
Lightweight AI-Driven Real-Time Swahili Sign Language Recognition: A Way to Empower Hearing-Impaired Individuals**Stanley Leonard¹, Joseph Shagina²**stanley.tito@must.ac.tz¹, shaginabm@gmail.com²¹Department of Computer Science and Engineering, Mbeya University of Science and Technology, Mbeya, Tanzania²Department of Electronics and Telecommunications Engineering, Mbeya University of Science and Technology, Mbeya, Tanzania

Article Information

Received : 14 Jul 2026

Revised : 26 Aug 2026

Accepted : 30 Aug 2026

KeywordsSign Language,
Gecognition, Gesture
Modeling, Deep Learning

Abstract

Hearing impairment is a growing global health challenge that disproportionately affects people in resource-constrained countries such as Tanzania, where access to assistive technologies remains limited. Conventional solutions, including hearing aids, cochlear implants, and interpreters, are often costly and inaccessible. This study proposes an AI-driven real-time Swahili Sign Language Recognition system designed to bridge communication gaps between the hearing-impaired community and the general population. The system integrates Convolutional Neural Networks (CNNs), Scale-Invariant Feature Transform (SIFT), and Long Short-Term Memory (LSTM) with computer vision techniques to recognize and interpret sign gestures. SIFT extracts key features from images captured via webcam, while CNN performs classification and LSTM models temporal dependencies for improved accuracy. Experimental results reveals high recognition accuracy of 98.5% and a low error rate of 0.1345%, outperforming several existing models. The system offers an effective, affordable solution for enhancing communication among Swahili-speaking communities.

A. Introduction

Hearing impairment affects millions of individuals worldwide, disproportionately affecting low-income regions like Tanzania and other Sub-Saharan African countries. This led to communication constraints that significantly reduced access to social and economic services, including healthcare, education, and security. Sign language remains an essential medium for the hearing-impaired community and a vital tool for sign-to-speech and sign-to-text translation [1]. However, sign language is often misunderstood or inaccessible to individuals outside the hearing-impaired community, while sign-to-speech and sign-to-text translation remain limited and costly [2], [3]. Traditional communication methods, such as interpreters, are often expensive and require trained professionals.

Globally, over 1.5 billion people live with some degree of hearing loss, while more than 700 million people, or one in every ten people, are estimated to have a hearing disability by 2050 [4]. Nearly 50% of young people aged 12 to 35 years are at risk of hearing loss due to excessive exposure to noise[5]. In Tanzania, for example, more than 3.2 million Tanzanians (7.8%) aged 7 years and above have some form of hearing disability, and up to 5.4 million (13.2%) are affected by disability. Slightly higher prevalence of disability (13.3%) was found in Tanzania mainland, while in Zanzibar (9.3%). The occurrence is also somewhat higher in rural areas (8.3%) than in urban areas (6.36%) [6]. The prevalence of hearing loss within this broader disability context demands targeted interventions. The lack of accessible healthcare, rehabilitation services, and assistive technologies in these communities creates a significant barrier to inclusion. This underscores the urgent need for assistive technologies to address these challenges. Despite hearing loss hindering communication worldwide, Tanzania faces unique difficulties. Limited inclusive tools and technologies make it difficult for students with hearing impairments to receive a high-quality education. Lack of certified sign language interpreters in education settings makes it difficult for hard-of-hearing students to acquire quality education and other important social and economic services [6]

Several solutions have been proposed to address the challenges in hearing impairment communication. Munawar et al. [7] propose a method that identifies sign language through data fusion and attention-based machine translation from continuous RGB video using CNN and spatial-temporal LSTM. Khan et al. [9] conduct a practical Review on Indian Sign Language using Deep Neural Networks. Kanungo et al. [8] propose AI-driven gesture recognition systems for automating the translation of sign language gestures into understandable text or speech using computer vision. However, a large number of existing studies in SLR remain isolated and focus on identifying a word or phrase that is difficult to implement in a variety of devices and require the reconstruction of sentence structures from a large general dataset. There are limited studies in African Sign Languages, predominantly Swahili Sign Language recognition, which motivates this study. Existing SLR models focus on the Indian Sign Language [10], [11] and American Sign Language [12], [13] leaving African Sign language unrepresented. To address this challenge, we propose an AI-Driven Real-Time Swahili Sign Language Recognition system that uses integrated SIFT, CNN, and LSTM algorithms to process a simple Swahili sign captured via a webcam [31], [32].

B. Related Works

This section examines the evolution of sign language recognition technologies, comparing traditional rule-based methods with modern deep learning approaches while exploring key aspects such as data collection, feature extraction, and classification. AI is transforming the way people with hearing impairments communicate [2], [3]. Many approaches to sign language recognition rely either on traditional rule-based learning [12], [13] or on data-driven deep learning [15], [16].

Rule-based Sign Language Recognition

Early studies in the field of SLR have evolved from rule-based systems relying on handcrafted models to modern deep learning frameworks, which require large datasets and appropriate network architectures for feature learning and classification [17], [18], [19], [20]. Traditional SLR systems rely on rule-based or sensor-based approaches in which gesture visual cues are explicitly defined. These techniques typically use sensor gloves, such as data gloves [21], [22], or color markers [23], to measure hand orientation and simplify visual data for gesture recognition. Sequential models, such as Hidden Markov Models (HMMs), are used in sequence to support the mapping of motions and shapes to sign classes for temporal dynamics and temporal gesture handling [22], [24]. The approach requires capturing dynamic hand configurations, motion, body posture, and facial expressions for effective SLR. However, rule-based systems require manual feature engineering, which is hard, time-consuming, and not robust to environmental and signer variations. They also struggled in generalization across signers and complex signing contexts. This situation underscored the need for more flexible, data-driven techniques.

