HandSpeak: A CNN-Based Deep Learning Framework for Sign Language Recognition

Banala Saritha
NVIDIA-CoE of Artificial Intelligence & Machine Learning
Department of Electronics and Communication Engineering
B V Raju Institute of Technology
Medak, India
saritha.b@bvrit.ac.in

G Purnachandrarao
Department of Electronics and Communication Engineering
BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India
pcrphd@gmail.com

Bandi Meghana
NVIDIA-CoE of Artificial Intelligence & Machine Learning
Department of Electronics and Communication Engineering
B V Raju Institute of Technology
Medak, India
22211a0427@bvrit.ac.in

Digwala Srivardhan
NVIDIA-CoE of Artificial Intelligence & Machine Learning
Department of Electronics and Communication Engineering
B.V. Raju Institute of Technology
Medak, India
22211a0463@bvrit.ac.in

Siddula Venkatesh
NVIDIA-CoE of Artificial Intelligence & Machine Learning
Department of Electronics and Communication Engineering
B.V. Raju Institute of Technology
Medak, India
22211a04N2@bvrit.ac.in

Rabul Hussain Laskar
Department of Electronics and Communication Engineering
National Institute of Technology Silchar Assam,
India
rhlaskar@ece.nits.ac.in

2025 Artificial Intelligence and Smart Technologies for Sustainability Conference (AISTS) | 979-8-3315-9852-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/AISTS66100.2025.11233056

***Abstract*— In the modern era, sign language plays a vital role in facilitating communication for hearing and speech-impaired individuals. However, a communication barrier still exists between sign language users and the general population. In this work, we propose a real-time sign language recognition system HandspeakNet, a deep learning framework using a convolutional neural network (CNN) architecture. The objective is to classify hand gestures corresponding to sign language alphabets with high accuracy using image data. The system is trained and tested on American sign language datasets, demonstrating significant potential for use in assistive technologies and human-computer interaction. The proposed system achieves a ~ 5% improvement in recognition accuracy over current state-of-the-art methods.**

Keywords— Deep Learning, Convolutional Neural Network, Hand Gesture Classification, Image Processing, Real-Time Recognition, American Sign Language.

I. INTRODUCTION

Communication is a fundamental human need, yet millions of people worldwide face challenges due to hearing and speech impairments. Sign language is the primary mode of communication for many individuals in the deaf and dumb community [1]. However, most of the population does not understand sign language, creating a significant communication barrier in education, employment, and social interaction.

Recent advancements in artificial intelligence (AI), particularly in deep learning, have opened new avenues for automatic sign language recognition. Traditional approaches relied on wearable sensors or manual feature extraction techniques, which were often costly, intrusive, or limited in flexibility. In contrast, deep learning enables the automatic extraction of relevant features from raw input data, such as images or video, providing a more robust and scalable solution. Convolutional Neural Networks (CNNs), a class of deep learning models particularly effective in image recognition tasks, have shown remarkable success in various domains, including medical imaging, facial recognition, and autonomous vehicles [2]. Leveraging CNNs for sign language recognition allows for real-time gesture classification using only a camera and a trained model, without the need for specialized hardware. The proposed model for sign language recognition is built on the foundation of Convolutional Neural Networks, a class of deep learning models particularly well-suited for visual data. CNNs are capable of learning hierarchical feature representations directly from raw pixel data, making them ideal for gesture recognition tasks where spatial and texture features are critical [3].

