Deep Learning-Based Recognition and Detection of the Filipino Sign Language Alphabet

Miguel Raphael Layos, Brix Anthony Manzanero, Kris Brian Diaz
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
layosm@students.national-u.edu.ph
manzanerobg@students.national-u.edu.ph
diazkv1@students.national-u.edu.ph

***Abstract*—The Filipino deaf community continues to face significant communication barriers in the Philippines. Developing a sign language recognition can promote more inclusive communication. This project presents a Filipino Sign Language Alphabet recognition and detection model utilizing a Convolutional Neural Network (CNN).**

***Index Terms*—Sign Language Recognition, Sign Language Detection, Filipino Sign Language, Convolutional Neural Network**

I. Introduction

Communication is fundamental to the growth of an emerging nation, as it fosters understanding and drives development across all sectors of society. However, individuals with hearing or speech impairments face significant barriers due to their inability to engage in conventional verbal communication [1]. To overcome this challenge, these individuals rely on sign language, a visual system of communication that utilizes hand gestures, facial expressions, body movements, and spatial orientation to convey meaning. Different sign languages have developed globally, each shaped by the culture and language of its users. [2].

In the Philippines, approximately 1.23% of the population is either deaf, mute, or hearing-impaired [3]. To support accessibility and inclusion, the country officially recognizes Filipino Sign Language or FSL as the national sign language and is used in educational institutions, legal proceedings, broadcast media, and other public services involving the deaf to ensure equal access to people with disabilities [4]. But despite the implementations, most Filipinos do not understand FSL. As a result, causes a significant communication gap between the deaf and hearing community [5].

With the continuous advancement of technology and research, have allowed better accessibility for deaf people [6]. Sign language recognition using computer vision and deep learning can be utilized to provide learning and assistive technologies to the deaf community [7].

This study aims to develop a system capable of identifying FSL letter gestures and detecting their spatial location within an image. To achieve this, the study will employ deep learning techniques, with a particular emphasis on Convolutional Neural Networks (CNNs), which are well-suited for image-based tasks. The scope is limited to FSL alphabet signs, specifically the letters A to Z. The primary objective is to establish a benchmark performance score for an image-based FSL letter recognition and detection model, which may serve as a reference for future, more advanced systems. Furthermore, this study seeks to deepen the understanding of deep learning architectures to support the development of assistive technologies that benefit the deaf community.

II. Review of Related Literature

With the rapid advancement of technology, a substantial amount of research has been directed toward the development of Sign Language Recognition (SLR) systems. A study [8] examined and reviewed the advancements in SLR across different languages. Their study classified SLR systems into two main categories: hardware-based and software-based. Hardware-based systems involve the use of specific wearable devices such as data gloves, Kinect sensors, or other sensor-based equipment. These systems have seen significant improvements through the integration of enhanced sensors into wearable devices like gloves, watches, and bands. While such systems offer high precision, they often present issues related to user discomfort during prolonged use and incur high maintenance costs.

In contrast, software-based SLR systems rely on computer vision and deep learning techniques. These approaches have shown promising performance, largely due to ongoing advancements in deep learning algorithms. Nevertheless, challenges persist in acquiring high-quality training data. Limitations such as small datasets, poor image quality, camera signal inconsistencies, background clutter, and variable lighting conditions can adversely impact the accuracy and robustness of these systems.

P. Bhatia et al. [9] observed that the number of studies focusing on Sign Language Recognition (SLR) has been steadily increasing each year. Their work predicts that research in this domain will continue to grow significantly, based on historical publication trends and increasing interest in the field. In their systematic review, it was found that the majority of research has been conducted on American Sign Language (21%), followed by Indian Sign Language (16%), Arabic Sign Language (13%), Chinese Sign Language (12%), Persian Sign Language (4%), with other sign languages collectively accounting for 34%. Although SLR research has been ongoing for several years, the study concludes that it is still in its early stages, as no large-scale, real-time system capable of interpreting a wide vocabulary of signs has been deployed. Major challenges include hand occlusion, limited and unscalable datasets, variations in background illumination, and high computational requirements. Furthermore, the study notes that most existing work has been limited to isolated sign recognition. Future research is encouraged to focus on continuous sign recognition to advance practical applications in the field.