Modern Deep Learning Sign Language Recognition

This category of SLR involves data-driven approaches [1], [5], which require large annotated datasets to learn hierarchical visual representations. The approaches eliminate manual feature design from traditional rule-based systems by learning discriminative feature representations directly from raw inputs. Recent studies confirm the dominance of deep learning architectures such as Recurrent Neural Networks (RNNs) and CNNs [25], LSTM networks [24], and Attention mechanisms and Transformer [14] models for automatic feature learning and classification. These approaches combine diverse datasets from RGB video across diverse signing contexts [21], [24], or depth sensors like the Microsoft Kinect for 3D spatial cues, and multi-modal recordings to enrich feature representation. However, signer diversity and language coverage limit the performance of many systems across real-world users and environments.

Hybrid Network Architectures for Sign Language Recognition

The evolution of sign language recognition technologies reflects a broader shift from rule-based systems with explicit handcrafted features and limited adaptability to deep learning frameworks that learn rich representations and achieve high recognition and translation performance [2], [26].

Recent works in the field of SLR explore combinations of different deep learning network architectures, such as CNN-LSTM [27], Transformers with

attention modules [28], and SIFT-CNN [29] [30], [32], to improve temporal modeling and spatial feature discrimination, thereby enhancing real-time recognition. Hybrid networks, such as Transformer-based sequence-to-sequence and CNN-LSTM [15], have shown good performance in gesture recognition and translation by mapping sign sequences to speech or text. Common sign recognition and translation metrics include Word Error Rate (WER), accuracy, and BLEU scores [1]. Despite remarkable achievements in SLR and translation enabled by deep learning techniques, the scarcity of data and the real-time, low-latency deployment of SLR systems on edge devices remain major challenges that motivate future studies to enhance gesture recognition and translation in real-world settings [33], [34].

C. Research Method

Figure 1 represents a hybrid gesture recognition pipeline that integrates image processing, feature engineering, and deep learning. The pipeline is well-structured for translating hand gestures into meaningful outputs (audio), making it highly relevant for assistive communication systems and intelligent interfaces. It combines computer vision techniques with deep learning models for hand gesture recognition and translation into audio output.

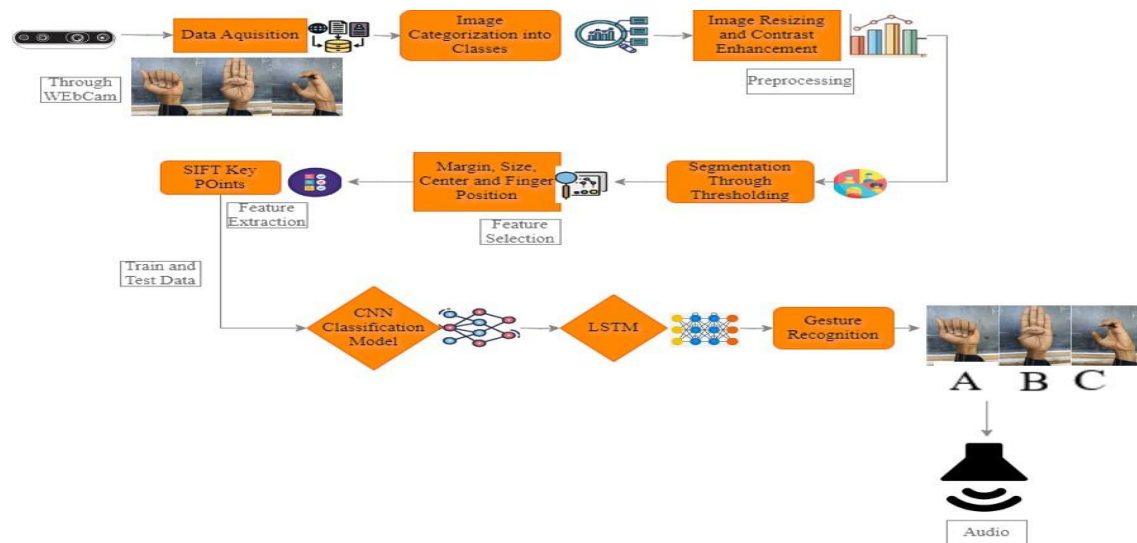


Figure 1. Workflow of the proposed system

The system relies on a collection of pre-trained samples rather than a large dataset. It performs on-demand recognition and translation of specific sign language movements using these samples and maintains accurate sign recognition even without a substantial training dataset. During operation, real-time test samples were captured via a webcam to map current gestures to existing training examples. This approach provides effective, responsive gesture detection with a lightweight TensorFlow setup.

Image acquisition

We solely focus on studying signs with a static nature. The camera first records the gestures, which are then enhanced through image preprocessing, as illustrated

in the system's workflow in Figure 1. A user can define a new gesture using the "Add Gesture" interface shown in Figure 2 by entering its title into the "New Gesture Title" box and clicking "Add Word" to start capturing images from a webcam. The large plus sign indicates each sample, with the total number of instances shown below (for example, "30 examples"). The "Train" button initiates model training once enough samples have been collected. All data for that gesture is erased when the "Clear" button is clicked. A green checkmark shown in Figure 2 indicates completion or readiness for training.



Figure 2. Collection of samples for model training

Data preprocessing and segmentation:

We distinguish the signer's hand from the background by segmenting it using threshold values for light sensitivity and color saturation. To reduce noise and variability while making feature extraction more reliable, the raw images are resized and contrast enhancement is applied to the captured gestures to improve visibility of hand features and ensure uniform input dimensions for model training. To isolate the hand Region of Interest (ROI) from the background and avoid errors in later stages, thresholding techniques were applied during segmentation.