II. LITERATURE SURVEY

This section reviews prominent studies in the domain of Sign Language recognition, emphasizing various deep learning (DL) techniques aimed at enhancing prediction accuracy and efficiency in Gesture-Recognition. Mahidar et al. [4] used a CNN followed by a temporal pooling layer for real-time gesture recognition. Rioux-Maldague et al.[5] introduced a novel method for extracting features aimed at recognizing hand poses, utilizing both depth and intensity data acquired from a Microsoft Kinect sensor. They further implemented this approach for classifying American Sign Language fingerspelling through the use of a Deep Belief Network. Pioneered the use of CNNs in sign recognition for temporal tasks. In 2017, Ameen and Vadera et al. [6] Investigated the use of deep learning techniques for sign language interpretation, designing a convolutional neural network to classify fingerspelling images by leveraging both intensity and depth information. Ma et al. [7] proposed a greedy strategy to determine the architecture of a Deep Belief Network tailored for gesture recognition tasks. Chong et al. [8] analyzed finger and hand movements to distinguish static gestures from dynamic ones, utilizing Support Vector Machines (SVM) and Deep Neural Networks (DNN) to perform the classification. Koller et al. [9] Combined CNNs with Recurrent Neural Networks (RNNs) for continuous sign language recognition. The RWTH-PHOENIX-Weather 2014 is achieved 82.9% frame-level accuracy. He is the first to demonstrate end-to-end deep learning models on continuous sign video data. Nagarajan et al. [10] Implemented a 3D

CNN for dynamic sign gesture recognition. Chalearn LAP IsoGD (dynamic gestures) as the dataset. Achieved a 86.4% accuracy on gesture sequences. Addressed spatial-temporal feature extraction using 3D convolution. Shankar et al.[11] Introduced a Transformer-based sign language recognition system for sequential gestures WLASL (Word-Level American Sign Language Dataset). Top-5 accuracy of 89.6% for large vocabulary gestures. Used self-attention for temporal dependency modeling across video frames. Singh et al. [12] Utilized 3D Convolutional Neural Networks (3D-CNNs) to model spatiotemporal features of dynamic hand gestures. Collected approximately 2,400 videos representing 20 dynamic ISL signs. Achieved an accuracy of 86.4% in recognizing dynamic gestures. Demonstrated the effectiveness of 3D-CNNs in capturing temporal dynamics inherent in sign language gestures. Aliya et al. [13] proposed user independent recognition system for American sign language alphabet using depth images captured from the low-cost Microsoft Kinect depth sensor.

III. Proposed Architecture

This work presents a HandspeakNet deep learning-based system for recognizing static hand gestures in American Sign Language (ASL) using CNNs. The model architecture consists of Input Layer, Convolutional Layers, Pooling Layers, Fully Connected Layers, Output Layer. This section deals with data acquisition and preprocessing, texture feature extraction, classification algorithm, model integration, and performance evaluation. The work flow of the proposed system is shown in figure 1.

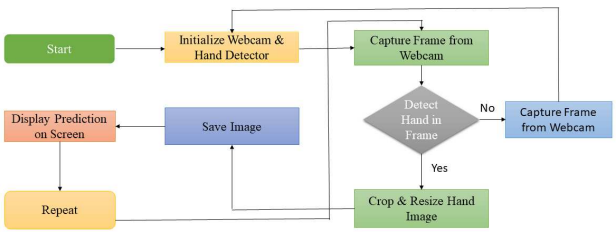

Fig 1. Workflow of sign language recognition using HandspeakNet.

Data Acquisition

We have considered an American sign language database [14] for this study, and a few sample gestures from this dataset are shown in Figure 2. This is the first step where data relevant to sign language gestures is collected, which serves as the foundation for further analysis and model training.

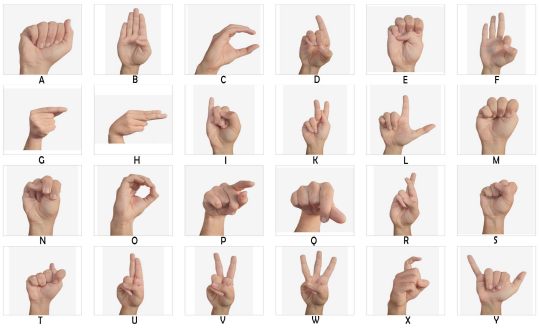

Fig. 2. Samples of sign language gestures

It involves in capturing videos or images of hand gestures representing signs. datasets custom recordings using PC cam [14]. The major challenges are Ensuring high quality, consistent lighting, and various sign types (static and dynamic). Several preprocessing steps were applied to ensure data quality and readiness for deep learning model.

