Video Level Sign Language Recognition with Key Frame Extraction Using Adaptive Convolution Neural Networks with a New Activation Function

Navyasri Mullapudi
Department of Computer Science and Engineering, JNTUK, Andhra Pradesh, 533003, India
navyasrimullapudi@gmail.com (corresponding author)

G. Jaya Suma
Department of Information Technology, Gurajada University, Andhra Pradesh, 535003, India
gjssuma@gmail.com

Received: 12 August 2025 | Revised: 27 August 2025, 16 September 2025, and 18 September 2025 | Accepted: 21 September 2025

 DOI: https://doi.org/10.48084/etasr.14023

ABSTRACT

This paper proposes a deep learning architecture with a novel activation function in video-level sign language recognition. Samples from a video dataset of deaf-mute people were divided into multiple frames, and a new extraction algorithm is proposed in order to select and extract key frames from the videos. Adaptive Convolution Neural Networks (CNNs) utilizing a novel activation function were trained with the extracted video frames. The the high accuracy of the proposed method was verified in terms of precision, recall, f1-score, and accuracy.

Keywords-sign language recognition; convolution neural networks; video frame extraction; activation function

I. INTRODUCTION

Sign language is represented with hand movements and is commonly used for communication with deaf people [1], but the communication gap with the people that do not understand it, remains [2]. Sign language data are recorded either in image or in video forms, with real-time video-based data carrying more information than static images. Deep learning methods in general and Convolution Neural Networks (CNNs) [3] in particular are commonly used for image or video classification. The DeepASL [4] demonstrates accurate sign language recognition with a deep learning approach. Authors in [5] integrated a leap motion controller with CNNs in sign language recognition. This study provides insights into the sensor technology to improve hand movement recognition. Authors in [6] proposed a hybrid (CNN and RNN) architecture for sign language recognition. Authors in [7] surveyed various proposed methodologies, leaving practical implementation gaps. Authors in [8-10] conducted research on video-based recognition but did not apply effective key-frame selection techniques, resulting in redundant frames, increased computational cost, and decreased accuracy. Earlier works were mainly depended on traditional activation functions such as ReLU, Sigmoid, or Tanh [11-13]. The activation function determines the effectiveness of a deep learning model. The existing activation functions like ReLU suffer from issues such as the vanishing gradient or the dying ReLU.

To bridge this gap, the current research proposes: (i) a new key-frame extraction algorithm based on histogram correlation to choose only highly expressive frames that carry the information of the class label, and (ii) a novel activation function proposed to overcome the limits of traditional activation functions and improve the recognition accuracy.

II. METHODOLOGY

The proposed methodology for Video Sign Language Recognition (VSLR) encompasses a systematic approach designed to leverage the capabilities of CNNs and transfer learning algorithms. The outlined process involves several key phases, each contributing to the expansion of a robust and accurate VSLR system.

A. Dataset Collection

The video dataset plays a pivotal role in the proposed VSLR system, influencing the effectiveness, generalization, and robustness of the trained model. In the proposed methodology, the Indian sign language dataset from [14] is used. This dataset contains more than 4000 videos of 263-word signs. A total of 17 words were selected from the dataset. The Proposed Sign Dataset Labels are: {BEAUTIFUL, BED, BOOK, CAT, CHAIR, COW, DOOR, DREAM, FISH, KITCHEN, LOUD, PEN, PHOTOGRAPH, SOAP, TABLE, UGLY, WINDOW}.

B. Key Frame Extraction

A video is a collection of sequential frames, but not every frame can be used to train the model. Key frame extraction [15, 16] is a process where significant frames are selected from a video to represent essential information or moments. These key frames capture the essence of the video content and are often used for summarization, indexing, or other applications.

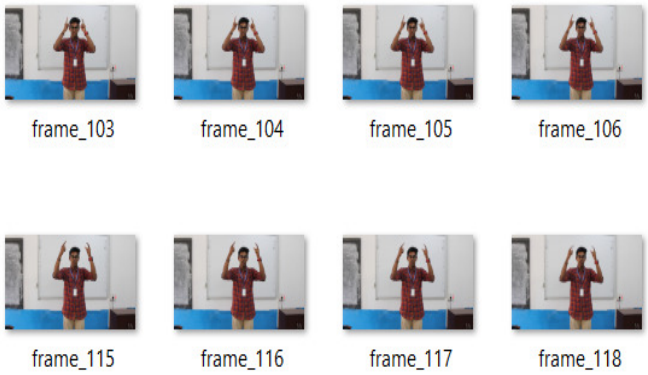

Fig. 1. Sample video frames of the word COW.

