Computing&AI Connect
ISSN: 3104-4719
2026, Article ID. 2026.0034, Cite as: https://doi.org/10.69709/CAIC.2026.130891

SCIFINITI PUBLISHING

Research Article

OPEN ACCESS

Attention-Based LSTM for Sign Language Recognition Leveraging Spatiotemporal Keypoints

Linus Tabari✉* **Kate Takyi**✉ **Rose-Mary Owusuaa Mensah Gyening**✉

Department of Computer Science, Faculty of Computational and Physical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi P.O. Box no. 643, Ghana

Article History
Received: November 01, 2025 Accepted: April 13, 2026 Published: May 11, 2026

Abstract
Sign language (SL) is a vital mode of communication for Deaf and hard-of-hearing communities. However, most individuals experience difficulty communicating with Deaf individuals in the absence of an interpreter. Advancements in technology, particularly in computer vision and deep learning, offer alternative approaches to addressing this problem. The literature indicates that Ghanaian Sign Language (GSL) remains under-researched due to the limited availability of large, publicly accessible datasets, as well as insufficient studies on the use of landmark keypoints for computational analysis of GSL. This study introduces the AkwaabaSign dataset, a large video-based collection that reflects the indigenous nature of Ghanaian Sign Language (GSL). The study employed two baseline models to evaluate the dataset: an attention-enhanced LSTM model and a ConvLSTM model, both of which extracted and normalized keypoints using MediaPipe. The attention-enhanced LSTM model achieved a test accuracy of 94.69%, along with balanced performance metrics: 93.32% precision, 92.70% recall, and 92.66% F1-score. In comparison, the ConvLSTM model achieved an accuracy of 90.28%, indicating lower performance than the attention-enhanced LSTM. The study achieved the objective of developing a large-scale dataset for sign language recognition, introduces a specialized normalization pipeline for dataset processing, and establishes a baseline model to support the dataset's practical application. The proposed model also outperforms several existing algorithms in computational research on Ghanaian Sign Language (GSL). This study further aims to expand the dataset to sentence-level data and develop a system for continuous GSL recognition.

Keywords:
convolutional neural network; deep neural network; long short-term memory; spatiotemporal; machine learning; attention-based architectures; support vector machine

1. Introduction

Communication is a fundamental mechanism for information exchange, encompassing both verbal and nonverbal modalities. Sign language (SL) is a natural visual language used by Deaf and Hard-of-Hearing (DHH) individuals for everyday communication. According to the World Health Organization (WHO), approximately 5% of the global population, or around 430 million people, experience hearing loss and require rehabilitation services [1]. There are several distinct sign languages, including American Sign Language (ASL) and German Sign Language (DGS), each influenced by regional variations. Therefore, sign language is not universal. Although it differs in form, it serves the same communicative functions as spoken languages [2]. In the contemporary world, most individuals do not understand sign language, which makes communication with Deaf and Hard-of-Hearing (DHH) individuals particularly challenging, especially in critical sectors such as healthcare and education. Sign language serves as a bridge to overcoming this communication barrier. Historically, linguists often misunderstood sign languages as merely gestural accompaniments to speech. However, William Stokoe's pioneering analysis of American Sign Language (ASL) demonstrated that sign languages pos-

* **Corresponding Author:**
Linus Tabari, Department of Computer Science, Faculty of Computational and Physical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi P.O. Box no. 643, Ghana; ltabari1@st.knust.edu.gh

sess their own phonological and grammatical structures, establishing them as fully developed human languages [3].

Researchers have since examined different forms of sign languages and the information conveyed through visual communication using both manual and non-manual modalities. Manual parameters include hand shape, hand orientation, hand location, and hand movement, while non-manual parameters involve head and body posture, facial expressions, gaze direction, and lip movements. Ghanaian Sign Language (GSL) serves as the primary communication medium for the hearing-impaired community in Ghana. Nevertheless, the lack of comprehensive datasets that reflect the indigenous structure of GSL has hindered large-scale computational research, highlighting the need for dedicated dataset curation and model development efforts. Most available datasets originate from high-income countries and focus on sign languages such as ASL and DGS, which do not capture the indigenous characteristics of GSL. The scarcity of available data has compelled researchers to develop custom datasets for model training [3].

