ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983

# A Computer Vision and Deep Learning Framework for Sign Language Recognition

Yila Amon[1]; Sunday Iliya[2]; Poyi G. T.[3]

[1,2,3]Department of Electrical and Electronics Engineering University of Jos-Plateau State, Nigeria.

Publication Date: 2026/10/27

**Abstract: Sign language is an important means of communication for Deaf and Hard of Hearing individuals, yet limited knowledge of sign language among hearing people can create substantial communication barriers. This study presents a computer vision and deep learning framework for sign language training and recognition. A custom dataset was developed from 50 selected high-frequency words and a NoSign control class, with image recordings obtained from two signers. MediaPipe Holistic was used to extract facial, pose, and hand landmarks from the captured video. The resulting frame sequences were processed by a hybrid CNN-BiLSTM model that combines spatial feature extraction with temporal sequence modelling. The developed model achieved an overall recognition accuracy of 95% and a macro-average F1-score of 0.95. The system recorded an inference time of 0.0941s for a 30-frame sequence, corresponding to approximately 10.63 sequences per second and 25 Frames per Second. Recognized signs were converted to English speech to provide immediate feedback during practice. The results demonstrate the feasibility of a lightweight, webcam based approach for interactive sign language training and recognition.**

***Keywords:*** *Sign Language; Computer Vision; MediaPipe Holistic; CNN-BiLSTM; Deep Learning; Gesture Recognition; Training and Recognition.*

**How to Cite:** Yila Amon; Sunday Iliya; Poyi G. T. (2026) A Computer Vision and Deep Learning Framework for Sign Language Recognition. *International Journal of Innovative Science and Research Technology*, 11(10), 1857-1863. https://doi.org/10.38124/ijisrt/26oct983

## I. INTRODUCTION

Sign language is a visual language used by Deaf and Hard of Hearing people to communicate through coordinated hand movements, body posture, facial expression, and other non-manual cues. Meaning is conveyed through structured signs rather than speech, and the form of a sign may depend on hand configuration, orientation, movement, and the surrounding visual context [1],[2].

Communication difficulties can arise when hearing individuals have little or no knowledge of sign language. The problem can affect participation in education, healthcare, employment, and everyday social interaction. In Nigeria, the need for accessible learning resources is particularly important because learners and teachers may have limited access to interactive tools for practicing sign language.

Globally, hearing loss affects a large population and an estimation of about 22 million people in Nigeria, including approximately three million school age children have severe to profound hearing loss [3]. These figures reinforce the importance of communication-support and educational technologies that can improve access for Deaf and Hard of Hearing communities.

Existing sign language learning practices often depend on instructors, printed materials, demonstrations, or limited digital resources. While these methods can introduce vocabulary and signing techniques, they may not provide continuous feedback while a learner is practicing. An interactive computer-vision system can address this limitation by observing a performed gesture and returning an immediate prediction.

Many sign language recognition systems reported in the literature have also been developed around datasets from other sign language communities. Their signing patterns, and operating assumptions may therefore not directly correspond to the Nigerian educational setting [4]. There is consequently a need for locally oriented datasets and tools that support teaching, practice, recognition, and feedback.

In this study, the term Nigerian Sign Language (NSL) is used for the localized sign vocabulary adopted for the Nigerian educational context. The study therefore does not claim to represent every regional or community variety of signing used in Nigeria, rather, it focuses on the selected educational vocabulary used to construct the experimental dataset.

For an educational recognition system, the distinction between a general sign-language translator and a training tool is important. A translator primarily seeks to convert an

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983

observed sign into another representation, whereas a training oriented system can support repeated practice by providing immediate recognition and spoken feedback. The latter is the focus of this work. The system is intended to provide a practical computer vision interface through which a learner can perform a selected sign, receive the recognized English word, and use the response as feedback during practice.