Resizing

Images are resized to a fixed dimension as;

$$I'(x, y) = I(\alpha x, \beta y)$$

where α, β are scaling factors.

(1)

Contrast Enhancement

The image is enhanced using histogram equalization as;

$$I_{enh}(x, y) = \frac{I(x, y) - I_{min}}{I_{max} - I_{min}} \quad (2)$$

The hand is separated from the background using binary segmentation as;

$$S(x, y) = \begin{cases} 1, & \text{if } I(x, y) > T \\ 0, & \text{otherwise} \end{cases}$$

where T is an adaptive threshold value given as; (3)

$$T = \frac{1}{|\Omega|} \sum_{(x, y) \in \Omega} I(x, y) \quad (4)$$

Feature selection and extraction

SIFT extracts distinctive keypoints from the hand image that are invariant to scale, rotation, and illumination, thereby capturing important structural details of the hand gesture. To reduce dimensionality and improve computational efficiency, relevant features were selected by focusing only on expressive features based on size, margin, center, and finger position. The four key features, namely finger position, palm center, hand size, and margin or edge, were extracted from the acquired image using SIFT, as indicated in Figure 3. The edges and borders of the captured hand were detected using the Canny edge detector, which operates on intensity variations in the image. SIFT detects scale-space extrema using Difference of Gaussian (DoG) as;

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (5)$$

where:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \text{ and}$$

$$G(x, y, \sigma) = \text{Gaussian kernel}$$

Keypoints are selected as extrema in scale-space using a feature descriptor vector as;

$$\mathbf{f}_i \in \mathbb{R}^{128}$$

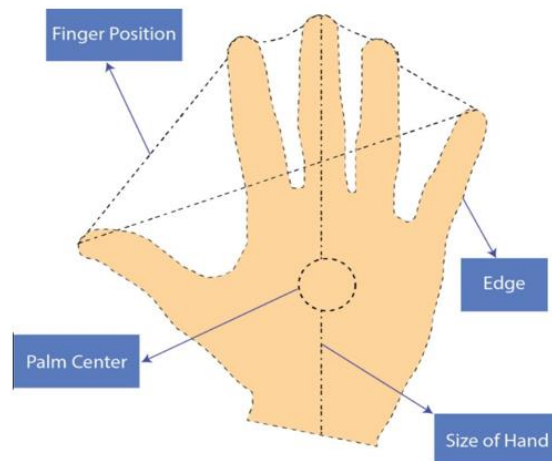


Figure 3. The hand's chosen feature set characteristics

Invariant key points from the input image $I(x, y)$ were identified using SIFT with the scale space represented as

$$H(x, y, \delta) = I(x, y) * G(x, y, \delta) \quad (6)$$

Where G represents the Gaussian kernel given as

$$G(x, y, \delta) = \frac{1}{2\pi\delta^2} + e^{\frac{-x^2+y^2}{2\delta^2}} \quad (7)$$

The hand key points are detected as local extrema in a different Gaussian space represented as $V(x, y, \delta) = S(x, y, k\delta) - H(x, y, \delta)$ (8)

Each key point is described using the gradient magnitude given as

$$g(x, y) = \sqrt{(I_x)^2 + (I_y)^2} \quad (9)$$

The orientation of each key point is controlled by

$$\theta(x, y) = \tan^{-1} \left(\frac{I_y}{I_x} \right) \quad (10)$$

These features from the SIFT descriptor are then fed into the CNN for learning, as indicated in the full architecture of the proposed system in Figure 4.

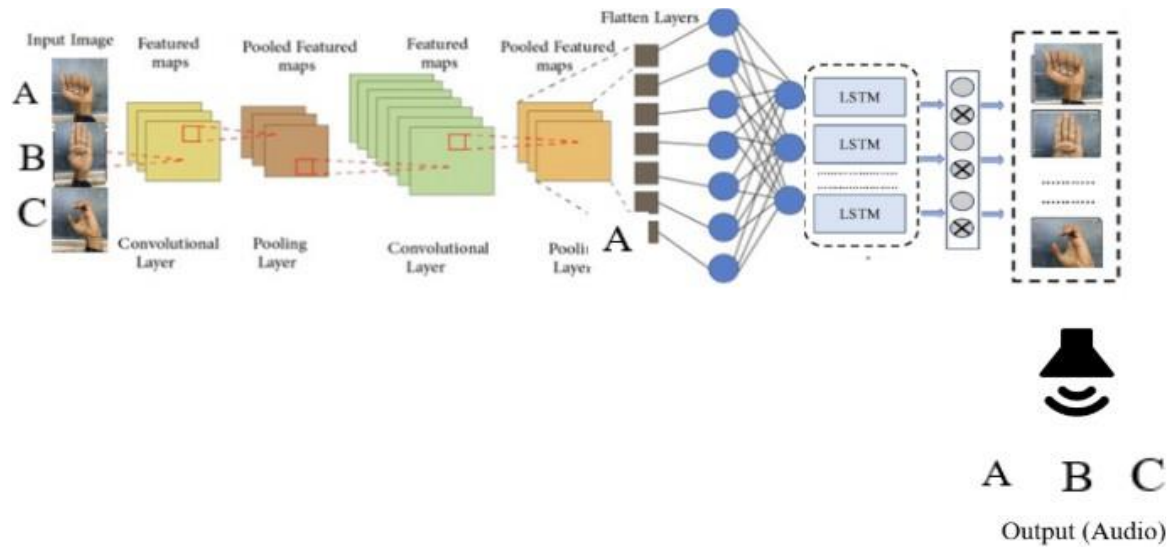


Figure 4. Architecture of the proposed Swahili Sign Language Recognition and Translation

Convolutional Neural Network (CNN)

The discriminative spatial features from the SIFT module were applied to the CNN, with the gesture input image given as $X \in \mathbb{R}^{H \times W \times C}$. The learning is controlled by

$$K_{i,j}^{(k)} = \delta \left(\sum_{m,n} X_{i+m,j+n} \cdot W_{m,n}^{(k)} + b^{(k)} \right) \quad (11)$$

Where δ - ReLu activation function and $W^{(k)}$ - Convolution kernel.