There are various key frame extraction methods. In the proposed methodology, histogram with correlation is applied to select the key frames. Histogram-based methods analyze the color distribution in a video and extract key frames based on the color diversity as given below:

- Collect the video samples. Let V be the Video.
- Extract all frames from the video V=F_1, F_2, F_3, F_4…. F_N where N is the total number of frames
- Let F_k be the reference frame.
- Calculate the correlation between the F_i frame and the F_k frame, where i=0,…, N. Figure 2 describes the comparison between two frames and Figure 3 gives the correlation value between the frames.
- The frame with the highest correlation value and the lower Mean Square Error (MSE) will be selected as the key frame.

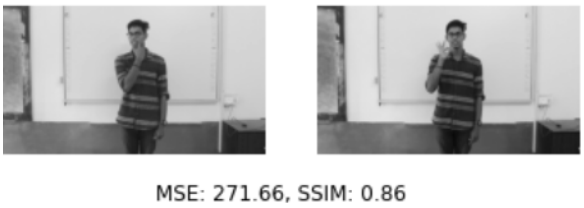

Fig. 2. MSE and Structure Similarity Index Measure (SSIM) between two frames.

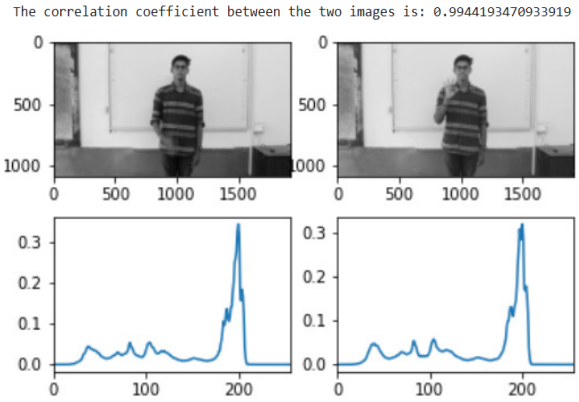

Fig. 3. Correlation calculation between F_i and F_k frames for the word BEAUTIFUL.

C. Model Training

The Proposed Methodology uses a CNN [17, 18] trained on the considered dataset, with a 80-20 split ratio, i.e. 80% of the data were used for training and 20% for testing. The tuned hyper parameter values are listed in Table I.

TABLE I. HYPER PARAMETER VALUES

Hyper Parameter	Set-Value
Epochs	15
Batch size	32
Optimizer	Adam
Dropout	0.1
Learning Rate	0.0001

D. Convolutional Neural Networks

A CNN is a specialized kind of artificial neural network particularly effective in image recognition tasks. The network consists of different layers, including convolutional layers which apply filters to input image dataset, detecting spatial hierarchies of features.

- Convolutional Layer: The convolutional layer [19] is responsible for extracting features from the input data. In image processing, these features might be edges, textures, or other visual patterns. Convolution involves sliding small kernels over the image data. At each and every position, the kernel performs element-wise multiplication with the input matrix, and the outcomes are summed to generate a feature map. Multiple kernels are used to acquire different features.
- Novel Activation Function: After the convolution operation, an activation function is applied element-wise to initiate non-linearity into the classification model. This makes the network to learn convoluted patterns and relationships in the data. ReLU (Rectified Linear Unit) is a commonly used activation function which often leads to the dying zero problem. The ReLU activation normalizes all the negative values in the feature map into zero while positive values will not be changed. In the proposed adaptive convolution model, a new activation function is used in order to reduce the dying zero problem.

$$f(x) = \begin{cases} x, & if\ x > 0 \\ \frac{e^{\alpha * x}}{\sqrt{1+x^2}}, & if\ x < 0 \end{cases} \quad (1)$$

- Pooling Layer (Max Pooling): Pooling layers reduce the spatial dimensions of the feature maps, decreasing computational complexity and controlling overfitting by focusing on the most essential information. Max pooling is used in the proposed model. It flattens pooled feature maps into a 1D vector.
- Fully Connected (Dense) Layer: Fully connected layers, also known as dense layers, learn to combine high-level features from the previous layers for making predictions. Neurons in a fully associated layer are bonded to all neurons in the preceding layer, forming a densely connected network. These connections allow the model to capture intricate relationships between features.
- Softmax Activation: The softmax activation function [20, 21] is typically applied to the output layer. Softmax converts large values to larger and small values to smaller. The soft-max function exponentiates each score and normalizes them to obtain a probability distribution across classes.