Frame Extraction and Resizing

Dynamic sign languages involve sequences of movements rather than static hand poses. Hence, videos must be broken down into frames to capture temporal information. It Converts a video into a series of images (frames) that represent different points in time of a gesture. It is done by Using libraries like OpenCV to extract frames at a fixed rate. The redundant or blurry frames are discarded. Images must be standardized in size and pixel values to ensure consistent model input. Resizing is done because the Deep learning models require fixed-size inputs. Each frame is resized using bilinear or bicubic interpolation. It Reduces computational load and ensures uniformity.

Data Augmentation and Labeling

Data augmentation is the process of artificially increasing the dataset size and variability by applying transformations. Common techniques are used like Rotation, Flipping are Zooming, Cropping, Shifting ,Brightness Adjustment. The Purpose is to Prevent overfitting and make the model robust to real-world variations. Each frame or video sequence must be labeled correctly to train the model in a supervised manner. Static gestures are One label per image and Dynamic gestures are One label per video or frame sequence.

Splitting the Dataset

To evaluate performance fairly and prevent data leakage, the dataset is split into: training set 70% used to fit the model. validation set 15% used to tune hyper parameters and check for overfitting and test set 15%.

Model Training

In this work, HandspeakNet a deep learning architecture based on CNN, is developed to predict sign-gesture based on tabular data features. Each model was designed with varying architectures and training parameters to evaluate their effectiveness under different optimization strategies.

The HandspeakNet model summary is presented in table1. It is a CNN model [15-16] taking the input of size 224x224x3. It is adapted for structured tabular data by reshaping the input features to a one-dimensional format suitable for convolutional operations. Convolutional Neural Networks are used commonly in image-based applications like sign language recognition since it can learn automatically spatial hierarchies of features via backpropagation utilizing elements such as convolutional layers, pooling layers, and fully connected layers.

Table 1: HandspeakNet model summary

Layer Type	Output Size / Filters	parameters	Activation
Input Layer	(None, 224,224,3)	896	—
Conv2D Layer	32 filters, kernel size 3	0	ReLU
MaxPooling2D Layer ,	(None, 112,112,32)	0	—

	Pool size 2		
Conv2D_1 Layer	(None,112,112, 64)	18496	ReLU
MaxPooling2D_1 Layer	(None,56,56, 64)	0	—
Flatten Layer	(None,200704)	0	—
Dense	(None,128)		ReLU
Dropout	(None,128)	0	—
Dense_1(output)	(None,27)	3483	Softmax

The output after the last pooling layer is then flattened into a one-dimensional vector and fed through a fully connected dense layer comprising 256 neurons with ReLU activation in order to learn abstract high-level features. A dropout layer is next applied in order to avoid overfitting through random disabling of neurons during training. Lastly, the output layer is utilizing a softmax activation function with 26 units to produce class probabilities that relate to alphabet letters, which provides precise gesture classification. This model design offers a balanced method to feature extraction, dimensionality reduction, and classification, thus being best suited for tasks in sign language recognition.

In sign language recognition, ReLU enables the model to detect and refine important gesture features at each layer—starting from simple edges in early layers to complex hand shapes in deeper layers. Softmax It is used in the output layer. When a hand gesture image is input to the model, the softmax function at the output layer helps determine whether the gesture corresponds to 'A', 'B', ..., or 'Z' by assigning the highest probability to the correct class. Together, these activation functions make the model powerful and accurate for real-time sign language recognition.

