128 neurones with ReLU activation for feature extraction, this layer helps the model learn complex representations of data.
- A final layer (softmax), 26 units (classes) one for each ASL letter, for multi-class classification to predict the probability of each ASL letter class.
- Compiled model using the Adam optimizer (learning rate of 0.001), categorical cross-entropy loss function, trained for up to 50 epochs
- The dataset split into (70% training, 15% validation, and test sets 15%).

The CNN Model Summary in Table1 as:

Table 1: Model Summary.

Layer Type	Output Shape	Para.	Importance
Input	(64, 64, 3)	-	Inputs the image
Conv2D	(32, 60, 60)	32 filters, 5x5 kernel, ReLU	Extracts essential features
Max Pooling 2D	(32, 30, 30)	2x2 Pooling	Reduces the dimensions while retaining important information
Conv2D	(64, 28, 28)	64 filters, 3x3 kernel, ReLU	Extracts essential features
Max Pooling 2D	(64, 14, 14)	2x2 Pooling	Reduces the dimensions while retaining important information
Dropout	(64, 14, 14)	25%	Reduces overfitting by randomly disabling some units
Conv2D	(64, 12, 12)	64 filters, 3x3 kernel, ReLU	Extracts essential features
Max Pooling 2D	(64, 6, 6)	2x2 Pooling	Reduces the dimensions while retaining important information
Flatten	(2304)	Convert to 1D	Converts the matrix into a single vector for processing by the dense layers
Dropout	(2304)	25%	Reduces overfitting by randomly disabling some units
Dense	(128)	128 Neurons, ReLU	Fully connected layers that learn relationships between extracted features
Dropout	(128)	25%	Reduces overfitting by randomly disabling some units
Dense	(64)	64 Neurons, ReLU	Fully connected layers that learn relationships between extracted features
Dense (Output)	(26)	Softmax Activation (26 Classes)	Final layer for classifies using Softmax activation

4. Results and Discussion

The proposed model developed in this paper showed remarkable performance in the ASL letter classification. The model's accuracy and high generalization to unseen data are demonstrated. This result enhances the expanding field of automated sign language recognition, especially in helping people with hearing impairments, by providing a robust and reliable system for real-time sign language recognition.

The model achieved a test accuracy of 99.97% and a validation accuracy that nearly matched the training accuracy. All classes' F1-scores, recall, and accuracy were perfect, according to the classification report in Table 2.

To visually analyze the classification results, a confusion matrix was plotted in

Figure 4. Lastly, random picture testing shows how applicable it is in the actual world as in Figure 5.

The suggested system runs on Python 3.6 and is programmed on a 64-bit Windows 10 OS with an Intel Core i7-7700HQ (7th Gen) CPU with 16GB of RAM and an Nvidia GEFORCE GTX with 6GB of RAM to speed up the neural networks' calculations. The performance of this system is assessed using a wide range of measures, including accuracy, precision, recall, and F1-score. A test's accuracy could be evaluated using the F-score, a statistic that balances recall and precision.

This metric may provide the most precise performance measurements by utilizing accuracy and recall. Importantly, a greater F-measure results in better outcomes.

Recall and accuracy are positively correlated with fewer false positives and false-negatives, respectively. These metrics are as follows:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F-meature = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Table 2: Classification Report for ASL letter.

Class	Precision	Recall	F1-Score
A-samples	1.00	1.00	1.00
B-samples	1.00	1.00	1.00
C-samples	1.00	1.00	1.00
D-samples	1.00	1.00	1.00
E-samples	1.00	1.00	1.00
F-samples	1.00	1.00	1.00
G-samples	1.00	1.00	1.00
H-samples	1.00	1.00	1.00
I-samples	1.00	1.00	1.00
J-samples	1.00	1.00	1.00
K-samples	1.00	0.99	1.00
L-samples	1.00	1.00	1.00
M-samples	1.00	1.00	1.00
N-samples	1.00	1.00	1.00
O-samples	1.00	1.00	1.00
P-samples	1.00	1.00	1.00
Q-samples	1.00	1.00	1.00
R-samples	1.00	1.00	1.00
S-samples	1.00	1.00	1.00
T-samples	1.00	1.00	1.00
U-samples	1.00	1.00	1.00
V-samples	0.99	1.00	1.00
W-samples	1.00	1.00	1.00
X-samples	1.00	1.00	1.00
Y-samples	1.00	1.00	1.00
Z-samples	1.00	1.00	1.00

The proposed CNN performs agreeably when the confusion matrix is studied. Every value is only found on the main diagonal; this means the model has

correctly classified the sample. Also, the model's perfect accuracy is verified by the absence of data outside the diagonal. These conclusions imply that the model operates well on the test data.

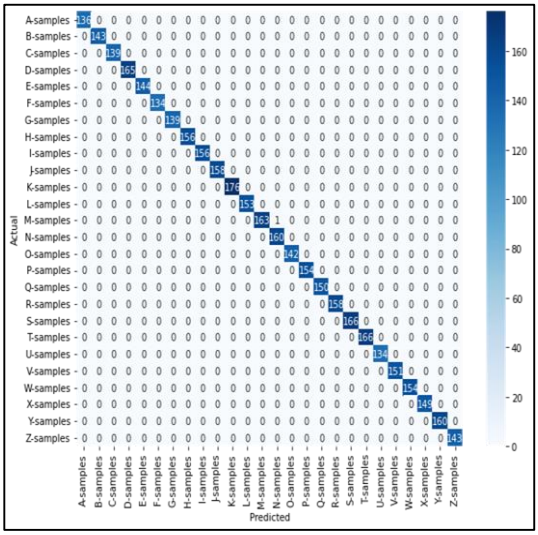

Fig. (4): Confusion matrix for ASL.