The architecture of the utilized CNN for video level sign language is shown in Figure 4.

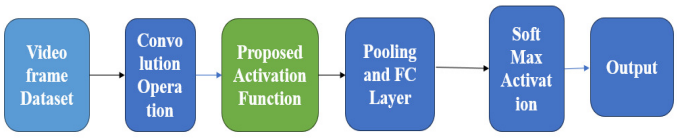

Fig. 4. Proposed architecture of video sign language classification using adaptive CNNs with the novel activation function.

- Mathematical properties of the proposed activation function:

The derivative of the proposed activation function is

$$f'(x) = \frac{\alpha e^{ax}\sqrt{1+x^2} - e^{ax}\frac{x}{\sqrt{1+x^2}}}{(1+x)^2} \quad (2)$$

The graph of the proposed activation function and its derivative are shown in Figure 5. As the curve shows, the proposed activation function is linear in the positive region and non-linear in the negative region. The derivative of the function is non-zero in both positive and negative regions.

The proposed activation function is continuous at $x > 0$ and monotonic and for negative values, it depends on the derivative. The output range of the proposed activation function is $(0, \infty)$. Unlike tanh and sigmoid activation functions [22, 23], the proposed activation function ensures stability for negative inputs. A comparison of the proposed activation function and ReLU can be seen in Table II.

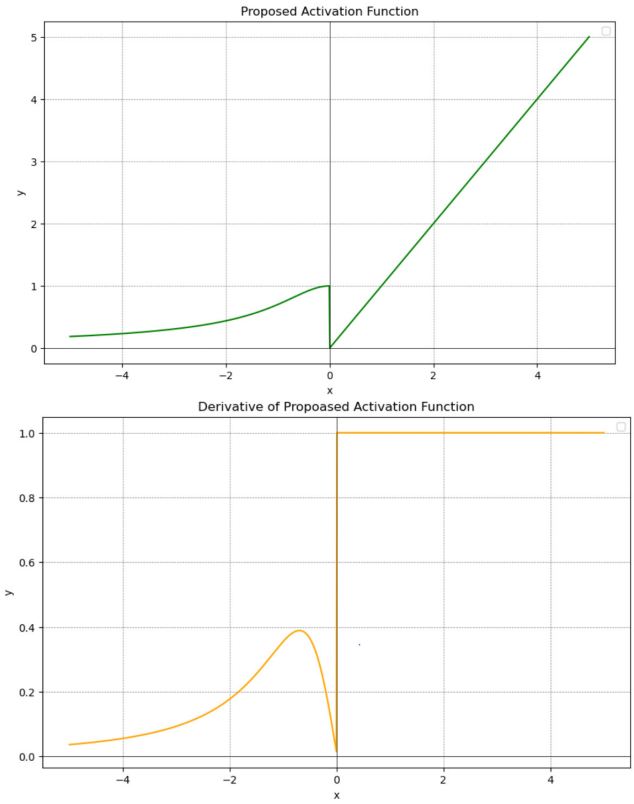

Fig 5. Graphical representation of the proposed activation function and its derivative.

TABLE II. COMPARISION BETWEEN RELU AND THE PROPOSED ACTIVATION FUNCTION FOR NEGATIVE VALUES OF x

f(x)	ReLU	Proposed
f(-1)	0	0.260
f(-2)	0	0.0605
f(-3)	0	0.0157
f(-4)	0	0.0044
f(-5)	0	0.0013
f(-10)	0	4.5×10^{-6}
f(-100)	0	3.7×10^{-46}
f(-500)	0	1.42×10^{-220}
f(-1000)	0	3.72×10^{-438}

III. PERFORMANCE EVALUATION

The performance of CNNs [24, 25] is commonly evaluated by Precision, Recall, F1-Score, and Accuracy. Randomly selected frames from the test dataset were used to examine the performance of the model. The above-mentioned evaluation metrics provide a comprehensive understanding of how well the models are performing in terms of different aspects of classification accuracy.