This study develops a computer vision and deep learning framework for sign language training and recognition. The system captures a learner's gesture through a webcam, extracts body, facial, and hand landmarks with MediaPipe Holistic, and classifies a sequence of frames with a hybrid CNN-BiLSTM architecture. The recognized word is displayed converted into English speech, allowing the learner to receive both visual and auditory feedback.

The study uses 50 selected high-frequency words together with a NoSign control class. The resulting system is intended primarily as an interactive learning and training aid for teachers and students, while also demonstrating the feasibility of real-time sign recognition using a relatively lightweight model.

## II. LITERATURE REVIEW

➢ *Related Studies*

Sign language recognition has progressed from conventional image-processing methods to machine and deep learning approaches. Early systems commonly used handcrafted descriptors such as contours, shape features, invariant moments, SURF and Bag-of-Visual-Words with classifiers including Support Vector Machine, K-Nearest Neighbour, Naïve Bayes and Random Forest. These methods established the feasibility of automatic recognition but depended strongly on manually selected features and controlled imaging conditions[5], [6], [7], [8], [9], [10]

Wearable approaches use data gloves, flex sensors and related devices to capture finger and hand movements directly[11], [12], [13].Their main limitations are additional hardware requirements, cost, maintenance and reduced naturalness of interaction. Camera-based systems avoid these requirements and are therefore more suitable for accessible learning applications.

Deep learning has improved recognition by allowing models to learn spatial representations directly from visual data. CNNs are effective for frame-level spatial feature extraction, while LSTM-based networks model temporal relationships across successive observations[14], [15], [16]. MediaPipe-based landmark representations provide a compact alternative to processing complete RGB frames and are useful when a lightweight real-time system is required.

For dynamic word recognition, spatial and temporal modelling are complementary. A CNN can learn the configuration of a gesture at individual time steps, whereas a BiLSTM can learn how those features evolve throughout a sequence. Previous LSTM, LSTM-GRU and skeletal/landmark-based studies have reported encouraging results[17], [18], [19] [17]–[19].. These findings support the CNN-BiLSTM design adopted in this study.

Image based deep networks can learn rich visual representations but may require greater computational resources, whereas landmark based representations reduce the input dimensionality and can be more appropriate for real time educational applications. The present framework follows the latter direction by extracting structured body, face and hand landmarks before temporal classification. This design provides a direct link between the visual gesture sequence and the temporal learning component while avoiding dependence on a large end-to-end image model.

Despite the progress reported in the literature, three limitations are relevant to the present work. First, many datasets represent American, Indian, Arabic or other sign languages, while locally developed resources for Nigerian educational applications remain limited[3], [15]. Second, many systems concentrate on recognition or translation rather than learner centered practice with immediate feedback. Third, relatively few lightweight systems combine landmark-based computer vision, temporal deep learning and speech output in one educational application.

This study addresses these gaps through a custom word-level dataset, MediaPipe Holistic landmark extraction, a CNN-BiLSTM recognition model and English speech feedback. Its contribution is therefore the integration of recognition and immediate feedback into a training oriented system intended for teachers and students.

## III. MATERIALS AND METHOD

➢ *Materials*

The system was implemented in Python using OpenCV for video acquisition, MediaPipe Holistic for landmark extraction, TensorFlow for model development, NumPy for numerical processing, and pyttsx3 for speech synthesis. A webcam provided video input and an external speaker provided audio feedback.

The software environment comprised OpenCV, MediaPipe Holistic, TensorFlow, NumPy and pyttsx3.