Performance Metrics

To quantitatively evaluate the performance of the proposed deep learning model in predicting Hand gestures, several standard classification metrics were employed. In this work we considered the metrics were accuracy, precision, recall (sensitivity) and F1-Score [17-18]. Accuracy measures the proportion of correctly classified instances over the total number of predictions:

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

Where TP is True Positives, TN is True Negatives , FP is False Positives and FNis False Negatives. Precision measures the accuracy of positive predictions, defined as the ratio of true positive instances to the total number of predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A high precision score indicates that the model returns more relevant results and fewer false alarms, which is critical in loan approvals to avoid granting loans to ineligible applicants. Recall (Sensitivity) reflects the model's ability to detect all relevant instances, calculated as the ratio of true positives to the total number of actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In the context of loan prediction, a high recall ensures that eligible applicants are correctly identified, minimizing the rejection of valid requests. The F1-score is the harmonic mean of precision and recall, offering a balanced metric that is particularly valuable when dealing with imbalanced class distributions.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix The confusion matrix is a 2×2 matrix used to evaluate the performance of a binary classifier:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

This matrix offers detailed insight into the types of prediction errors made by the model, facilitating more targeted improvements.

IV. Experimental Settings and Performance Evaluation

This section presents a comprehensive evaluation of HandspeakNet model for the task of sign language recognition. The HandspeakNet is evaluated using a ASL on an NVIDIA RTX 3080 GPU with TensorFlow/Keras/OpenCv and VScode, for audio processing, and Scikit-learn for evaluation. The models were evaluated using multiple classification metrics, including precision, recall, F1-score.

Figure 3. Detection of sign language alphabet gestures using the proposed HandSpeak framework.

Performance of Convolutional Neural Network

The Convolutional Neural Network (CNN) model achieved perfect classification on the test dataset, which included the following sign classes: A, B, C, D, E, F, H, I, K and M. One of the predicted label doesn't matched the true labels, demonstrating the model's strong generalization and accuracy. This indicates that the CNN model is highly effective in recognizing static hand gestures for sign language with the current dataset. Accuracy is 0.9344, Precision is 0.9277, Recall is 0.9634 and finally the F1-Score is also 0.9163.

Table 3. Performance of HandspeakNet on test data

Accuracy	Precision	Recall	F1-Score
0.9346	0.9279	0.9635	0.9164

The performance of the CNN model for real-time sign language recognition was evaluated based on standard classification metrics including accuracy, precision, recall, and F1-score. The model achieved an overall accuracy of 93.44%, a precision of 92.77%, a recall of 96.34%, and an F1-score of 91.63%, indicating strong classification ability across the gesture classes.

Comparative Analysis and Observations

Table 4. Comparison of sign language recognition Accuracy with existing methods

Reference/year	Accuracy (%)
Rioux-Maldague et al.(2014) [2]	77
Ameen and Vadera et al. (2017) [3]	80.34
Ma et al. (2018)[4]	83.72
Chong and Lee et al. (2018) [9]	83.78
Aliya et al. (2019)[10]	88.7
Proposed	93.46

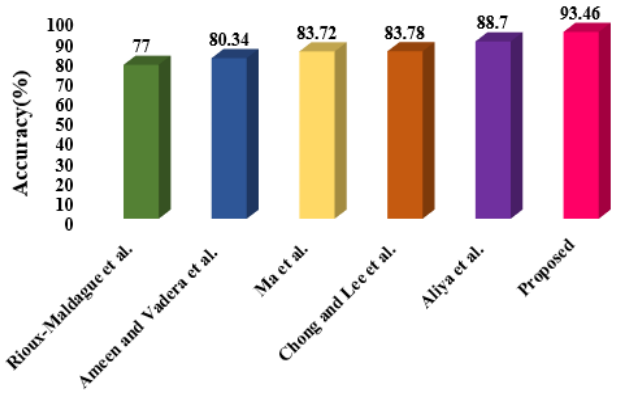

Figure 4. Accuracy comparison of existing systems with proposed model

Notably, the recall of 96.34% is a significant strength of the model, suggesting that it is highly effective at minimizing false negatives, which is critical in sign language interpretation where missed gestures can lead to loss of semantic meaning. The precision of 92.77% also confirms that most predicted signs were relevant, though the presence of some misclassifications slightly lowers the F1-score. Despite these challenges, the CNN model maintained a robust balance between precision and recall, demonstrating its capacity to generalize well across a variety of sign classes without extensive pre-processing.