In terms of deep learning applications, most SLR research has been concentrated on American Sign Language (ASL). For instance, a study [10] proposed a multimodal fusion-based SLR system that collected 1,400 static single-handed signs using Kinect and Leap Motion sensors. The system utilized Convolutional Neural Networks (CNNs) and achieved a recognition accuracy of 97% by combining color, depth, and motion data. Similarly, another study [11] developed a vision-based static hand recognition system. By collecting 2,040 alphabet signs and applying median filtering for preprocessing, the researchers achieved a CNN-based accuracy of 91.33%. Notably, a study [12] introduced a real-time SLR system aimed at enhancing communication for the deaf and mute. Implemented in Python, the system employed MediaPipe for precise hand tracking and Bi-directional Long Short-Term Memory (BiLSTM) networks for classification. Using a dataset of 676 images representing 80 static hand gestures, the model achieved an accuracy of 98.35%, a precision of 90.91%, a recall of 99.09%, and an F1 score of 90.91%. Compared to traditional models such as YOLOv4 and Support Vector Machines (SVM), the BiLSTM-MediaPipe approach demonstrated superior performance, particularly in capturing fine-grained gesture variations while maintaining real-time efficiency. The authors suggested further improvements through the inclusion of dynamic gestures and integration with assistive technologies for broader applicability.

Research on Filipino Sign Language (FSL) remains limited but notable contributions have emerged. Montefalcon et al. [13] developed a deep learning-based system for recognizing FSL numeric gestures (0–9). Utilizing a fine-tuned ResNet-50 model trained on images depicting various hand postures and orientations, the system achieved a validation accuracy of 86.7% after 15 training epochs. This work presents a foundational approach for more complex FSL systems. The authors noted that the pre-trained and fine-tuned ResNet-50 model outperformed prior techniques that relied heavily on manual feature extraction or operated only under ideal conditions. Its adaptability and reduced training time make it suitable for future development in FSL recognition.

After 15 epochs of training, the system achieved 86.7% validation accuracy. This indicates it could properly identify most of the movements it learned from the training data. While the initial focus is on number signs, the researchers consider it is an effective foundation for developing more complex tools to meet the communication demands of the Filipino deaf community. The ResNet-50 machine learning model improved before techniques, which relied on human feature extraction or just functioned within ideal conditions, in terms of both accuracy and adaptability. Because the model was pre-trained and fine-tuned, it took less time to train while maintaining impressive results. This makes it feasible for future developments in sign language recognition. The authors propose to improve the system by include other indicators such as letters of service settings. This effort is a crucial first step toward developing accessible technology for the deaf in Philippines.

The same researcher, Montefalcon et al [14] created another research that focuses on the communication issues that the Deaf population in the Philippines has, which are mostly due to the general public's low comprehend of Filipino Sign Language (FSL). To help overcome this gap, the researchers created an automated system that recognizes FSL words using computer vision and deep learning techniques. The system uses MediaPipe Advanced to extract key details from video clips of three FSL signers studying Filipino texts. Those collected features—which include hand, body, and facial movements—are put into a Long Short-Term Memory (LSTM) neural network. This model is most appropriate for processing data that is ordered, such as sign language motions. The system was trained on a proprietary video dataset and was built to detect 15 continuous Filipino words in real time, making it a dynamic tool for understanding sign language as it develops naturally. In regards to performance, the LSTM-based model had an amazing average test accuracy of 94demonstrating its ability to recognize FSL. In real-world testing with 10 participants, the system achieved an average accuracy of 72.38% after two rounds, with an average forecast time of only 0.3 seconds. These findings demonstrate the model's stability, efficiency, and capacity to generalize between users. The comparison was also performed with different deep learning structures, especially Residual Networks (Res Net). The best-performing model, ResNet-34, obtained 87LSTM model outperformed it with greater accuracy, demonstrating its advantage in dealing with ordered patterns built into sign language communication.