➢ *System Architecture*

The overall processing sequence begins with webcam video acquisition. Each captured sequence is processed to obtain landmark features, which are supplied to the CNN-BiLSTM model. The model generates a gesture-class prediction, after which the recognized class is displayed and converted to speech for real-time feedback.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983

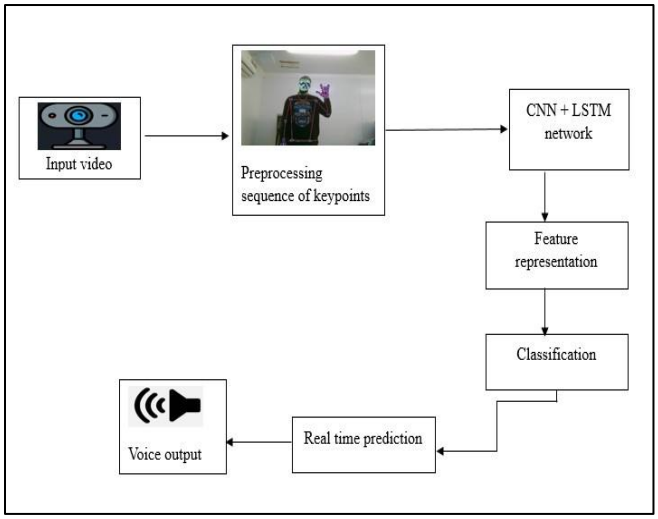


Fig 1 The Flow Chart of the System Processing Architecture

➢ *Dataset Development and Data Acquisition*

A custom NSL dataset was created from 50 selected high-frequency English words and a NoSign control class to make a total of 51 classes. The word list was derived from the high-frequency word resources cited in [20] and [21]**,** and the first 50 words were selected for the study. The signs recorded for these words represent the localized Nigerian educational vocabulary described above. Two signers participated in data collection. Each signer performed 20 sequences for each class, with every sequence containing 30 frames. This produced 2,040 sequences and 61,200 frames across the 51 classes. The dataset was designed to capture temporal gesture information rather than isolated still images.

The inclusion of the NoSign class to make a total of 51 classes was intended to provide a non-target state for the recognition pipeline. This is relevant to continuous interaction because a learner does not perform a target word at every moment. A non-sign class can therefore help the application distinguish an intended vocabulary gesture from periods in which no target gesture is being presented.

The selected 50 words are presented in Table 1. During data collection, attention was given to signer variation, lighting, background, and consistency of the performed gestures. The experimental setup consisted of a High Definition webcam connected to a laptop as shown in Fig 2, with the captured video processed by OpenCV and MediaPipe. The recorded frame sequences were stored on the computer for subsequent preprocessing and model development.

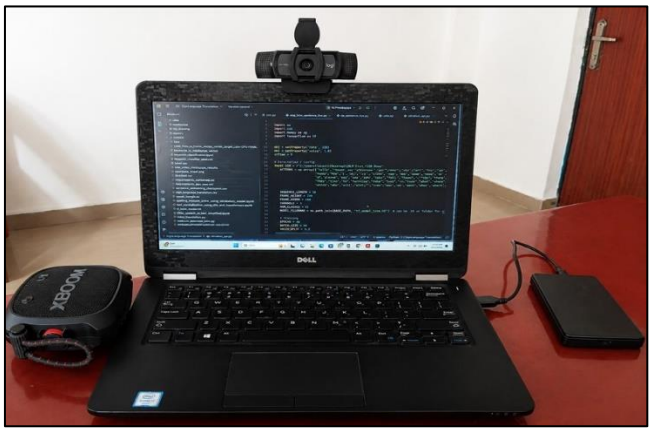

Fig 2 Set up for Image Capturing for Dataset Formation

Table 1 Adopted High Frequency Word List (Source: *[21]*)

| **Adopted List of High Frequency Words** | | | | |
|---|---|---|---|---|
| 1.the | 11.then | 21.his | 31.we | 41.far |
| 2.and | 12.said | 22.went | 32.cat | 42.saw |
| 3.a | 13.lion | 23.go | 33.when | 43.come |
| 4.he | 14.in | 24.that | 34.it | 44.monkey |
| 5.is | 15.there | 25.tortoise | 35.so | 45.tree |
| 6.to | 16.they | 26.food | 36.her | 46.father |
| 7.I | 17.eat | 27.rat | 37.once | 47.say |
| 8.my | 18.day | 28.children | 38.will | 48.school |
| 9.one | 19..man | 29.house | 39.are | 49.she |
| 10.was | 20.boy | 30.me | 40.dog | 50.you |