Globally, research in sign language recognition has evolved significantly, with early systems relying heavily on handcrafted features and sensor-based gloves. These systems commonly employed Hidden Markov Models (HMMs) to capture temporal dynamics [4]; however, their scope and practical applicability remained limited [5]. Comprehensive surveys of recent developments in sign language recognition highlight a clear shift toward deep learning approaches, particularly convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and hybrid architectures. These methods have consistently achieved high accuracy on isolated datasets, underscoring the growing dominance of deep learning in advancing the field [6–8].

Despite these global advances, many sign languages, including Ghanaian Sign Language (GSL), which is linguistically distinct with its own phonological and morphological structures, remain under-researched. Studies have indicated a significant lack of publicly available datasets for Ghanaian Sign Language (GSL), as noted in [3]. In response to this limitation, a previous study developed a self-constructed dataset for model training, which achieved an accuracy of approximately 96% using convolutional neural networks (CNNs) and transfer learning approaches. However, the dataset was relatively small, non-generalizable, and not available to the broader research community. This scarcity of data presents a dual challenge. First, it impedes the development of robust Ghanaian Sign Language (GSL) recognition systems. Second, it delays research into contextually appropriate sign language technologies tailored to the needs of Ghana's Deaf community. Addressing this gap requires the development of comprehensive datasets and the implementation of technological solutions specifically tailored to Ghanaian Sign Language (GSL).

Recent advancements in computer vision and sequential modelling have opened new possibilities. Frameworks such as MediaPipe and other modern pose estimation tools can reliably and efficiently extract human keypoints, enabling lightweight feature representations from video frames. Models integrating Long Short-Term Memory (LSTM) units with attention mechanisms, as well as transformer-based architectures, have demonstrated strong performance in sign and gesture recognition by effectively capturing both temporal and spatial dependencies in sequential data [8–10].

With the growing interest in African sign languages, Ghanaian Sign Language (GSL) remains significantly underrepresented in computational linguistics and machine learning research. Most existing sign language recognition systems are developed using American or British Sign Language datasets, while available indigenous datasets are often limited in size or require labor-intensive collection methods, thereby failing to meet the robustness and scalability demands of modern sign language recognition applications [3,11].

Our key contributions are as follows:

- Introduction of a novel sign language recognition dataset, AkwaabaSign, a custom video-based dataset comprising isolated GSL words signed by multiple individuals
- Designed a pipeline that extracts and utilizes 2D keypoint coordinates (face, hands, and body pose) for gesture recognition
- Proposed a robust normalization strategy that enhances recognition accuracy by aligning keypoints with the canonical representation
- Designed a deep learning model architecture that utilizes spatial and temporal features to better recognize isolated sign words from videos

The rest of the paper is organized as follows: Section 2 provides a theoretical, conceptual, and empirical review of related work. Section 3 describes the data acquisition process, data preprocessing steps, normalization approach, and model architecture. Section 4 presents a summary of the results, including an analysis of the baseline models, a comparison with existing studies, and a description of the evaluation metrics used. It also presents a detailed discussion of the findings. Section 5 concludes the paper and outlines recommendations for future research directions.

2026, Vol. 3, Article ID. 2026.0034
https://doi.org/10.69709/CAIC.2026.130891

2. Literature Review

This section reviews studies related to sign language recognition. The review begins with theoretical insights into the linguistic and visual attributes of sign languages, and then examines the conceptual building blocks of SLR systems. It provides an empirical synthesis of key studies across various sign languages, highlighting advances in machine learning, deep learning, and computer vision. It also emphasizes the marginalization of Ghanaian Sign Language (GSL) within these domains, particularly the challenges of limited datasets, signer variability, and poor model generalization. These issues are directly addressed in this study through the development of the custom AkwaabaSign dataset and the use of efficient modelling techniques.