Published systems for Filipino Sign Language (FSL) recognition remain limited. Existing works predominantly utilize Convolutional Neural Networks (CNNs) for the classification of static hand signs and Long Short-Term Memory (LSTM) networks for real-time recognition. However, there is a noticeable gap in research focused on detecting dynamic FSL gestures. This gap presents an opportunity to explore object detection techniques in combination with CNN-based classification to enhance recognition capabilities. Conducting a comprehensive evaluation of such systems can help identify key areas for improvement. These insights will be valuable in guiding future research efforts aimed at improving the performance and applicability of FSL recognition systems.

III. Methodology

This section of the paper discusses the development of the Filipino Sign Language (FSL) Alphabet recognition and detection model using deep learning. This section includes the data collection, data description, data preprocessing, the CNN architecture, experimental setup, and evaluation metrics used. in structuring the methodology to ensure the research flow is correct.

A. Data Collection

Fig. 1 shows sample letter gestures from the Filipino Sign Language (FSL) alphabet, arranged from left to right corresponding to letters A to Z. These samples are part of the dataset used for training and evaluating the CNN model. It consists of images representing Filipino Sign Language (FSL) Alphabet. These images were collected through both self-captured performances and online sources, primarily from a publicly available Kaggle dataset titled FSL Dataset [15]. The images images are validated to ensure it represents the correct sign for a letter. The images are manually anointed to represent its location in the image, with each image containing a bounding box marking the location of the hand sign. The annotations will include the class name (representing the specific FSL letter), as well as four numerical values describing the bounding box: the horizontal position (x), vertical position (y), width, and height. Due to the limited data in FSL, the dataset is synthetically expanded using augmentation techniques The dataset gathered is limited, in this case data augmentation technique can be utilized to artificially expand the size and diversity of a dataset by creating modified versions of existing data samples [16].The final gathered data is composed of 6800 images with each letter containing 208 images. 5460 is used for training and 1340 for training. Each image is associated with a text file containing its class label and annotation information.

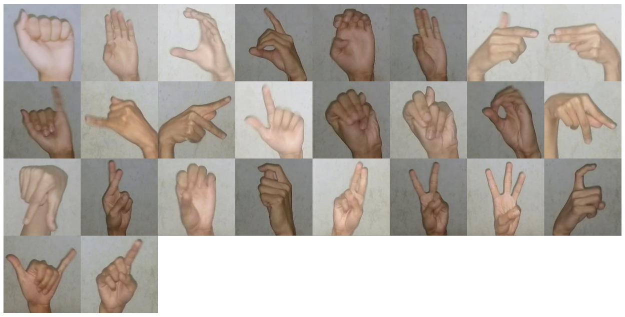


Fig. 1. FSL Alphabet

B. Data Preprocessing

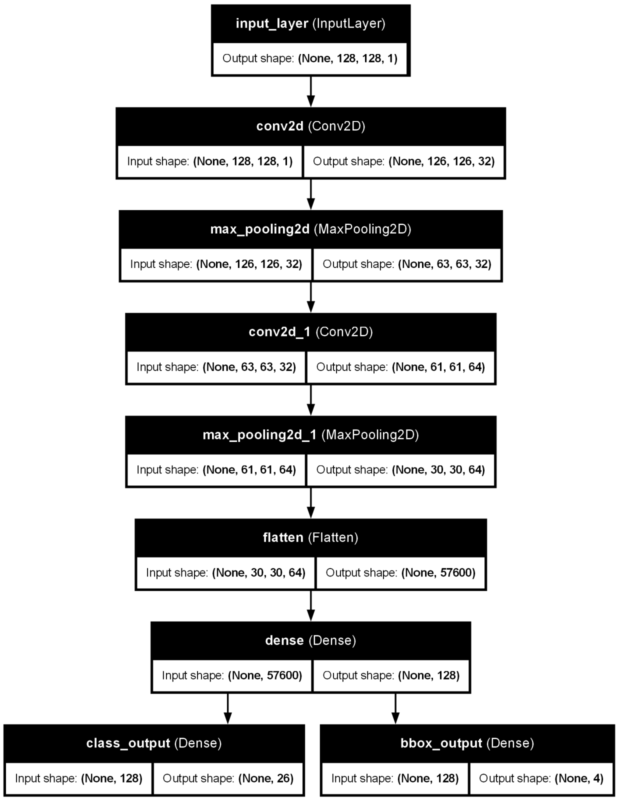
To prepare the image data for the CNN architecture, the following preprocessing techniques were implemented:

- The entire dataset was converted to grayscale to reduce the images to a single channel and simplify their complexity.
- A Gaussian blur filter was applied to help extract useful features from the hand gesture images.
- Image augmentation techniques, such as flipping, rotation, and brightness change, were used to balance the dataset.
- For consistency, images were resized to a fixed dimension of 128 × 128 pixels.
- To speed up model convergence, image pixel values were normalized to the range [0, 1].

C. Convolutional Neural Network

After undergoing preprocessing, Convulutional Neural Network or CNN is implemented A Convolutional Neural Network (CNN) is a specialized type of artificial neural network (ANN) optimized for extracting features from grid-structured data. It is especially effective for visual inputs like images and videos, where recognizing spatial patterns is essential. CNNs are extensively used in computer vision tasks because of their strong performance in analyzing and interpreting visual information [17].

As seen in Fig. 2, CNNs consist of multiple layers such as input layer, convolutional layer, pooling layer, and fully connected layers. For the classification and detection of sign language, the CNN network is customized to have two different outputs, one for the classification of the sign language, and the second for the detection of the sign gesture in the image using the anointed locations in the text files.



input_layer (InputLayer)
Output shape: **(None, 128, 128, 1)**
conv2d (Conv2D)
Input shape: **(None, 128, 128, 1)**
Output shape: **(None, 126, 126, 32)**
max_pooling2d (MaxPooling2D)
Input shape: **(None, 126, 126, 32)**
Output shape: **(None, 63, 63, 32)**
conv2d_1 (Conv2D)
Input shape: **(None, 63, 63, 32)**
Output shape: **(None, 61, 61, 64)**
max_pooling2d_1 (MaxPooling2D)
Input shape: **(None, 61, 61, 64)**
Output shape: **(None, 30, 30, 64)**
flatten (Flatten)
Input shape: **(None, 30, 30, 64)**
Output shape: **(None, 57600)**
dense (Dense)
Input shape: **(None, 57600)**
Output shape: **(None, 128)**
class_output (Dense)
Input shape: **(None, 128)**
Output shape: **(None, 26)**
bbox_output (Dense)
Input shape: **(None, 128)**
Output shape: **(None, 4)**

Fig. 2. CNN Architecture

D. Experimental Setup

In the initial experimentation, the goal is to establish benchmark scores of FSL letter recognition and detection model on FSL dataset. For this goal, the model will be evaluated based on its recognition rate and generalization ability. The chosen cross validation tech nique is a train test split of (80:20) wherein n in training set is equal to . The experiments were conducted on a Acer Nitro Laptop with Intel-I5 10th Gen and 8 Gb ram.