➢ *Preprocessing Using MediaPipe Holistic*

Each video frame was processed with MediaPipe Holistic to obtain landmarks from the body pose, face, left hand, and right hand. The extracted coordinates contain spatial information describing the detected landmark positions. The sequence was maintained as a 30-frame time series so that the model could learn changes in the gesture across time.

The raw landmark representation contains 33 pose landmarks represented with four values per landmark, 468 facial landmarks represented with three coordinates each, and 21 three-dimensional landmarks for each hand. This gives a raw feature vector of 1,662 elements per frame: (33 × 4) + (468 × 3) + (21 × 3) + (21 × 3) = 1,662.

For model input, the extracted sequence was converted into the fixed spatial representation used by the convolutional network, producing an input tensor of shape (30, 46, 36, 1). The processed sequences were stored in keypoints.npy file format, with image or video files retained where required for visualization.

➢ *Hybrid CNN-BiLSTM Model*

The recognition architecture combines a TimeDistributed 2D convolutional network with Global Average Pooling and stacked Bidirectional LSTM layers. The CNN processes the spatial representation at each time step, while the BiLSTM learns temporal dependencies across the 30-frame sequence. The final representation is passed through dense layers and a Softmax classifier.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983

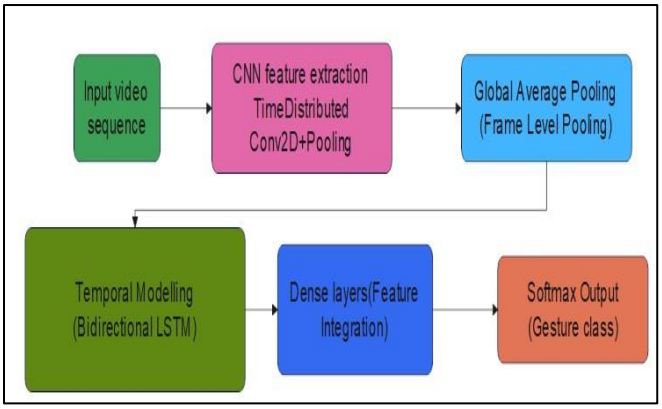


Fig 3 Work Flow of the Hybrid CNN-LSTM Model Layers

➢ *Input Representation*

Let $X_{raw,t}$ denote the feature vector at time *t*. It is formed by concatenating the Pose (P), Facial (F), Left-hand (L), and Right-hand (R) landmark features:

$$X_{raw,t} = [P_{33\times4} \parallel F_{468\times3} \parallel L_{21\times3} \parallel R_{21\times3}] \quad (1)$$

The resulting landmark vector contains 1,662 values for each frame. The sequence is then arranged into the fixed model input representation (30, 46, 36, 1), where 30 denotes the temporal length, 46 the spatial height, 36 the spatial width, and 1 the input channel.

➢ *Convolutional Feature Extraction*

The TimeDistributed convolutional layers apply the same convolutional feature extractor to each frame while preserving the temporal order. For a filter k at time t, the convolutional operation can be represented as:

$$M_t = \varphi\left(\sum_{m,n} K_{m,n}\right).I_t(i+m, j+n) + b_k \quad (2)$$

Where $M_t$ is the resulting feature map, $K_{m,n}$ denotes the convolution kernel, $I_t$ is the input representation at time *t*, $b_k$ is the filter bias, and $\varphi$ denotes the ReLU activation function.

Max-pooling layers reduce the spatial dimensions of the feature maps and remove less dominant spatial information. This reduces computational demand while retaining the strongest learned features.