Historically, sign languages were often misinterpreted as mere pantomime or simplified gestural systems. This perception was subsequently challenged by William Stokoe's pioneering work, which established American Sign Language (ASL) as a legitimate linguistic system. Like spoken languages, sign languages are fully developed natural languages with phonological, morphological, syntactic, and semantic structures expressed through the visual-manual modality [12]. While spoken language relies on auditory signals to convey information, sign language uses manual parameters such as hand shape, hand location, hand orientation, and hand movement, along with non-manual markers including facial expressions, gaze direction, and body posture. These components form the linguistic system of sign languages worldwide; however, research coverage remains uneven across different sign languages [11]. Although Ghanaian Sign Language (GSL) holds significant sociolinguistic importance for the Deaf and Hard-of-Hearing (DHH) community in Ghana, it remains underrepresented both as an independent linguistic system and in computational research. Most existing studies rely on datasets from American Sign Language (ASL), British Sign Language (BSL), or German Sign Language (DGS), which are comparatively more accessible. In contrast, the absence of large-scale GSL datasets and established deep learning benchmarks continues to limit its computational exploration and model development [3]. For indigenous African sign languages, this poses a challenge in developing effective sign language recognition systems, thereby curtailing opportunities for comparative linguistic study with other sign languages.

Earlier sign language recognition (SLR) systems employed data gloves and motion sensors to capture hand and finger positions, facilitating gesture identification [13]. Coupled with statistical and mathematical models, along with advancements in machine learning and computer vision, research in sign language recognition has evolved in multiple directions. Feature extraction techniques such as Histogram of Oriented Gradients (HOG) [14] and Scale-Invariant Feature Transform (SIFT) [15] have been widely used in combination with classifiers such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs), achieving accuracies of up to 78.85% on RGB images [16]. Although this approach yielded promising results, its performance is limited when exposed to variability in signers, lighting conditions, and background environments.

Advancements in deep learning have introduced modern architectures such as Convolutional Neural Networks (CNNs), which are widely used for spatial feature extraction. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have also been employed to capture temporal dynamics and model sequential data, leading to more robust and reliable performance. Shin et al. applied a CNN model to the KSL dataset and achieved improved performance compared with existing systems [17]. Pigou et al. demonstrated an end-to-end learning approach using raw video inputs without relying on handcrafted features, achieving a high recognition rate [18]. This approach enables the development of lightweight, language-agnostic SLR pipelines by leveraging 2D or 3D keypoints of body, hand, and facial landmarks. In addition, pose estimation frameworks such as OpenPose [19] and MediaPipe [20] can be employed to generate efficient models, reducing the impact of variations in background and lighting conditions [21,22].

In recent years, self-attention-based architectures [23], particularly transformers, have demonstrated strong performance in sequence modelling tasks, often matching or surpassing recurrent neural networks (RNNs) in capturing long-range dependencies within sequential data [24]. These advancements have also been applied to sign language recognition (SLR), where transformer-based models have achieved state-of-the-art performance in recognition tasks [10,25]. However, their high computational complexity has limited their deployment in resource-constrained environments.

The intersection of linguistics, computer vision, and machine learning has been regarded as the conceptual foundation of sign language recognition [26]. The process of developing standard SLR systems follows a pipeline of data acquisition, preprocessing, feature extraction, sequence modeling, and classification. The choice of methodology is shaped by decisions made at each stage, which in turn influence recognition performance, computational efficiency, and adaptability [9]. Data collection is typically carried out using RGB cameras to record signs, as they are more cost-effective and provide sufficiently accurate results compared with depth or infrared sensors [5,10]. Preprocessing involves transforming acquired data through

several operations, including frame resizing, background subtraction, segmentation of signing regions, and normalization of spatial coordinates, thereby converting the data into a structured form that facilitates efficient processing and analysis [27]. In pose-based approaches, preprocessing is closely integrated with the outputs of landmark detection algorithms such as MediaPipe Holistic, which generate high-dimensional keypoint coordinates for the face, hands, and upper body. Early sign language recognition (SLR) research relied on handcrafted descriptors such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Motion History Images for feature extraction. However, modern approaches increasingly employ deep neural networks, where convolutional layers automatically learn hierarchical spatial features [28,29]. Recently, skeletal keypoint-based representations have emerged as a lightweight alternative, particularly advantageous for datasets with limited samples, as they preserve essential motion and posture information while reducing computational complexity [30].