E. Evaluation Metrics

The proposed study aims to accurately recognize FSL gestures and identify their location within an image. To evaluate the model's performance in classifying FSL letter gestures, A common way is to check the training and validation accuracy [18]. The evaluation considers the standard classification indicators: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy is computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

To assess the model's capability to detect and accurately localize FSL gestures within an image, we employ the *Intersection over Union (IoU)* metric. IoU is a widely used evaluation criterion in object detection tasks that quantifies the degree of overlap between the predicted bounding box (B_p) and the ground truth bounding box (B_{gt}) [19]. It is defined as:

$$\text{IoU} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \tag{2}$$

A higher IoU value indicates a greater overlap between the predicted and actual bounding boxes, thus reflecting more precise localization. In general, an IoU score of 0.5 or higher is considered a correct detection in many object detection benchmarks, while values above 0.75are regarded as highly accurate. Therefore, IoU serves as a critical measure to evaluate the spatial precision of the model's gesture detection performance.

IV. Results and discussion

The FSL Alphabet recognition model, based on the CNN architecture, was trained using a multi-head approach, comprising two output heads: one for classification and another for bounding box regression. The classification head, responsible for recognizing FSL letters, was optimized using the cross-entropy loss function, while the detection head, tasked with predicting bounding box coordinates, utilized the mean squared error (MSE) loss. Both heads were jointly optimized using the Adam optimizer with a learning rate of 0.001. The model's performance was evaluated using classification accuracy for the letter recognition task and Intersection over Union (IoU) for the bounding box regression, in accordance with the specified evaluation metrics.

This section presents the results obtained before and after hyperparameter tuning. Following the tuning process, misclassifications and IoU scores were reassessed to analyze the improvements in model performance.

A. Model Accuracy

As shown in Table I, the initial results were gathered from the CNN model trained without data augmentation or Gaussian blur filtering. In this setting, the model achieved a training accuracy of 97% and a validation accuracy of 81.09%. This significant gap between training and validation performance indicates a clear case of overfitting, suggesting that the model memorized the training data—likely due to the high similarity among the input images—rather than learning generalizable patterns.

To address this issue, various data augmentation techniques were applied, including random rotations, brightness adjustments, hue variations, and Gaussian blur on a subset of the training data. These augmentations aimed to improve the model's generalization ability by diversifying the training samples. As presented in Table II, the augmented model achieved an improved training accuracy of 97.97% and a validation accuracy of 85.55%.

Although the validation performance still lags behind the training performance, the application of data augmentation techniques significantly reduced overfitting and enhanced the model's ability to generalize to unseen data.

TABLE I
INITIAL TRAINING AND VALIDATION ACCURACY PER EPOCH

Epoch	Training Accuracy (%)	Validation Accuracy (%)
1	25.39	62.71
2	83.60	76.19
3	93.60	78.10
4	96.18	79.46
5	97.86	80.79
6	97.00	81.09

TABLE II
TUNED TRAINING AND VALIDATION ACCURACY PER EPOCH

Epoch	Train Accuracy (%)	Validation Accuracy (%)
1	33.57	65.12
2	78.15	75.59
3	89.79	83.78
4	94.80	84.37
5	97.09	85.55
6	97.97	85.55

B. Missclassifications

An assessment of the CNN model after applying data augmentation techniques and Gaussian blur by the sixth training epoch is presented in Table III, which shows the misclassification rates for individual FSL letter classes. Notably, certain letters such as *A* and *B* exhibit very low misclassification rates of 2.94% and 4.90%, respectively. This may be attributed to the distinctiveness of their visual features, such as unique edge patterns and finger orientations, which make them easier for the model to differentiate.

In contrast, letters such as *S* and *W* show significantly higher misclassification rates, at 30.95% and 31.82%, respectively. A possible explanation is that the hand gestures corresponding to these letters share strong visual similarities with other gestures, leading to confusion during classification. Specifically, the letter *S* may be visually similar to multiple other signs, while *W* may share finger orientation patterns with other classes. Another contributing factor could be the model's sensitivity to slight variations in finger positioning, suggesting it is less robust to minor variations in gesture presentation.