➢ *Global Average Pooling*

Global Average Pooling converts each spatial feature map into a single value by averaging the activations over its spatial dimensions. For channel *c*, at time *t*:

$$V_{t,c} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,k} \quad (3)$$

Where $H$ and $W$ are the height and width of the feature map. The operation produces a compact feature vector for each frame and avoids the large parameter count that would result from flattening the entire feature map.

➢ *Temporal Modelling with BiLSTM*

The pooled frame-level representations are passed to stacked Bidirectional LSTM layers. An LSTM maintains a hidden state and a cell state and controls information flow through forget, input, and output gates. The bidirectional structure processes the sequence in both forward and backward directions, allowing the model to use information from the complete gesture sequence. The first bidirectional layer preserves the sequence structure for subsequent temporal processing, while the final recurrent representation summarizes the 30-frame gesture.

➢ *Classification*

The final representation is fed into a dense layers and a Softmax classifier. The classifier contains 51 output classes, representing the 50 target words and the NoSign control class. Dropout of 40% is applied to reduce overfitting.

## IV. RESULTS

➢ *Captured Gestures and Landmark Detection*

The recorded gestures were processed with MediaPipe Holistic, producing landmark representations for the face, body pose, and both hands. This representation preserves manual and non-manual information that can contribute to word-level recognition. The examples in Fig. 4, 5, 6 and 7 illustrate some of the captured gesture patterns for two of the classes (Food and Father) by the two signers. Using landmarks rather than raw image pixels also provides a compact representation suitable for the proposed real time recognition pipeline.

The landmark sequence also provides a common representation across the recorded samples. Rather than relying only on raw appearance, the model receives coordinates describing the spatial configuration of relevant landmarks at each frame. The sequence structure is then preserved for temporal modelling, allowing the recurrent component to learn changes in hand, body and facial configurations over time.