Sign language is inherently temporal and therefore requires sequential modelling, which plays a central role in recognition. Hidden Markov Models (HMMs) have been used for real-time recognition by capturing state transitions; however, they are limited in their ability to model long-range dependencies. RNNs, such as LSTM, address issues related to HMMs, including the vanishing gradient problem [31]. The incorporation of attention mechanisms and transformer architectures enables models to dynamically assign weights to temporal segments and selectively focus on the most informative portions of a sequence [24].

Compared with traditional methods, deep learning has achieved remarkable success in computer vision tasks. It has been widely applied to sign language and gesture recognition using architectures such as CNNs, LSTMs, and RNNs. Chen et al. proposed a 3D-CNN-based approach for human action recognition, which was later adapted for sign language recognition (SLR), enabling direct learning of both spatial and temporal features from raw video frames [29]. In practice, hybrid architectures are more commonly used, such as CNN–LSTM models for Arabic sign language recognition, as well as approaches incorporating temporal convolution and bidirectional RNNs [22,32].

Instead of processing large and computationally intensive raw images, pose estimation based on keypoints has emerged as an effective alternative. This approach utilizes 2D or 3D skeletal landmarks of the face, hands, and body, significantly reducing input dimensionality while preserving essential motion cues, which are critical for sign language recognition (SLR) [9]. Compared with CNN-based, RNN-based, and pose-based SLR methods, these approaches are more computationally efficient and more robust to background variations, making them particularly suitable for low-resource settings. Kumar et al. conducted a comparative analysis of various sign language recognition (SLR) techniques, including traditional vision-based methods, deep learning architectures, and, where applicable, hybrid approaches [5]. The review demonstrated that keypoint-based data representations are advantageous in scenarios with limited computational resources, as they help mitigate challenges such as signer dependency and background variability.

There have been significant global advancements in SLR studies, but research on SLR in Africa remains relatively underexplored. The review of this study indicates that pose-based approaches, when combined with attention-enhanced recurrent architectures, often provide an effective solution for isolated sign recognition, particularly in resource-constrained environments. These approaches help address common challenges in Ghanaian Sign Language (GSL), including signer variability, background clutter, and limited datasets. By utilizing MediaPipe for pose estimation and an attention-based LSTM for temporal modeling, this study leverages the strengths of effective empirical methods while addressing key gaps in GSL research.

3. Methodology and Experiments

This work proposes a pose-based approach for Ghanaian Sign Language (GSL) recognition. The step-by-step framework for developing the pose-based GSL recognition system is outlined below.

3.1. Research Design

Human keypoints extracted from video frames are used instead of raw pixel data to reduce dimensionality while preserving motion and structural information essential for sign language recognition [9]. This research proposes a methodology to improve generalization across signers by reducing signer dependence and minimizing overfitting to variations in background and clothing. It also ensures computational efficiency for deployment on low-resource hardware while integrating advanced sequential modelling with attention mechanisms for spatiotemporal tasks [24,33]. Table 1 compares input approaches for sign language recognition (SLR), highlighting the advantages of pose-based methods. Figure 1 illustrates the proposed pipeline, comprising data curation, keypoint extraction using MediaPipe Holistic, normalization, model training (including an attention-enhanced LSTM baseline and a ConvLSTM comparative model), and evaluation.

2026, Vol. 3, Article ID. 2026.0034
https://doi.org/10.69709/CAIC.2026.130891

Table 1: Comparison of input approaches for SLR.