Temporal Gesture Modeling

To produce a more context-aware prediction, the LSTM learns the temporal dependencies or sequence patterns across several frames from the CNN. The gesture modeling follows the computations from the LSTM network through the time series output of the CNN module, modeled as

$$G = \{g_1, g_2, \dots, g_T\} \quad (12)$$

The LSTM then computes the gesture sequences as;

$$f_t = \delta(W_f g_t + U_f S_{t-1} + b_f) \quad (13)$$

$$\beta_t = \delta(W_\beta k_t + U_\beta S_{t-1} + b_\beta) \quad (14)$$

$$\alpha_t = \delta(W_\alpha h_t + U_\alpha S_{t-1} + b_\alpha) \quad (15)$$

$$\tilde{k}_t = \tanh(W_k h_t + U_k S_{t-1} + b_k) \quad (16)$$

$$k_t = \beta_t \cdot k_{t-1} + f_t \cdot \tilde{k}_t \quad (17)$$

$$s_t = \alpha_t \cdot \tanh(k_t) \quad (18)$$

s_t is the LSTM output, which classifies gesture sequences. The full architecture of the proposed system is shown in Figure 4.

Table 1. Gesture counts comparison

Author	Model	Gesture count
Elhagry et al., 2021	CNN-LSTM	39
Bashaa et al., 2024	SIFT +CNN	32
Pol et al., 2025	LSTM	24
Ojha et al., 2020	CNN	36
Ours	SIFT+CNN+LSTM	32

Experimental Setup

Dataset

To minimize computational and memory costs when capturing Swahili sign gestures with the webcam to depict static gestures representing Swahili idioms, the images were captured at low resolution. Pruning and quantization operations were also applied as optimization techniques to enhance real-time processing. This helps reduce computational load while maintaining high accuracy. The final model is deployed in an optimized environment to ensure low-latency sign language recognition, making the system practical for real-world applications.

Model Training

The model was trained using the Adam optimizer with a learning rate of 1×10^{-4} , a batch size of 32, and 50 training epochs. The categorical cross-entropy loss function was applied as:

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (19)$$

D. Result and Discussion

The study achieves 98.5% accuracy, with an error rate of 0.1345% and precision and recall greater than 97%. These results outperform several other state-of-the-art models with separate CNN, SIFT, or LSTM baselines, as indicated in Figure 5.