The reliability of the model in accurately recognizing ASL letters was demonstrated by random image testing from the test set.

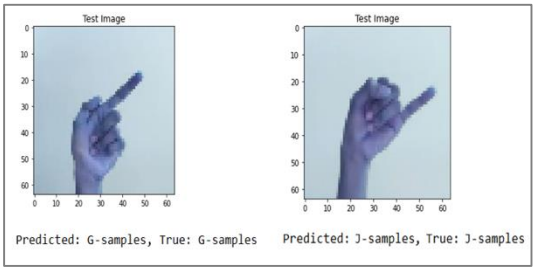

Fig. (5): Example of Test Image with Predicted and True Labels.

To make comparisons with previous studies that used this dataset. This dataset was sourced from the Kaggle website and was made public in 2024; as of the completion of this research work, no other published papers have used it.

5. Conclusion

Sign language blends intricate facial expressions, bodily postures, and hand gestures. Only a few, though, can comprehend and utilize it. To lessen the burden, a convolutional neural network (CNN)-based computer help for sign language identification with a fingerspelling manner is suggested.

This paper has developed a CNN model for classifying ASL letters with near-perfect test accuracy (99.97%). The performance model of the test set shows how well CNNs function in visual recognition tasks, mainly when it comes to classifying sign language.

To make the model relevant in real-world settings, future work will concentrate on extending it to incorporate dynamic hand gestures (words and sentences) and real-time video input.

References

[1] Cheok, Ming Jin, Zaid Omar, and Mohamed Hisham Jaward. "A review of hand gesture and sign language recognition techniques." International Journal of Machine Learning and Cybernetics 10 (2019): 131-153.

[2] Carmona-Arroyo, G.; Rios-Figueroa, H.V.; Avendaño-Garrido, M.L. Mexican Sign-Language Static-Alphabet Recognition Using 3D Affine Invariants. In Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches; Scrivener Publishing LLC: Beverly, MA, USA, 2021; pp. 171–192.

[3] Sánchez-Vicinaiz, Tzeico J., Enrique Camacho-Pérez, Alejandro A. Castillo-Atoche, Mayra Cruz-Fernandez, José R. García-Martínez, and Juvenal Rodríguez-Reséndiz. "MediaPipe Frame and Convolutional Neural Networks-Based Fingerspelling Detection in Mexican Sign Language." Technologies 12, no. 8 (2024): 124.

[4] Bayrak, Selda, Vasif Nabiyev, and Celal Atalar. "American Sign Language Recognition Model Using Complex Zernike Moments and Complex-Valued Deep Neural Networks." IEEE Access (2024).

[5] Sundar, B., and T. Bagyammal. "American sign language recognition for alphabets using MediaPipe and LSTM." Procedia Computer Science 215 (2022): 642-651.

[6] Rahman, Md Moklesur, Md Shafiqul Islam, Md Hafizur Rahman, Roberto Sassi, Massimo W. Rivolta, and Md Aktaruzzaman. "A new benchmark on american sign language recognition using convolutional neural network." In 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1-6. IEEE, 2019.

[7] Vaillant, Régis, Christophe Monrocq, and Yann Le Cun. "Original approach for the localisation of objects in images." IEE Proceedings-Vision, Image and Signal Processing 141, no. 4 (1994): 245-250.

[8] Neto, Geovane M. Ramos, Geraldo Braz Junior, João Dallyson Sousa de Almeida, and Anselmo Cardoso de Paiva. "Sign language recognition based on 3d convolutional neural networks." In Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15, pp. 399-407. Springer International Publishing, 2018.

[9] Cui, Runpeng, Hu Liu, and Changshui Zhang. "A deep neural framework for continuous sign language recognition by iterative training." IEEE Transactions on Multimedia 21, no. 7 (2019): 1880-1891.

[10] Lee, Carman KM, Kam KH Ng, Chun-Hsien Chen, Henry CW Lau, Sui Ying Chung, and Tiffany Tsoi. "American sign language recognition and training method with recurrent neural network." Expert Systems with Applications 167 (2021): 114403.

[11] Saiful, Md Nafis, Abdulla Al Isam, Hamim Ahmed Moon, Rifa Tammana Jaman, Mitul Das, Md Raisul Alam, and Ashifur Rahman. "Real-time sign language detection using cnn." In 2022 International Conference on Data Analytics for Business and Industry (ICDABI), pp. 697-701. IEEE, 2022.

[12] Li, Lanxi, Da Liu, Chenlin Shen, and Jing Sun. "American Sign Language Recognition based on machine learning and Neural Network." In 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 452-457. IEEE, 2022.

[13] Ansar, Hira, Naif Al Mudawi, Saud S. Alotaibi, Abdulwahab Alazeb, Bayan Ibrahimm Alabdullah, Mohammed Alonazi, and Jeongmin Park. "Hand gesture recognition for characters understanding using convex Hull landmarks and geometric features." Ieee Access 11 (2023): 82065-82078.

[14] Sreemathy, R., Mousami Turuk, Isha Kulkarni, and Soumya Khurana. "Sign language recognition using artificial intelligence." Education and Information Technologies 28, no. 5 (2023): 5259-5278.

[15] Bhaumik, Gopa, and Mahesh Chandra Govil. "SpAtNet: A spatial feature attention network for hand gesture recognition." Multimedia Tools and Applications 83, no. 14 (2024): 41805-41822.

[16] Gu, Yutong, Hiromasa Oku, and Masahiro Todoh. "American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System." Sensors 24, no. 2 (2024): 453.

[17] Chaitanya_Kakade, 2024, American Sign Language Dataset, https://www.kaggle.com/datasets/chaitanyakakade77/american-sign-language-dataset/.