Table 4 presents a comparative analysis of the proposed HandSpeak: A CNN-Based Deep Learning Framework for Sign Language Recognition against several state-of-the-art methods. Early work by Rioux-Maldague et al. (2014) achieved an accuracy of 77%, while subsequent improvements by Ameen and Vadera et al. (2017) and Ma et al. (2018) reported 80.34% and 83.72% respectively. More recent approaches by Chong and Lee et al. (2018) and Aliya et al. (2019) achieved accuracies of 83.78% and 88.7%, reflecting steady progress in the field.

In contrast, the proposed HandSpeak framework achieves a significantly higher accuracy of 93.46%, outperforming all previously reported methods. The significant improvement in accuracy of the proposed model is due to its lightweight CNN architecture with convolutional and max-pooling layers for feature extraction, followed by dense layers with ReLU, dropout, and softmax for effective classification.

V. CONCLUSION

This work presents and evaluates the effectiveness of a proposed HandspeakNet architecture for real-time sign language recognition from a webcam-based input system recorded an accuracy of 93.44%, precision of 92.77%, recall of 96.34%, and an F1-score of 91.63%, with strong performance in identifying varied hand gestures even under real-world scenarios like changing lighting, background noise, and hand positioning inconsistencies. The CNN architecture leveraged in this work employed the hidden layers utilized the Rectified Linear Unit (ReLU) activation function to add non-linearity so that the network could learn complex gesture patterns effectively, while the last classification layer utilized the Softmax activation function for multi-class prediction, enabling probabilistic multi-class prediction essential for distinguishing between multiple sign classes. The results affirm that CNNs are well-suited for gesture-based classification tasks, offering robust generalization and practical usability. The proposed HandspeakNet surpasses these with an accuracy of 93.46%, demonstrating a significant performance gain of over ~5% compared to the most recent comparable study. Future work may explore the impact of various optimization algorithms (such as Adam, RMSprop, and SGD) to identify the most robust training configuration for improving performance further in sign language recognition systems.

ACKNOWLEDGMENT

We sincerely thank electronics and communication department and management of BVRIT Narsapur for providing high NVIDIA – 3090Ti 24 GB GPU machine for our research work.

REFERENCES

[1] M. Al-Qurishi, T. Khalid and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," in IEEE Access, vol. 9, pp. 126917-126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[2] B. Saritha, M. A. Laskar, A. M. Kirupakaran, R. H. Laskar, M.Choudhury and N. Shome, "Deep learning-based end-to-end speaker identification using time-frequency representation of speech signal",Circuits Syst. Signal Process., vol. 43, pp. 1-23, Nov. 2023.https://doi.org/10.1007/s00034-023-02542-9 .

[3] Manakitsa, N., Maraslidis, G. S., Moysis, L., & Fragulis, G. F. (2024). A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision. *Technologies*, *12*(2), 15. https://doi.org/10.3390/technologies12020015

[4] B.Saritha, M.A.Laskar, RH.Laskar, R.H. (2023). A Comprehensive Review on Speaker Recognition. In: Biswas, A., Wennekes, E., Wieczorkowska, A., Laskar, R.H. (eds) Advances in Speech and Music Technology. Signals and Communication Technology. Springer, Cham. doi:10.1007/978-3-031-18444-4 1

[5] B. Saritha, R. H. Laskar, M. Choudhury et al., "Optimizing speaker identification through sincsquarenet and sincnet fusion with attention mechanism" Procedia Computer Science, vol. 233, pp. 215-225, 2024, doi: 10.1016/j.procs.2024.03.211.