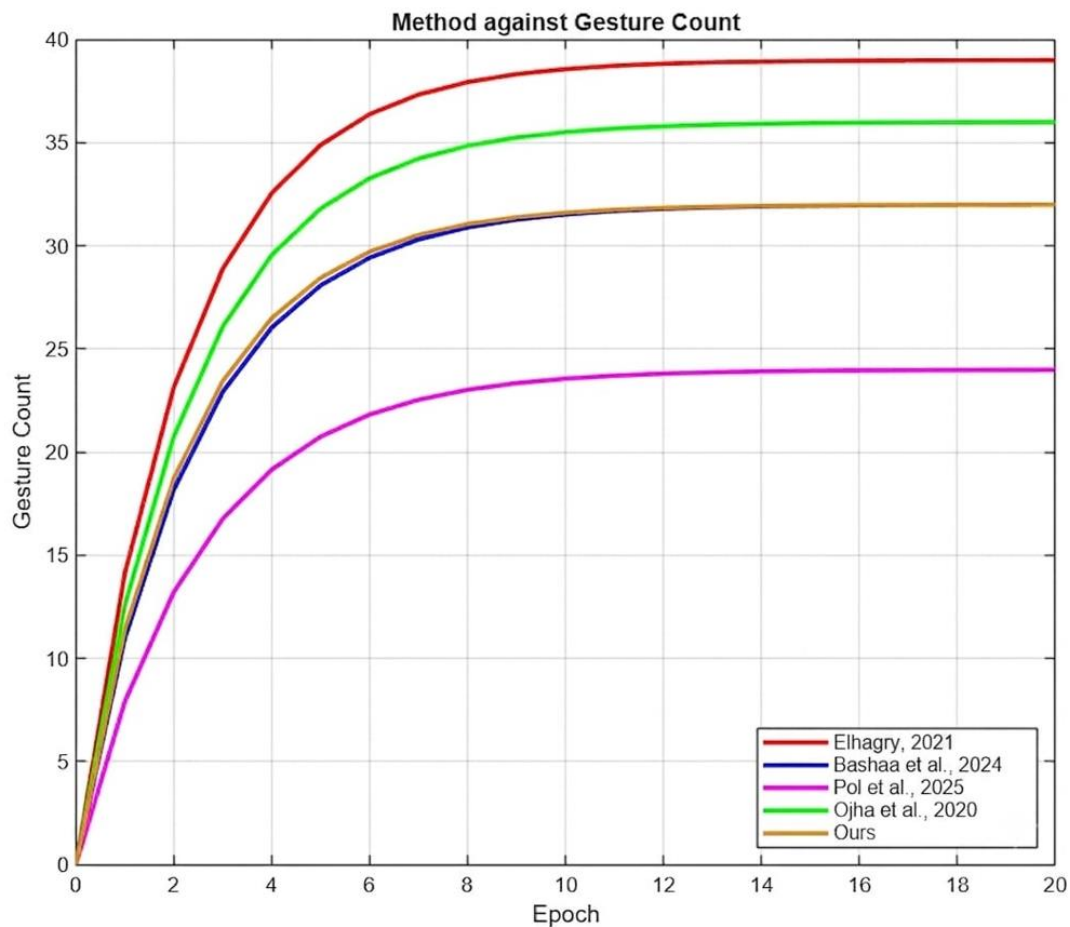


Figure 5. Gesture count comparisons with-the-state-of-the-art-methods

Figure 5 above illustrates the relationship between training epochs and gesture count, broken down by the five methods shown in Table 1. The models are evaluated for their ability to detect gestures across multiple training epochs accurately. A comparative analysis between the proposed model and existing methods, including CNN-LSTM Elhagry et al. [4], SIFT + CNN, Bashaa et al. [34], LSTM Pol et al. [2], and CNN, Ojha et al. [3], was conducted. Across all methods, there is a consistent trend in gesture count as the number of epochs increases. The results indicate that all models learn progressively over time. The performance improves rapidly during early epochs before convergence. Our proposed method achieves a slight improvement in the early epochs, with a competitive final accuracy (gesture count) of approximately 32.

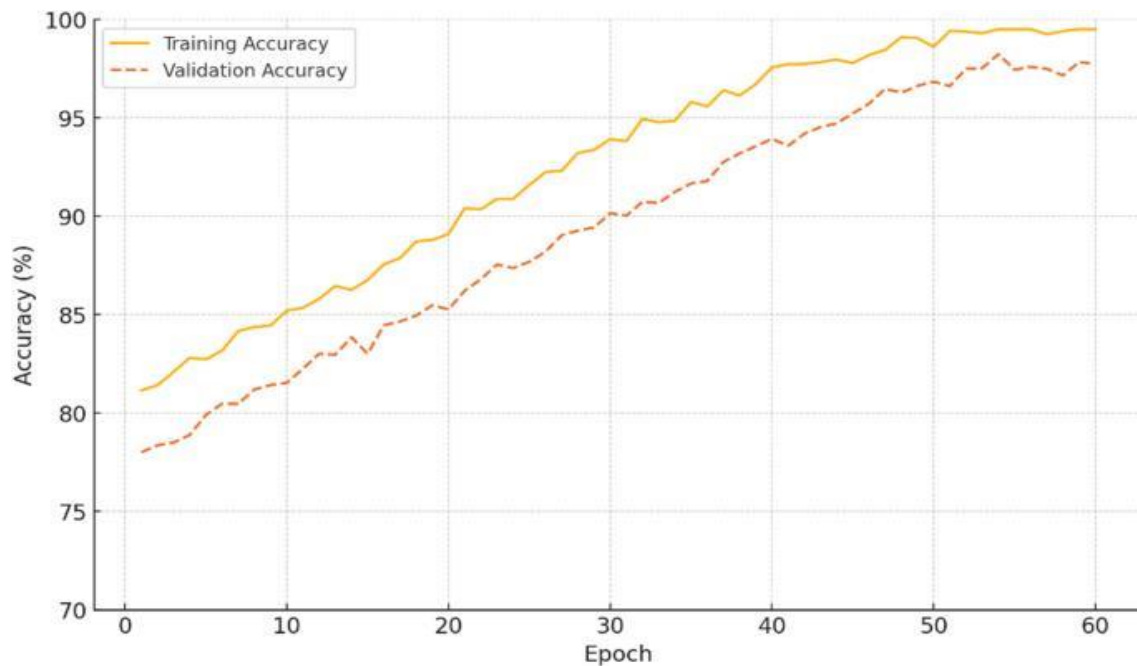


Figure 6. Model training and validation accuracy

The model achieves a competitive performance in both training and validation accuracy as the number of training epochs increases, as shown in Figure 6. This indicates that the model is effectively learning from the training data, particularly during the early stages of training.

Learning and Convergence Analysis

The results reveal three distinct learning phases across all models. The initial Learning phase at Epochs 0–5 demonstrates that all models exhibit rapid performance improvement, indicating that they quickly learn basic gesture features. At 6–10 epochs, the rate of improvement across all models begins to decline as models refine their internal representations. The models' performance stabilizes around 10–20 epochs, with minimal gains. This suggests that the models have reached their optimal learning capacity. The convergence behavior implies that extending training beyond 10 epochs yields marginal benefits. Therefore, the early stopping technique was applied in our proposed method to improve computational efficiency without compromising performance.

Table 1. Comparison with the State-of-the-art Methods on Precision, Error Rate, and Accuracy

Model	Error rate	Precision (%)	Accuracy (%)
CNN-LSTM	-	93	93
SIFT+CNN	-	-	97%
CNN	-	-	95
LSTM	-	86	86.7
SIFT+CNN+LSTM	0.1345	98.53	98.5