Fig 4 Sample Sign Gesture Representing Food by Signer 1

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983



Fig 5 Sample Sign Gesture Representing Food by Signer 2

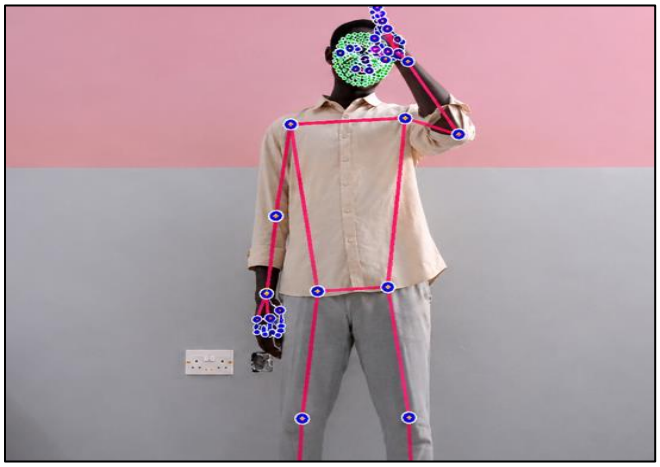


Fig 6 Sample Sign Gesture Representing Father by Signer 1

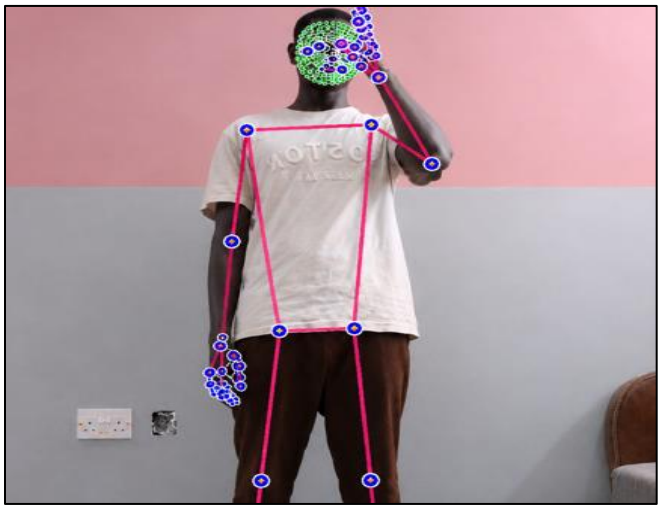


Fig 7 Sample Sign Gesture Representing Father by Signer 2

➢ *Model Accuracy and Loss*

The model was trained for 60 epochs using the training configuration described in the methodology. The training and validation accuracy curves show progressive learning, with performance approaching the reported final level toward the later epochs. The corresponding loss curves decrease as training progresses, indicating that the network learned discriminative spatial-temporal representations of the gesture sequences. The small separation between the training and validation trends suggests that the model did not exhibit severe overfitting under the experimental setting.

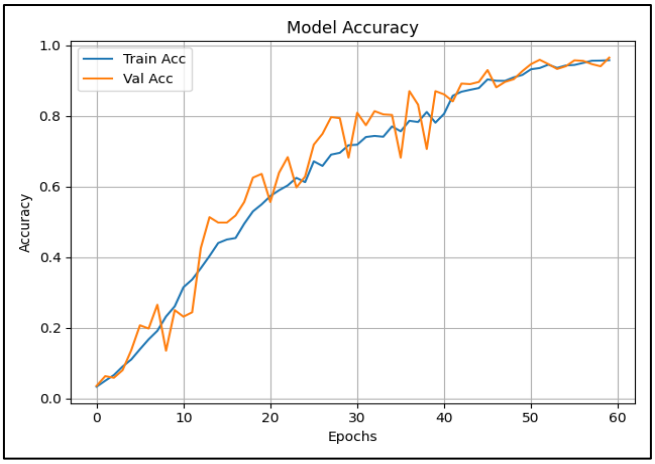



Fig 8 Model Accuracy Plot

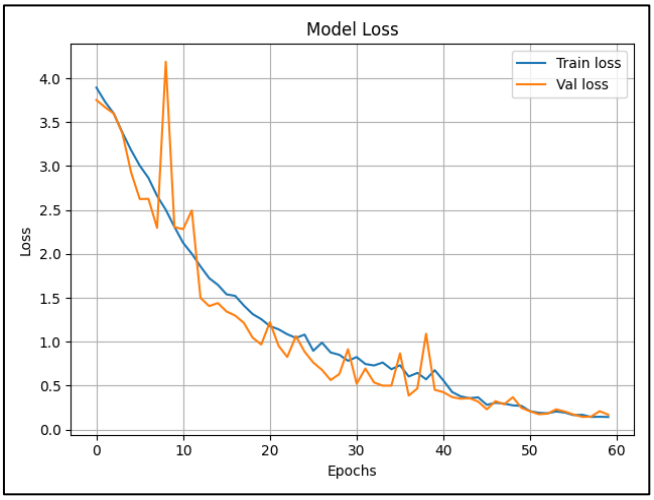



Fig 9 Model Loss Plot

The close relationship between the training and validation curves suggests that the model reached a stable learning state without prolonged divergence. Nevertheless, because only two signers were used for dataset development, stronger claims about generalization to unseen signers should be treated cautiously. A larger signer-independent evaluation would provide stronger evidence of generalization.

➢ *Real-Time Prediction*

After training, the model was integrated into the real-time application. Live webcam frames are processed through the same landmark extraction procedure used during dataset preparation. A sequence is then passed to the trained model to obtain a class prediction. The recognized sign is displayed as text and converted to English speech using pyttsx3.