TABLE III
Misclassification Rates per Class

Class	Rate (%)	Class	Rate (%)
A	2.94	N	11.63
B	4.90	O	11.63
C	8.82	P	19.05
D	10.42	Q	13.95
E	17.31	R	9.30
F	17.31	S	30.95
G	18.60	T	11.90
H	4.65	U	20.93
I	13.95	V	16.67
J	11.90	W	31.82
K	27.91	X	14.29
L	21.95	Y	16.67
M	19.05	Z	28.57

C. IoU Result

The Intersection over Union (IoU) results for the CNN model are presented in Table IV. Letter *A* achieved the highest IoU score of 76.27%, while letter *Z* had the lowest score of 62.23%. The overall average IoU across all letters was 69.4%. An IoU above 0.5 is generally considered acceptable for object detection models, while scores above 0.7 are often regarded as indicative of strong localization performance. Although the average IoU approaches 0.7, it is important to note that IoU metrics are influenced by the quality and consistency of the annotated data.

Since bounding boxes were manually annotated, variability in box size and placement can introduce inconsistencies in IoU calculations. For instance, some bounding boxes may be drawn larger or smaller, which can affect the overlap measurement. Additionally, the training data may be biased toward certain object positions, such as centered objects, while others may be slightly shifted, potentially impacting localization accuracy. Despite these factors, the model's IoU scores remain relatively high and consistent, suggesting it is effective for tasks such as hand gesture recognition.

D. Conclusion

This study presented a CNN-based model for recognizing and detecting Filipino Sign Language (FSL) letters. The dataset was compiled from multiple validated sources to ensure accurate representation of the letter gestures. After applying data augmentation techniques and a Gaussian blur filter, the model achieved a training accuracy of 97.97% and a validation accuracy of 85.55%.

TABLE IV
Average IoU per Class

Class	IoU (%)	Class	IoU (%)
A	76.27	N	68.59
B	70.75	O	71.30
C	76.02	P	71.14
D	72.37	Q	73.67
E	69.47	R	66.02
F	71.80	S	62.95
G	63.73	T	65.82
H	56.60	U	66.16
I	67.47	V	66.74
J	70.71	W	65.95
K	65.47	X	67.03
L	75.44	Y	66.51
M	69.50	Z	62.37
Overall Average IoU		**69.40**	

The analysis of the model's misclassification rates revealed significant errors in recognizing letters S and W, with misclassification rates of 30.95% and 31.82%, respectively. These results suggest that certain letters may share visual similarities or that the model is sensitive to slight variations in hand gestures. For detection, the regression head achieved an overall average Intersection over Union (IoU) of 69.4%, which is considered a strong result for a CNN-based approach.

Overall, the proposed system demonstrates good performance and serves as a solid baseline for future improvements. However, it is limited to static image recognition of individual letters. For future work, this system will be extended by incorporating temporal information through video frame analysis using models such as CNN+LSTM or transformer-based architectures, as some FSL letters involve motion-based gestures.

In addition, we aim to expand the dataset to include not only a greater number of examples of FSL letters, but also to incorporate the recognition of commonly used phrases, greetings, and eventually a wider range of signs. This will allow the system towards understanding and interpreting a more comprehensive portion of Filipino Sign Language. Further model optimization and fine-tuning techniques will also be explored to enhance the accuracy and generalizability of the system. This work provides a foundation for the development of more advanced FSL recognition tools, contributing to accessibility for the deaf community.

References

[1] J. Levy, ”Breaking the Sound Barrier: Understanding the Profound Impact of Hearing Loss on Communication,” *Hearing First*, Jan. 23, 2023. [Online]. Available: https://www.hearingfirst.co.uk/breaking-the-sound-barrier-understanding-the-profound-impact-of-hearing-loss-on-communication/

[2] ”Sign language,” *Encyclopædia Britannica*. [Online]. Available: https://www.britannica.com/topic/sign-language [Accessed: May 6, 2025].

[3] Philippine Senate, *Republic Act No. 7277: Magna Carta for Disabled Persons*. [Online]. Available: https://legacy.senate.gov.ph/lisdata/1868815815!.pdf [Accessed: May 6, 2025].