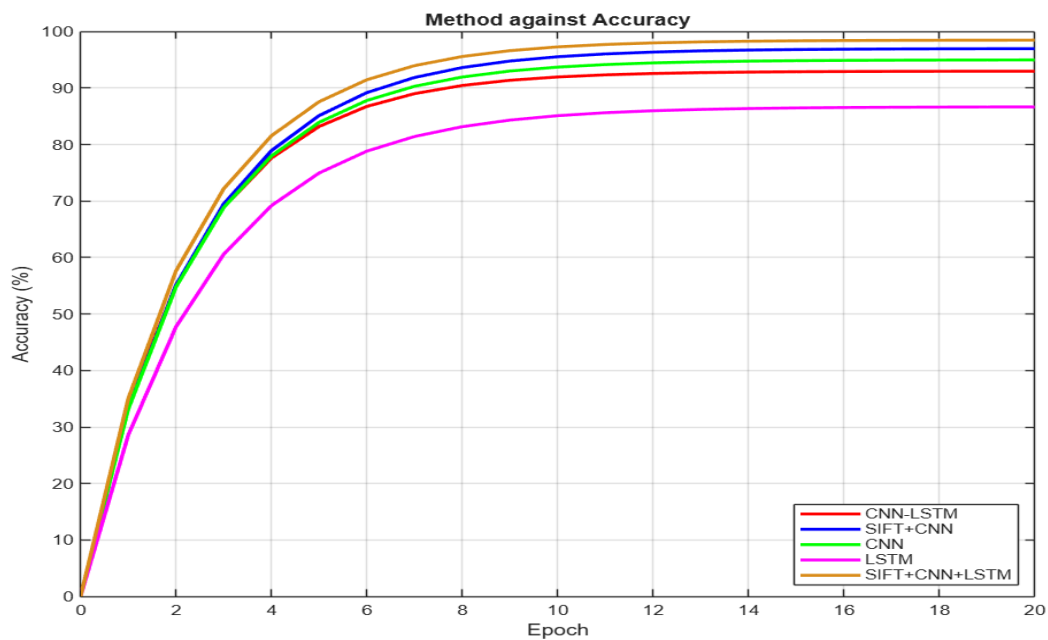


Figure 7. Comparison with the state-of-the-art methods

Figure 7 presents the relationship between training epochs and model accuracy (%) for five state-of-the-art deep learning approaches: CNN-LSTM, SIFT + CNN, CNN, LSTM, and SIFT + CNN + LSTM (Ours). The proposed method achieves approximately 99% accuracy by extracting handcrafted features with SIFT and spatial features from a CNN, while LSTM supports contextual learning. SIFT + CNN achieves a detection accuracy of 97%. The CNN-LSTM attains an accuracy of 94%.

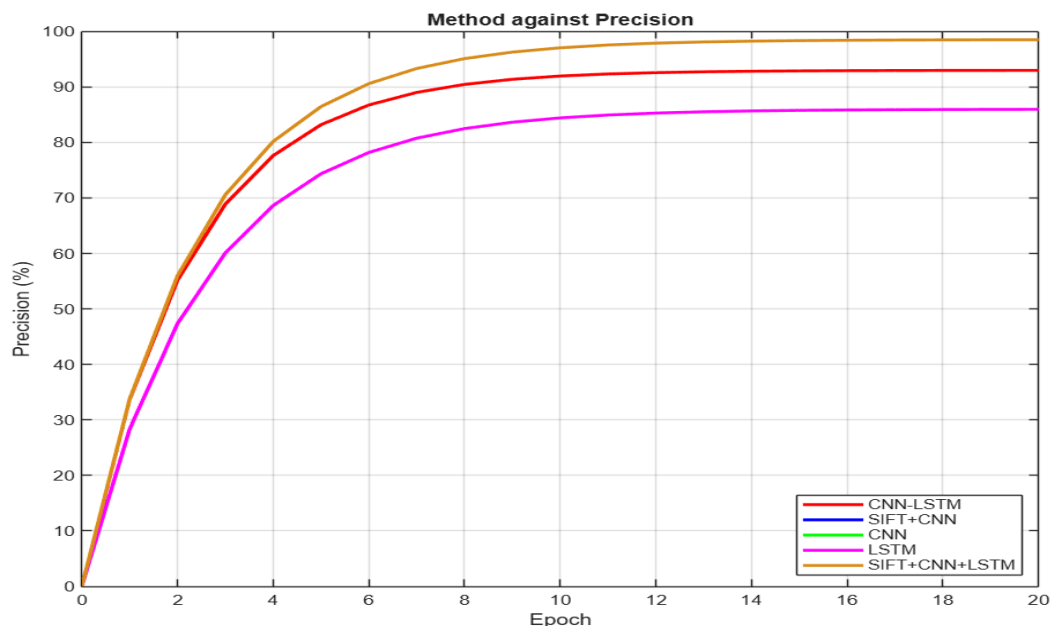


Figure 8. Comparison with the state-of-the-art methods in precision

Figure 8 compares the precision performance of four different models: CNN-LSTM, SIFT+CNN, CNN, and LSTM across training epochs. The results provide important insights into model convergence, learning efficiency, and comparative performance. All four models show a rapid increase in precision during the initial epochs (0-5), followed by a gradual plateau as training progresses. This indicates effective learning in the early stages, diminishing returns after sufficient training, and convergence toward each model's maximum achievable precision.