Approach	Input Type	Data Source	Generalization Across Signers	Computational Efficiency	Temporal Modeling Capability
Pose-Based (Keypoints)	Keypoints data	Extracted using pose estimation models	High; pose sensitivity helps reduce background noise	High; produces lightweight models for recognition	Strong (attention-enhanced)
Raw Video	Sequential video frames	Direct video feed from cameras	Low, as it is susceptible to environmental factors (sensitive to background)	Low (processing extensive video data requires high power resources for models like CNNs or RNNs)	Moderate (Sensitive to environmental factors like lighting)
Image-Based	Static Images	Single frame capture	Low (Mostly sensitive to signer gestures)	Low to moderate, as it uses single-frame processing	Low (Lack of temporal context)
Sensor-Based	Data from Sensors (gloves, EMG, etc.)	Wearable device	Moderate to high (depends on the device and its data quality)	Moderate to high (depending on sensor data complexity)	Strong (motion capture)

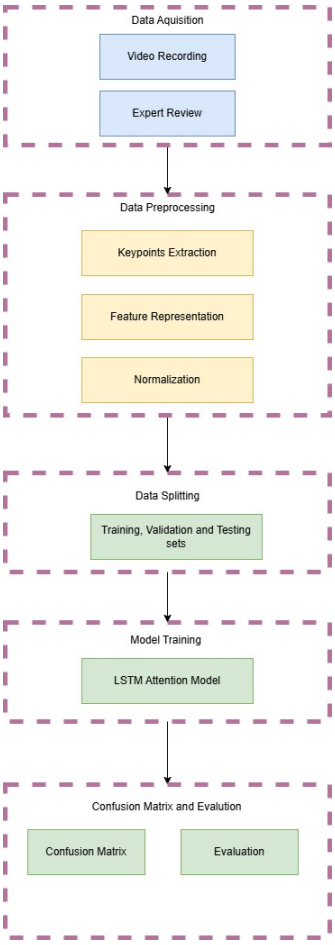

Figure 1: Flowchart of the proposed methodology pipeline for GSL recognition.

3.2. Data Description

The dataset used in this study is a novel, custom-compiled, publicly available Ghanaian Sign Language (GSL) dataset, developed to address the scarcity of resources for sign language research in Ghana. It comprises 5750 videos covering 115 words, with each word signed by five indigenous signers (with at least 10 videos per word–signer pair). Each video is approximately 5 seconds in duration, recorded at 30 FPS (150 frames per video), and stored in .avi format. The videos were recorded in a controlled environment under standardized lighting conditions using a Sony camera and an Ikan teleprompter setup. The words were selected in consultation with Deaf educators and linguists, with a focus on educational (71 words) and health-related (44 words) categories. Figure 2 illustrates the category-wise distribution per signer, while Figure 3 presents the hierarchical directory structure, consisting of word-level classes organized into subfolders per signer, with each signer directory containing 10 videos. A graphical representation of the total number of videos signed by each signer is shown in Figure 4.

3.3. Data Preprocessing and Feature Engineering

In this study, the raw video dataset is preprocessed into a suitable format for model training through scaling, augmentation, and normalization. This ensures that the data reflects real-world variations and improves the model's ability to generalize.

3.3.1. Keypoint Extraction

Using the MediaPipe Holistic framework developed by Google Research as the feature extraction tool, the videos are processed frame by frame, generating 543 keypoints that capture facial landmarks, hand positions, and full-body pose information [20]. For the proposed study, a subset of landmarks was selected to capture the most relevant features, enabling robust sign language detection. The extracted landmarks include 33 pose landmarks, 21 landmarks per hand, and a subset of the 468 facial landmarks. The use of keypoints offers two main advantages over traditional approaches: it reduces input dimensionality compared to raw video frames and significantly lowers computational cost.