[4] Supreme Court of the Philippines, *Republic Act No. 11106: The Filipino Sign Language Act*, Dec. 20, 2018. [Online]. Available: https://elibrary.judiciary.gov.ph/thebookshelf/showdocs/2/85265 [Accessed: May 6, 2025].

[5] D. Lozada, ”Filipino Deaf community strives for inclusivity through Filipino Sign Language,” Rappler, 27 Jan. 2015. [Online]. Available: https://www.rappler.com/moveph/filipino-deaf-community-strives-inclusivity-filipino-sign-language/. [Accessed: 26 May 2025].

[6] M. Mairona-Basas and C. Pagliario, ”Technology use among adults who are deaf and hard of hearing: A national survey,” The Journal of Deaf Studies and Deaf Education, vol. 19, no. 3, pp. 400–410, Mar. 2014. [Online]. Available: https://academic.oup.com/jdsde/article/19/3/400/2937196

[7] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, et al., ”Sign language recognition, generation, and translation: An interdisciplinary perspective,” in Proc. 21st Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS), Pittsburgh, PA, USA, Oct. 2019, pp. 16–31.

[8] B. A. Al Abdullah, G. A. Amoudi, and H. S. Alghamdi, ”Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects,” IEEE Access, vol. 12, pp. 128871–128895, Sep. 2024, doi: 10.1109/ACCESS.2024.3457692. DBLP +4

[9] P. Bhatia and A. Wadhawan, “Sign language recognition systems: A decade systematic literature review,” Arch. Comput. Methods Eng., vol. 26, no. 4, pp. 785–813, 2019, doi: 10.1007/s11831-018-9288-4.

[10] P. M. Ferreira, J. S. Cardoso, and A. Rebelo, ”Multimodal learning for sign language recognition,” in Proc. Iberian Conf. Pattern Recognit. Image Anal., Cham: Springer, 2017, pp. 313–321.

[11] O. K. Oyedotun and A. Khashman, ”Deep learning in vision-based static hand gesture recognition,” Neural Comput. Appl., vol. 28, no. 12, pp. 3941–3951, Dec. 2017.

[12] A. Singh, B. Mathai, S. Silas, and J. Princess, ”Real time sign language translator for deaf and mute,” in *Proc. IEEE Int. Conf. Engineering, Research, and Computer Science (ICERCS)*, 2023, pp. 1–7. doi: 10.1109/ICERCS57948.2023.10433971.

[13] M. D. Montefalcon, J. Padilla, and R. Rodriguez, ”Filipino Sign Language Recognition using Deep Learning,” in *Proc. ACM Conf. Computing and Data Science (CoDS)*, 2021, pp. 219–225. doi: 10.1145/3485768.3485783.

[14] M. D. Montefalcon, J. Padilla, and R. Rodriguez, ”Sign Language Recognition of Selected Filipino Phrases Using LSTM Neural Network,” in *Proc. Int. Conf.*, Aug. 2022, pp. 633–641. doi: 10.1007/978-981-19-2397-5_56.

[15] J. A. Porton, ”FSL Dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/japorton/fsl-dataset

[16] P. Christiano, ”What Is Data Augmentation? The Complete Guide for 2024,” ExpertBeacon, Nov. 4, 2023. [Online]. Available: https://expertbeacon.com/data-augmentation/

[17] GeeksforGeeks, ”Convolutional Neural Network (CNN) in Machine Learning,” https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/ (accessed May 30, 2025).

[18] A. F. Gad and J. Skelton, ”How to Evaluate Deep Learning Models: Key Metrics Explained,” DigitalOcean, 2021. [Online]. Available: https://www.digitalocean.com/community/tutorials/deep-learning-metrics-precision-recall-accuracy. [Accessed: May 28, 2025].

[19] G. Boesch, ”What is Intersection over Union (IoU)?”, Viso.ai, Jan. 4, 2024. [Online]. Available: https://viso.ai/computer-vision/intersection-over-union-iou/. [Accessed: May 28, 2025].