A non-blocking audio mechanism was used so that speech generation does not unnecessarily interrupt visual processing and prediction. OpenCV overlays provide the user with the detected landmarks and recognition output during operation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983



Fig 9 Real Time Prediction Output Window

➢ *Performance*

The developed system achieved an overall recognition accuracy of 95% and a macro-average F1-score of 0.95. The reported inference time was 0.0941 s for a 30-frame sequence. This corresponds to approximately 10.63 sequences per second and 25 frames per second. The trained model occupied approximately 5.03 MB of storage. These measurements indicate that the model is relatively lightweight and capable of supporting real-time interaction on the development platform.

Misclassification was observed for some gestures with similar visual or movement characteristics. Such errors are expected in a word-level gesture-recognition task when classes share related hand positions or movement patterns. Expanding the dataset with more signers, additional examples, and more varied conditions would provide a stronger basis for assessing robustness.

## V. CONCLUSION

The study nevertheless has defined boundaries. The vocabulary contains a selected set of 50 words rather than the complete vocabulary used in Nigerian signing communities, and the dataset was collected from two signers. Consequently, the reported performance should be interpreted as performance on the developed experimental vocabulary and not as a measure of complete Nigerian Sign Language proficiency.

This study presented a computer vision and deep learning framework for sign language training and recognition. A custom dataset comprising 2,040 labelled video sequences and 61,200 frames was developed from 50 selected high-frequency words and a NoSign control class. MediaPipe Holistic was used to extract body, facial, and hand landmarks from the captured sequences.

A hybrid CNN-BiLSTM architecture was developed to combine spatial feature extraction with temporal sequence modelling. The CNN component learns spatial characteristics from individual frames, while the BiLSTM component models the temporal features across the 30-frame sequence. The model achieved an overall recognition accuracy of 95% and a macro-average F1-score of 0.95.

The application provides both visual and English audio feedback, enabling a learner to see the recognized sign and hear the corresponding word during practice. The measured inference time of 0.0941 s per 30-frame sequence, throughput of 10.63 sequences per second, and processing rate of 25 FPS indicate that the system can support real-time interaction. The model size of approximately 5.03 MB also indicates a relatively lightweight implementation.

The main limitation of the present study is the small number of participating signers. Future work should therefore expand the dataset to include more signers, broader vocabulary categories, different signing styles, investigate additional continuous signing scenarios and more varied environmental conditions. Future evaluations should also include signer-independent testing and investigate more advanced deep learning architectures. Such extensions would improve the robustness and educational coverage of the proposed framework.

## REFERENCES

[1.] K. Renz, N. C. Stache, S. Albanie, and G. Varol, “Sign language segmentation with temporal convolutional networks,” Feb. 12, 2021, arXiv: arXiv:2011.12986. doi: 10.48550/arXiv.2011.12986.

[2.] L. Zdrazilova, D. M. Sidhu, and P. M. Pexman, “Communicating abstract meaning: concepts revealed in words and gestures,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 373, no. 1752, p. 20170138, Jun. 2018, doi: 10.1098/rstb.2017.0138.

[3.] I. Nwogu, R. Peiris, K. Dantu, R. Gamta, and E. Asonye, “AI-Driven Sign Language Interpretation for Nigerian Children at Home,” in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 6395–6404. doi: 10.24963/ijcai.2023/710.

[4.] S. Kolawole, O. Osakuade, N. Saxena, and B. K. Olorisade, “Sign-to-Speech Model for Sign Language Understanding: A Case Study of Nigerian Sign Language,” in Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 5924–5927. doi: 10.24963/ijcai.2022/855.

[5.] N. S. Chourasia and S. Barman, “Hand Gesture Spotting Using Sign Language through Computer Interfacing,” vol. 4, no. 1, 2014.

[6.] Y. M. S. S. S. Manikanta, V. R. S. M. Reddy, and R. S. Ponmagal, “Sign language training tool using machine learning techniques,” presented at the 4TH INTERNATIONAL CONFERENCE ON INTERNET OF THINGS 2023: ICIoT2023, Kattankalathur, India, 2024, p. 020125. doi: 10.1063/5.0217576.