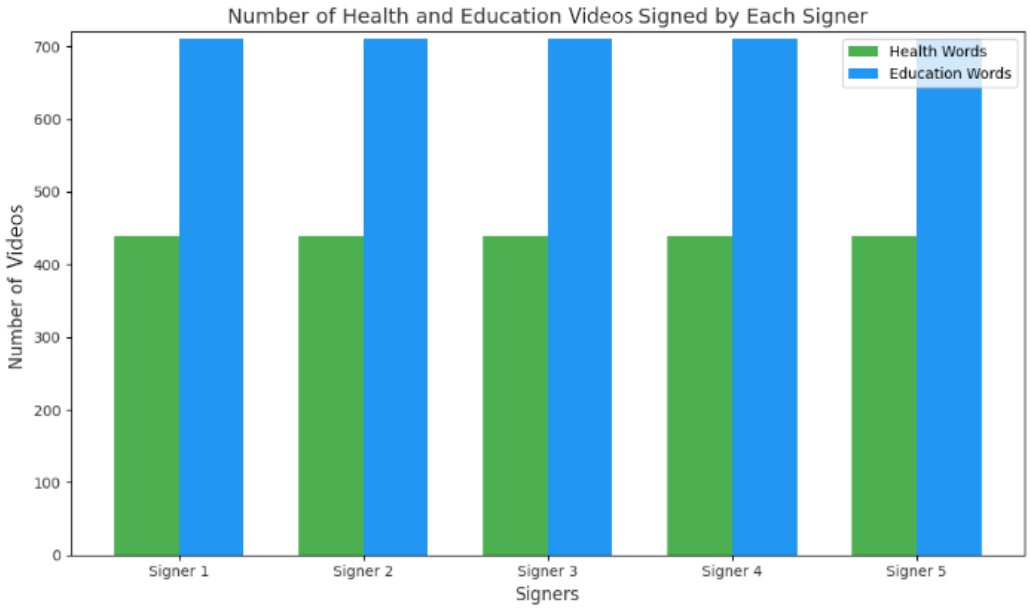

Figure 2: Distribution of videos signed by each signer per category.

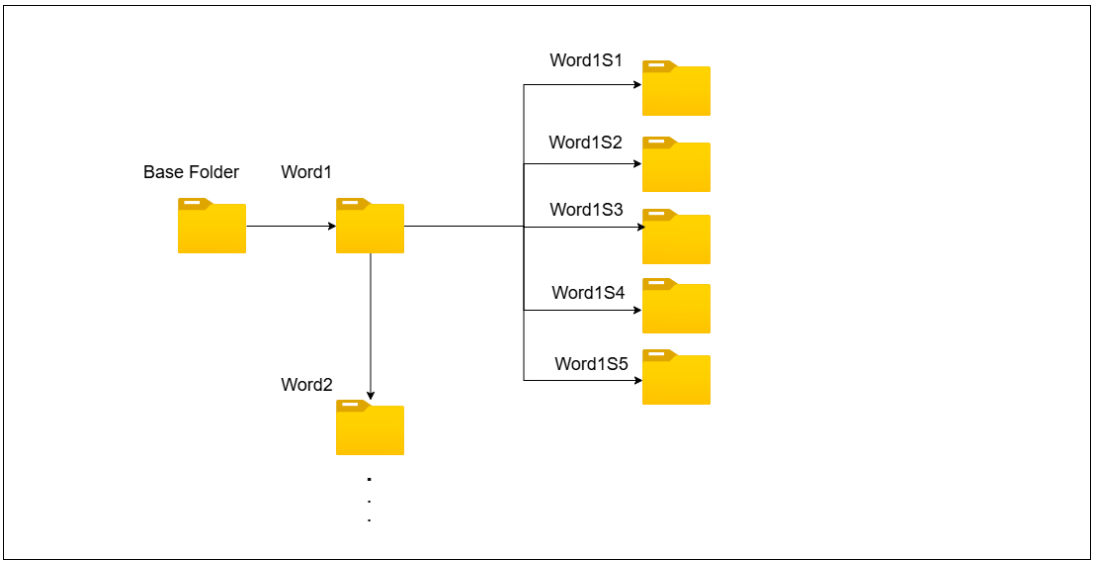

Figure 3: Hierarchical representation of dataset structure.