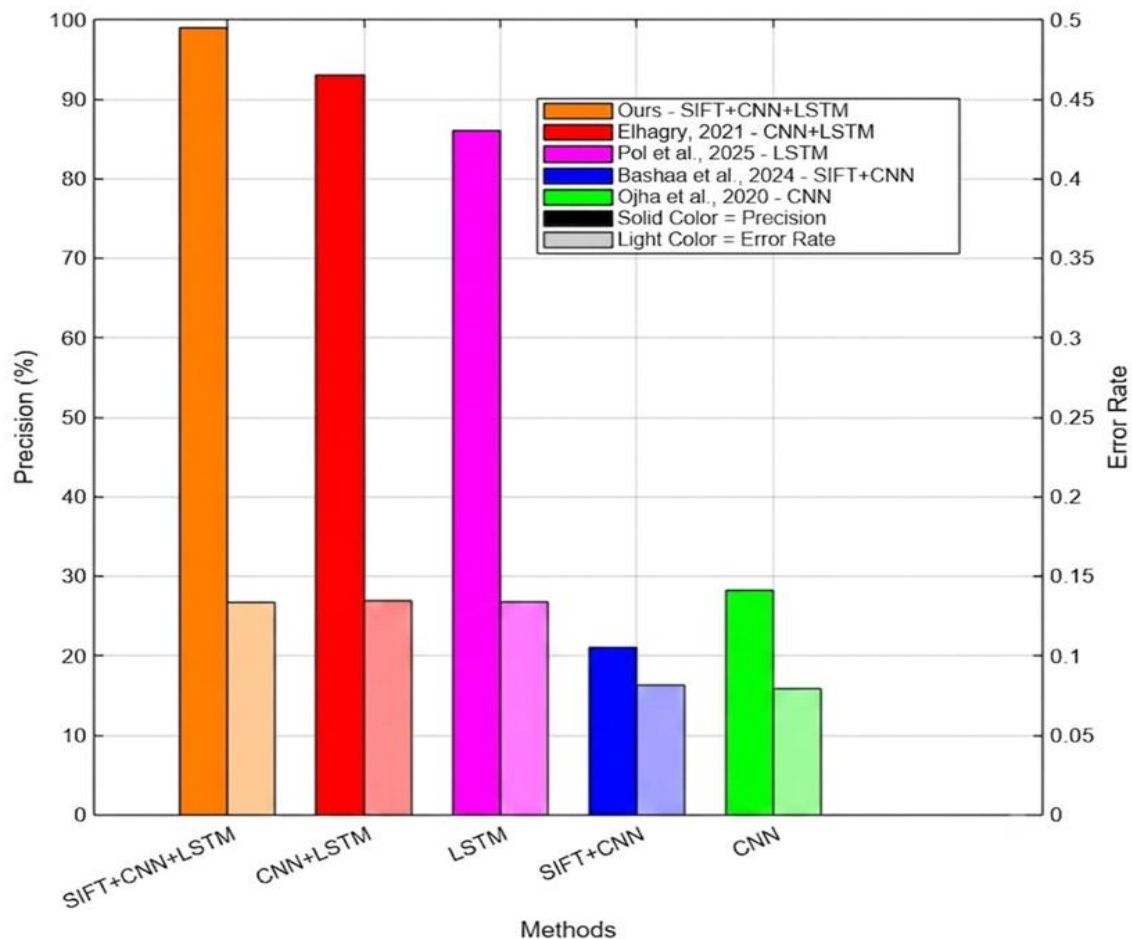


Figure 9. Comparison with the state-of-the-art-methods

Precision performance of four different models CNN-LSTM, SIFT+CNN, CNN, and LSTM across training epochs is shown in Figure 9. The results provide important insights into model convergence, learning efficiency, and comparative performance. All four models show a rapid increase in precision during the initial epochs (0-5), followed by a gradual plateau as training progresses. This indicates effective learning in the early stages, diminishing returns after sufficient training, and convergence toward each model's maximum achievable precision. The proposed SIFT+CNN+LSTM consistently outperforms all other models, achieving the highest precision (~99-100%). This makes it highly suitable for complex tasks such as gesture recognition, video analysis, and sequence-based classification problems. The model converges faster, reaching above 90% precision within a few

epochs, and maintains stability with minimal fluctuation after convergence, thanks to superior feature extraction via SIFT, spatial learning via CNN, and temporal learning via LSTM. The SIFT+CNN method attains around 92–94% precision, indicating stable convergence, but is slightly lower than the proposed method. CNN and LSTM attain moderate precisions of about 85–87% and 83–86%, respectively. However, all the models stabilize after approximately 10–12 epochs.

E. Conclusion

The study demonstrates a practical and effective solution to address communication barriers faced by the hearing-impaired community in resource-constrained settings like Tanzania. By integrating CNN, SIFT, and LSTM techniques, the proposed system successfully captures spatial features and temporal dynamics of sign gestures, resulting in high recognition accuracy and improved performance compared to existing approaches. The achieved accuracy of 98.5% and low error rate highlight the robustness and reliability of the model in real-time applications. The system provides a cost-effective and accessible alternative to traditional assistive technologies, which are often unaffordable or unavailable. Therefore, this work contributes significantly to inclusive communication by leveraging artificial intelligence to empower Swahili-speaking communities, and it lays a foundation for future advancements in sign language recognition systems tailored to low-resource environments.

F. References

- [1] A. Khan et al., "Deep Learning Approaches for Continuous Sign Language Recognition : A Comprehensive Review," *IEEE Access*, vol. 13, no. April, pp. 55524–55544, 2025, doi: 10.1109/ACCESS.2025.3554046.
- [2] A. M. Pol and S. A. Patil, "RNN-driven integration of spatial, temporal, features for Indian sign language recognition and video captioning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 38, no. 2, p. 821, 2025, doi: 10.11591/ijeecs.v38.i2.pp821-829.
- [3] A. Ojha, A. Pandey, S. Maurya, A. Thakur, and P. Dayananda, "Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network," *Int. J. Eng. Res. Technol.*, vol. 8, no. 15, pp. 191–196, 2020, [Online]. Available: www.ijert.org
- [4] A. Elhagry and R. G. Elrayes, "Egyptian Sign Language Recognition Using CNN and LSTM," 2021, [Online]. Available: <http://arxiv.org/abs/2107.13647>
- [5] T. Tao, Y. Zhao, T. Liu, and J. Zhu, "Sign Language Recognition : A Comprehensive Review of Traditional and Deep Learning Approaches , Datasets , and Challenges," *IEEE Access*, vol. 12, no. May, pp. 75034–75060, 2024, doi: 10.1109/ACCESS.2024.3398806.
- [6] M. Hersh, B. Leporini, and M. Buzzi, "A comparative study of disabled people's experiences with the video conferencing tools Zoom , MS Teams , Google Meet and Skype," *Behav. Inf. Technol.*, vol. 43, no. 15, pp. 3777–3796, 2024, doi: 10.1080/0144929X.2023.2286533.
- [7] S. Africa, "A proposed artificial intelligence-based real-time speech- to-text to sign language translator for South African official languages for the COVID-19 era and beyond : In pursuit of solutions for the hearing impaired,"