A. Precision

Precision is a evaluation metric used in classification model tasks to evaluate the correctness of positive predictions attained by a model. It is particularly relevant once the rate of False Positives (FP) is high. Precision is calculated as the ratio of True Positive (TP) predictions to the sum of TP and FP:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

B. Recall

Recall, also termed as sensitivity or true positive rate, is a performance metric used in classification tasks to estimate a model's ability to capture all the relevant positive instances. Recall is calculated as the ratio of TP predictions to the sum of TP and False Negatives (FN):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

C. F1-Score

F1-Score is the harmonic mean of Precision and Recall and provides a balanced degree of a model's performance, especially in situations where there is a disproportion between positive and negative instances.

$$\text{F1} - \text{Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (5)$$

D. Model Accuracy and Loss

Model accuracy graphs and model loss graphs are visual representations commonly used to evaluate classification models during training and evaluation. These graphs provide insight into how well the classification model is learning from the training data and testing the unseen data. The model accuracy graph illustrates the percentage of correctly classified instances by the model on both the training and validation datasets. Accuracy is defined by:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (6)$$

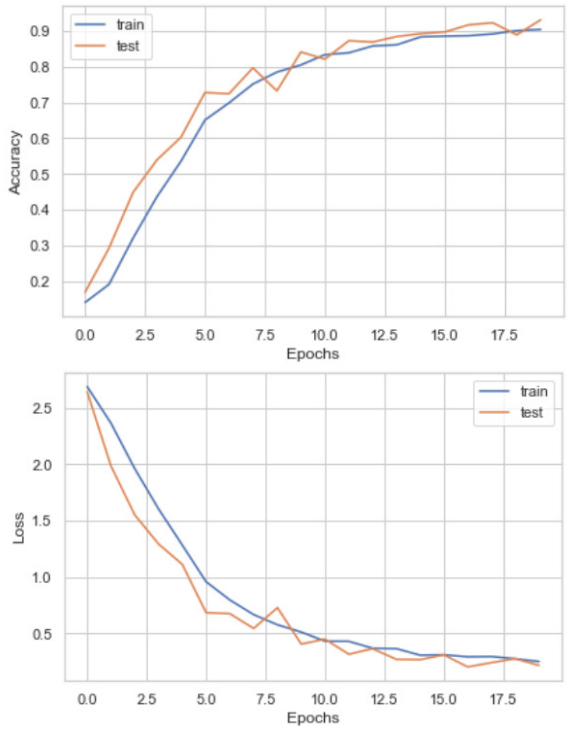

Fig. 5. Model accuracy and loss graphs.

E. Confusion Matrix

The confusion matrix provides a comprehensive view of the model's performance. Its diagonal elements are the TP of each class. The performance metrics are calculated based upon the values of the confusion matrix. Figure 7 shows the confusion matrix of the proposed model (multiclass classification).

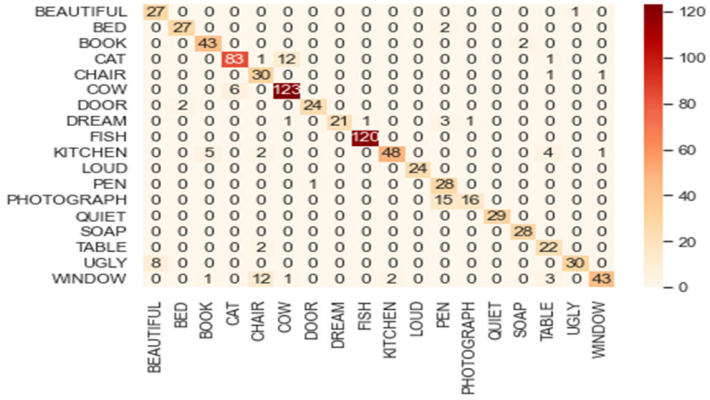

Fig 7. Confusion matrix of video level SLR using Adaptive CNN with Novel Activation Function

F. Summary

Table III shows the values of precision, recall and F1-Score of the proposed model for every class for the CNN with the proposed activation function.