2026, Vol. 3, Article ID. 2026.0034
https://doi.org/10.69709/CAIC.2026.130891

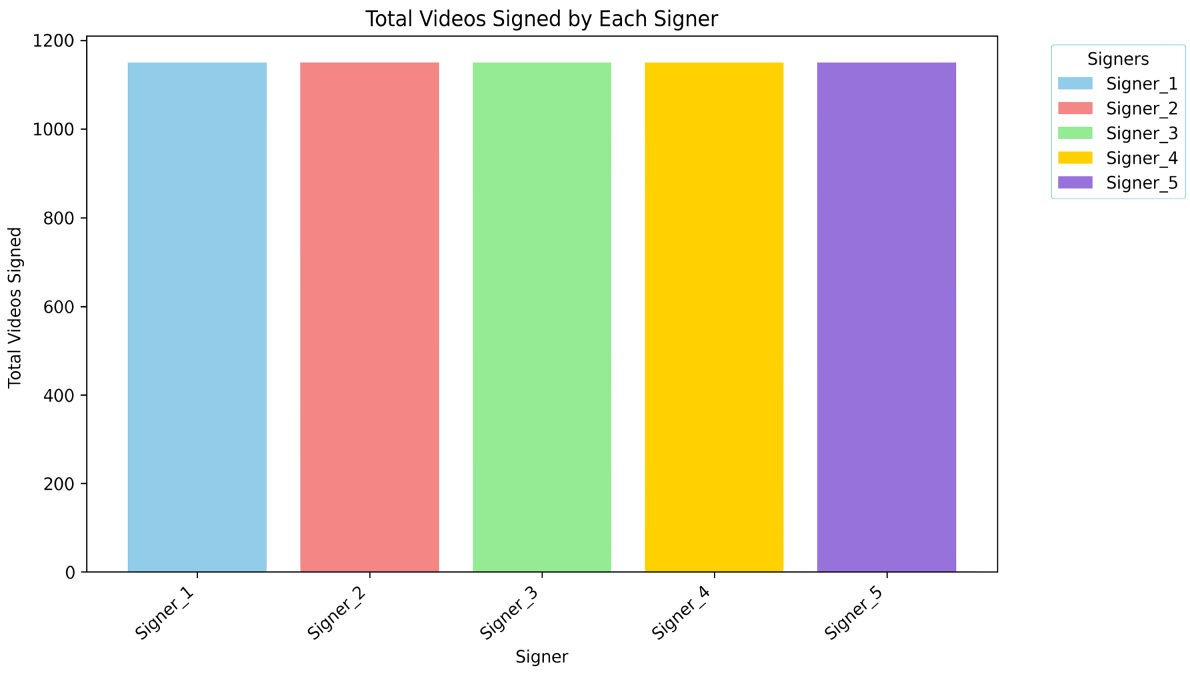

Figure 4: Total number of videos signed by each signer.

3.3.2. Pose-Relative Normalization

Normalization is particularly critical in this context, as the data is sensitive to signer distance from the camera, positional differences during recording, varying heights, arm lengths, and signer styles. Without normalization, models may overfit to signer-specific geometric characteristics, thereby reducing overall generalization performance [26]. The data is normalized by rescaling the coordinates to a fixed range of $[-1, 1]$, ensuring consistency across different signers. Equations (1) and (2a,2b) formalize translation (relative to the nose) and scaling (to the range of $[-1, 1]$), providing a straightforward and reproducible method for normalization.

Let $p_i = (x_i, y_i)$ represent the raw coordinates of keypoints *i*, and $p_{nose} = (x_{nose}, y_{nose})$ the nose's coordinates. The normalized coordinates $p'_i = (x'_i, y'_i)$ are computed as:

$$x'_i = \frac{x_i - x_{nose}}{s},\ y'_i = \frac{y_i - y_{nose}}{s} \tag{1}$$

where s is the scaling factor (e.g., average shoulder distance), and coordinates as rescaled to $[-1, 1]$ using:

$$x''_i = \frac{x'_i - \min(x')}{\max(x') - \min(x')} \tag{2a}$$

$$y''_i = \frac{y'_i - \min(y')}{\max(y') - \min(y')} \tag{2b}$$

3.3.3. Feature Representation

To reduce computational overhead, frame-level landmarks were extracted from each 5-s video recorded at approximately 30 frames per second, with each frame represented as a 244-dimensional feature vector. Sequence padding was applied using zero vectors during model training to ensure uniform batch processing without introducing artificial motion.