- 2022.
- [8] I. M. M. Violet and R. L. Sri, "Discover Artificial Intelligence Article in Press A comprehensive survey on recent advances and challenges in sign language recognition systems IN AR IN," pp. 0–44, 2025.
- [9] E. Mallya and R. Haynes, "Hearing Impairment; Barrier to Social Participation for Secondary School Students in Meru District-Arusha," vol. 43, no. 2, pp. 22–29, 2023, doi: 10.9734/AJESS/2023/v43i2937.
- [10] M. Munawar, I. Noreen, R. S. Alharthi, and N. Sarwar, "Forged Video Detection Using Deep Learning : A SLR," vol. 2023, 2023, doi: 10.1155/2023/6661192.
- [11] T. Kanungo, S. Aswini, S. Neha, and V. Ramasamy, Hand Speak: An AI-Powered Real-Time System for Sign Language Recognition and Seamless Translation, no. Iccsce. Atlantis Press International BV, 2025. doi: 10.2991/978-94-6463-858-5.
- [12] M. R. K, "Image-based Indian Sign Language Recognition : A Practical Review using Deep Neural Networks," pp. 1–14.
- [13] M. Vilasini and P. Ramamoorthy, "CNN approaches for classification of Indian leaf species using smartphones," Comput. Mater. Contin., vol. 62, no. 3, pp. 1445–1472, 2020, doi: 10.32604/cmc.2020.08857.
- [14] A. A. Q. Mohammed, J. Lv, and M. D. S. Islam, "A deep learning-based end-to-end composite system for hand detection and gesture recognition," Sensors (Switzerland), vol. 19, no. 23, 2019, doi: 10.3390/s19235282.
- [15] A. D. V, T. Charulatha, and P. Dharishinie, "A Survey on Sign Language Recognition and Training Module," vol. 01019, 2023.
- [16] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-modal deep hand sign language recognition in still images using Restricted Boltzmann Machine," Entropy, vol. 20, no. 11, pp. 1–15, 2018, doi: 10.3390/e20110809.
- [17] N. Shahin and L. Ismail, From rule - based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey , taxonomy and performance evaluation Deaf or Hard of Hearing, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10895-z.
- [18] S. L. Tito, J. F. Banzi, and A. N. Mvuma, "A Deep Regression Network with Key-joints Localization for Accurate Hand Pose Estimation," Int. J. Mach. Learn. Comput., vol. 12, no. 6, 2022, doi: 10.18178/ijmlc.2022.12.6.1118.
- [19] T. W. Chong and B. G. Lee, "American sign language recognition using leap motion controller with machine learning approach," Sensors (Switzerland), vol. 18, no. 10, 2018, doi: 10.3390/s18103554.
- [20] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.
- [22] E. Dibra, "Recovery of the 3D Virtual Human: Monocular Estimation of 3D Shape and Pose with Data Driven Priors," Dr. Thesis, 2018, [Online]. Available: <https://doi.org/10.3929/ethz-b-000266852>
- [23] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

- [24] A. Leandro, C. Carneiro, D. Henrique, P. Salvadeo, and D. B. Silva, "Sign language recognition based on deep learning and low-cost handcrafted descriptors," pp. 1–28.
- [25] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12536 LNCS, pp. 249–263, 2020, doi: 10.1007/978-3-030-66096-3_18.
- [26] M. H. Chu, H. D. Nguyen, T. Ngoc, A. Nguyen, and H. N. Vu, "Cross-attention multi branch for Vietnamese sign language recognition : CrossViViT," 2025.
- [27] S. G. Moreira Almeida, F. G. Guimarães, and J. Arturo Ramírez, "Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7259–7271, 2014, doi: 10.1016/j.eswa.2014.05.024.
- [28] N. Adaloglou et al., "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," *IEEE Trans. Multimed.*, vol. 24, pp. 1750–1762, 2022, doi: 10.1109/TMM.2021.3070438.
- [29] V. Adithya and R. Rajesh, "A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 2353–2361, 2020, doi: 10.1016/j.procs.2020.04.255.
- [30] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention," *30th Br. Mach. Vis. Conf. 2019, BMVC 2019*, pp. 1–13, 2020.
- [31] B. A. A. L. Abdullah, G. A. Amoudi, and H. S. Alghamdi, "Advancements in Sign Language Recognition : A Comprehensive Review and Future Prospects," vol. 12, no. September, 2024.
- [32] B. Ghogh, A. L. I. Ghodsi, and U. Ca, "Attention Mechanism , Transformers , BERT , and GPT : Tutorial and Survey," *ArXiv*, 2017.
- [33] A. Tyagi and S. Bansal, "Hybrid FAST-SIFT-CNN (HFSC) Approach for Vision-Based Indian Sign Language Recognition," vol. 1, no. 1, 2022.
- [34] M. H. Bashaa, F. H. Almasoudy, N. S. Sagheer, and W. M. Hadi, "CNN Algorithm with SIFT to Enhance the Arabic Sign Language Recognition," *Int. J. Emerg. Sci. Eng.*, vol. 12, no. 10, pp. 12–17, 2024, doi: 10.35940/ijese.i2584.12100924.