[7.] A. M.Suresh and A.Kumaresan, “[An Integrated Two Way ISL (Indian Sign Language) Translation System – A New Approach],” ijarcs, 2013.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct983

[8.] S. Pramada, "Intelligent Sign Language Recognition Using Image Processing," IOSRJEN, vol. 03, no. 02, pp. 45–51, Feb. 2013, doi: 10.9790/3021-03224551.
[9.] S. S. Shinde, R. M. Autee, and V. K. Bhosale, "Real time two way communication approach for hearing impaired and dumb person based on image processing," in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai: IEEE, Dec. 2016, pp. 1–5. doi: 10.1109/ICCIC.2016.7919572.
[10.] K. M. Tripathi, P. Kamat, S. Patil, R. Jayaswal, S. Ahirrao, and K. Kotecha, "Gesture-to-Text Translation Using SURF for Indian Sign Language," ASI, vol. 6, no. 2, p. 35, Mar. 2023, doi: 10.3390/asi6020035.
[11.] A. Abraham and V. Rohini, "Real time conversion of sign language to speech and prediction of gestures using Artificial Neural Network," Procedia Computer Science, vol. 143, pp. 587–594, 2018, doi: 10.1016/j.procs.2018.10.435.
[12.] M. S. Amin, S. T. H. Rizvi, A. Mazzei, and L. Anselma, "Assistive Data Glove for Isolated Static Postures Recognition in American Sign Language Using Neural Network," Electronics, vol. 12, no. 8, p. 1904, Apr. 2023, doi: 10.3390/electronics12081904.
[13.] Pausa Ghazi and T. Noralam, "Data Glove for American Sign Language Alphabet and Numbers (1-9) Translation System," IJATCSE, vol. 8, no. 1.4, pp. 1128–1133, Sep. 2019, doi: 10.30534/ijatcse/2019/20842019.
[14.] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "SIGN LANGUAGE RECOGNITION BASED ON HAND AND BODY SKELETAL DATA," in 2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Helsinki: IEEE, Jun. 2018, pp. 1–4. doi: 10.1109/3DTV.2018.8478467.
[15.] E. Okorie, N. V. Blamah, and G. D. Bibu, "An Intelligent Support System for Speech and Hearing-Impaired Persons Using Deep Learning for Nigerian Sign Language," 2022, SSRN. doi: 10.2139/ssrn.5088026.
[16.] R. E. Rwelli, O. R. Shahin, and A. I. Taloba, "Gesture based Arabic Sign Language Recognition for Impaired People based on Convolution Neural Network," IJACSA, vol. 12, no. 12, 2021, doi: 10.14569/IJACSA.2021.0121273.
[17.] N. Amangeldy, A. Ukenova, G. Bekmanova, B. Razakhova, M. Milosz, and S. Kudubayeva, "Continuous Sign Language Recognition and Its Translation into Intonation-Colored Speech," 2023.
[18.] K. Navendu and V. Sahula, "Word Level Sign Language Recognition using MediaPipe and LSTM-GRU Network," 2024 IEEE International Symposium on Smart Electronic Systems (iSES), Dec. 2024, doi: 10.1109/iSES63344.2024.00014.
[19.] Ch. Rakesh, G. Madhumitha, S. Meghana, T. S. Niharika, M. Rohith, and A. Ram. K, "Two Way Indian Sign Language Translator using LSTM and NLP," IJRASET, vol. 11, no. 5, pp. 5978–5983, May 2023, doi: 10.22214/ijraset.2023.53085.
[20.] E. U. Ikwen, "Effect of Concentrated Language Encounter Method in Developing Sight Word Recognition Skills in Primary School Pupils in Cross River State," 2013.
[21.] T. O. Oyetunde, "SECOND LANGUAGE READING: INSIGHTS FROM NIGERIAN PRIMARY SCHOOLS," May 2002.