TABLE III. PRECISION, RECALL AND F1-SCORE VALUES OF THE PROPOSED MODEL

Label Name	Precision	Recall	F1-Score
BEAUTIFUL	0.77	0.96	0.88
BED	0.93	0.93	0.83
BOOK	0.88	0.96	0.97
CAT	0.93	0.86	0.93
CHAIR	0.64	0.94	0.96
COW	0.9	0.95	0.97
DOOR	0.96	0.92	0.89
DREAM	1	0.78	0.71
FISH	0.99	1	1.00
KITCHEN	0.96	0.8	0.72
LOUD	1	1	0.97
PEN	0.58	0.97	0.73
PHOTOGRAPH	0.94	0.52	0.45
SOAP	1	1	0.97
TABLE	0.93	1	0.97
UGLY	0.71	0.92	0.81
WINDOW	0.97	0.79	0.93
Average Value	0.91	0.89	0.86

The proposed activation function proved to be highly accurate in comparison with known existing activation functions, as can be seen in Table IV.

TABLE IV. ACCURACY COMPARISION AT THE ACTIVATION FUNCTION LEVEL

Activation function	Accuracy
ReLU	86.123%
Tanh	84.321%
Sigmoid	86.19%
Proposed	91%

IV. CONCLUSION

In this study, the challenging problem of video-level sign language recognition has been addressed. The dying ReLU problem and the difficulty of identifying key frames have been addressed. A novel activation function is proposed, which utilizes an adaptive convolution neural network architecture and a key frame extraction algorithm was developed to identify the key frames that carry the respective sign information.

The experimental results proved that the proposed model achieved high accuracy of the order of 91%, outperforming the same CNN architecture utilizing traditional activation functions. The future scope of this research is to extend the proposed model to real-time sign recognition of continuous sign language including sentence-level and context-level gestures, providing more natural interaction systems. This paper is a significant step toward a robust, efficient, and accurate sign language recognition system which will have a high impact on the accessibility and communication opportunities of deaf people.

REFERENCES

[1] N. K. Kahlon and W. Singh, "Machine translation from text to sign language: a systematic review," *Universal Access in the Information Society*, vol. 22, no. 1, pp. 1–35, Mar. 2023, https://doi.org/10.1007/s10209-021-00823-1.

[2] A. H. Kugate *et al.*, "Efficient Key Frame Extraction from Videos Using Convolutional Neural Networks and Clustering Techniques," *EAI Endorsed Transactions on Context-aware Systems and Applications*, vol. 10, Jul. 2024, https://doi.org/10.4108/eetcasa.5131.

[3] A. Ajit, K. Acharya, and A. Samanta, "A Review of Convolutional Neural Networks," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, Oct. 2020, pp. 1–5, https://doi.org/10.1109/ic-ETITE47903.2020.049.

[4] B. Fang, J. Co, and M. Zhang, "DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, New York, NY, USA, Aug. 2017, pp. 1–13, https://doi.org/10.1145/3131672.3131693.

[5] A. Khan *et al.*, "Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive Review," *IEEE Access*, vol. 13, pp. 55524–55544, 2025, https://doi.org/10.1109/ACCESS.2025.3554046.

[6] S. Renjith and R. Manazhy, "Indian Sign Language Recognition: A Comparative Analysis Using CNN and RNN Models," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India, Dec. 2023, pp. 1573–1576, https://doi.org/10.1109/ICCPCT58313.2023.10245525.

[7] K. Nimisha and A. Jacob, "A Brief Review of the Recent Trends in Sign Language Recognition," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, Jul. 2020, pp. 186–190, https://doi.org/10.1109/ICCSP48568.2020.9182351.

[8] R. K. Vasanthakumari, R. V. Nair, and V. G. Krishnappa, "Improved learning by using a modified activation function of a Convolutional Neural Network in multi-spectral image classification," *Machine Learning with Applications*, vol. 14, Dec. 2023, Art. no. 100502, https://doi.org/10.1016/j.mlwa.2023.100502.

[9] B. Khagi and G.-R. Kwon, "A novel scaled-gamma-tanh (SGT) activation function in 3D CNN applied for MRI classification," *Scientific Reports*, vol. 12, no. 1, Sep. 2022, Art. no. 14978, https://doi.org/10.1038/s41598-022-19020-y.

[10] R. Avenash and P. Viswanath, "Semantic Segmentation of Satellite Images using a Modified CNN with Hard-Swish Activation Function," presented at the 14th International Conference on Computer Vision Theory and Applications, Nov. 2025, pp. 413–420, Nov. 03, 2025, https://doi.org/10.5220/0007469604130420.