[6] S. Ameen and S. Vadera, A convolutional neural network to classify American sign language ngerspelling from depth and colour images, Expert Syst., vol. 34, no. 3, Jun. 2017, Art. no. e12197.

[7] B. Saritha, M. A. Laskar, A. M. K. Anish, R. H. Laskar, and M. Choudhury, "CACRN-Net: A 3D log Mel spectrogram based channel attention convolutional recurrent neural network for few-shot speaker identification," Comput. Electr. Eng., vol. 115, Art. no. 109100, 2024. https://doi.org/10.1016/j.compeleceng.2024.109100

[8] B. Saritha, A. M. K, R. HussainLaskar and M. Choudhury, "FSIR: Few-Shot Speaker Identification using Reptile Algorithm," 2023 8th International Conference on Computers and Devices for Communication (CODEC), Kolkata, India, 2023, pp. 1-2, doi: 10.1109/CODEC60112.2023.10466164.

[9] B. Saritha, M. A. Laskar, A. M. Kirupakaran, R. H. Laskar,"ReptoNet: A 3D Log Mel Spectrogram-Based Few-Shot Speaker Identification with Reptile Algorithm." Arab J Sci Eng (2024). https://doi.org/10.1007/s13369-024-09426-3.

[10] Nagarajan, S., & Soundararajan, R. (2020). Dynamic Hand Gesture Recognition Using 3D-CNN and LSTM Networks. *Computers, Materials & Continua*, 70(3), 4985–5000. https://doi.org/10.32604/cmc.2022.014494

[11] B. Saritha, M. A. Laskar, R. H. Laskar and M. Choudhury, "Raw WaveformBased Speaker Identification Using Deep Neural Networks," 2022 IEEE SilcharSubsection Conference (SILCON), Silchar, India, 2022, pp. 1-4, doi:10.1109/SILCON55242.2022.10028890.

[12] Singh Seher, Abedi, K., Fatima, K., & Houda, T. (2024). 3D-CNN based Dynamic Gesture Recognition for Indian Sign Language Modeling. *ResearchGate*.

[13] B. Saritha, G. Purnachandrarao, B. Sravanthi, C. S. Reddy, K. Karthik and A. Harshitha, "Deep Learning-Based Optimized YOLOv5 for Enhanced Surface Defect Detection in Ceramic Tiles," *2025 7th International Conference on Signal Processing, Computing and Control (ISPCC)*, SOLAN, India, 2025, pp. 407-411, doi: 10.1109/ISPCC66872.2025.11039522.

[14] Aly, S. Aly, and S. Almotairi, User-independent American sign lan guage alphabet recognition based on depth image and PCANet features, IEEE Access, vol. 7, pp. 123138123150, 2019.

[15] B. Saritha, N. Shome, R. H. Laskar and M. Choudhury, "Enhancement in Speaker Recognition using SincNet through Optimal Window and Frame Shift," 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1-6. doi: 10.1109/CONIT55038.2022.9848231.

[16] B. Saritha, S. K. Sharma, T. Thiru Venkata Naga Manoj, A. M. K, R. Hussain Laskar and M. Choudhury, "DNN Based Speaker Identification System Under Multi-Variability Speech Conditions," *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, Naya Raipur, India, 2022, pp. 233-236, doi: 10.1109/WIECON-ECE57977.2022.10150728.

[17] Sevcikova Sehyr, Z., Caselli, N., Cohen-Goldberg, A. M., & Emmorey, K. (2021). The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Langua.

[18] B. Saritha, T. T. V. N. Manoj, S. K. Sharma, R. H. Laskar, M. Choudhury, and K. A. Monsley, "Intelligent speaker identification system under multi-variability speech conditions," in *Recent Advances in Electrical and Electronic Engineering*, B. P. Swain and U. S. Dixit, Eds., Lecture Notes in Electrical Engineering, vol. 1071, Springer, Singapore, 2024, doi: https://doi.org/10.1007/978-981-99-4713-3_35