[11] I. D. Khan, O. Farooq, and Y. U. Khan, "Automatic Seizure Detection Using Modified CNN Architecture and Activation Layer," *Journal of Physics: Conference Series*, vol. 2318, no. 1, Dec. 2022, Art. no. 012013, https://doi.org/10.1088/1742-6596/2318/1/012013.

[12] R. ZahediNasab and H. Mohseni, "Neuroevolutionary based convolutional neural network with adaptive activation functions," *Neurocomputing*, vol. 381, pp. 306–313, Mar. 2020, https://doi.org/10.1016/j.neucom.2019.11.090.

[13] V. S. Bawa and V. Kumar, "Linearized sigmoidal activation: A novel activation function with tractable non-linear characteristics to boost representation capability," *Expert Systems with Applications*, vol. 120, pp. 346–356, Apr. 2019, https://doi.org/10.1016/j.eswa.2018.11.042.

[14] P. Shet, M. Srinivas, C. Madhav, and R. Likhith, "Indian Sign Language Video Dataset." [Online]. Available: https://www.kaggle.com/datasets/prasadshet/indian-sign-language-video-dataset.

[15] M. S. Aiswarya and R. Arockia Xavier Annie, "Keyframe Extraction Algorithm for Continuous Sign-Language Videos Using Angular Displacement and Sequence Check Metrics," *International Journal of Intelligent Systems*, vol. 2024, no. 1, 2024, Art. no. 4725216, https://doi.org/10.1155/2024/4725216.

[16] N. Devabathini and P. Mathivanan, "Sign Language Recognition Through Video Frame Feature Extraction using Transfer Learning and Neural Networks," in *2023 International Conference on Next Generation Electronics (NEleX)*, Vellore, India, Sep. 2023, pp. 1–6, https://doi.org/10.1109/NEleX59773.2023.10421383.

[17] M. Navyasri and G. J. Suma, "Digit Recognition of Hand Gesture Images in Sign Language Using Convolution Neural Network Classification Algorithm," in *Recent Advances in Electrical and Electronic Engineering*, (ICSTE 2023), 2024, pp. 337–345, https://doi.org/10.1007/978-981-99-4713-3_32.

[18] M. Navyasri and G. J. Suma, "A novel key frame extraction using a deep learning model for sign language recognition on videos," *Multimedia Tools and Applications*, Oct. 2025, https://doi.org/10.1007/s11042-025-21134-0.

[19] L. Xiangyang, Q. Xing, Z. Han, and C. Feng, "A Novel Activation Function of Deep Neural Network," *Scientific Programming*, vol. 2023, no. 1, 2023, Art. no. 3873561, https://doi.org/10.1155/2023/3873561.

[20] P. Bohra, J. Campos, H. Gupta, S. Aziznejad, and M. Unser, "Learning Activation Functions in Deep (Spline) Neural Networks," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 295–309, 2020, https://doi.org/10.1109/OJSP.2020.3039379.

[21] M. A. Mohamed, H. A. Hassan, M. H. Essai, H. Esmaiel, A. S. Mubarak, and O. A. Omer, "Modified state activation functions of deep learning-based SC-FDMA channel equalization system," *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, no. 1, Nov. 2023, Art. no. 115, https://doi.org/10.1186/s13638-023-02326-4.

[22] A. O. Hashi, S. Z. M. Hashim, and A. B. Asamah, "Dynamic Adaptation in Deep Learning for Enhanced Hand Gesture Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15836–15841, Aug. 2024, https://doi.org/10.48084/etasr.7670.

[23] W. H. Lee, J. L. Tan, Z. A. A. Salam, H. Y. Teoh, Q. J. Lee, and L. T. S. Suzanne, "Sign language recognition based on CNN with optimized activation function," *Journal of Applied Technology and Innovation*, vol. 8, no. 1, pp. 9–14, 2024.

[24] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks," in *Computer Vision - ECCV 2014 Workshops*, 2015, pp. 572–578, https://doi.org/10.1007/978-3-319-16178-5_40.

[25] Md. M. Rahman, Md. S. Islam, Md. H. Rahman, R. Sassi, M. W. Rivolta, and M. 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)*, Dhaka, Bangladesh, Sep. 2019, pp. 1–6, https://doi.org/10.1109/STI47673.2019.9067974.