3.4. Architecture Design

Figure 5 illustrates the proposed model architecture, and a summary of the functions and parameters of the architectural layers is presented in Table 2. The input layer receives frame-wise feature vectors of size 244, which are compatible with the normalized keypoint coordinates extracted from each video frame, as represented in Equation (3).

$$X_t \epsilon R^d,\ d = 244 \tag{3}$$

where d denotes the number of coordinate features extracted per frame, comprising selected body, hand, and facial keypoint coordinates. For a video with T frames, the complete sequence can be represented by Equation (4).

$$X = \{x_1, x_2, \ldots, x_T\},\ X_t \epsilon R^d \tag{4}$$

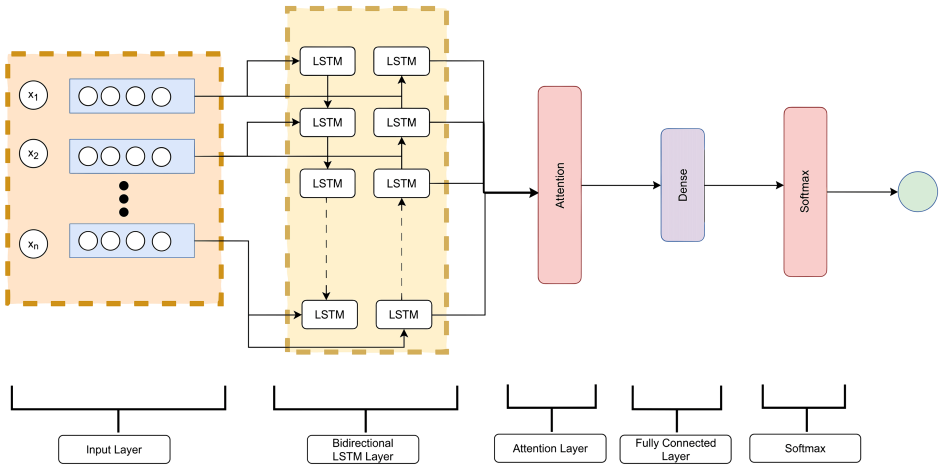

Figure 5: Architectural design of model.

Table 2: Components of the proposed model architecture.

Layer	Role	Size/Parameters	Hyperparameters
Input	Receives normalized keypoint sequences	T frames × 244 coordinate features	None
BiLSTM Layer 1	Capture temporal dependencies in both forward and backward directions.	256 units × 2	batch_first=True, bidirectional=True
BiLSTM Layer 2	Captures higher-level temporal patterns.	256 units × 2	batch_first=True, bidirectional=True
Attention Layer	Computes time-step importance and forms a context vector.	513 parameters	Softmax over sequence length
Fully Connected Layer	Prevents overfitting and improves generalization	Dropout = 0.5 after attention	None
Output Layer	Classifies the context vector into sign probabilities	115 units, Categorical Cross-Entropy	Softmax activation

The input layer is followed by stacked LSTM layers designed to effectively learn long-range dependencies within sequential data [34]. They are especially effective for sign language recognition, where there is a need to capture the entire duration of a sign gesture. The architecture employs two layers of bidirectional Long Short-Term Memory (BiLSTM) networks, with hidden units designed to capture temporal dependencies across sequences. Mathematically, the LSTM formulation is represented by Equations (5)–(11). Each LSTM cell preserves memory through the input, forget, and output gates:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{7}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{8}$$

$$c_t = f_t \bigodot c_{t-1} + i_t \bigodot \tilde{c}_t \tag{9}$$

$$h_t = o_t \bigodot \tanh(c_t) \tag{10}$$

where the relevant variables should be clearly defined, including the gates, cell state, and hidden state.

In the bidirectional model, each time step has both forward and backward hidden states, which help to incorporate past and future context while modeling sign gestures:

2026, Vol. 3, Article ID. 2026.0034
https://doi.org/10.69709/CAIC.2026.130891

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \tag{11}$$

This is followed by a Bahdanau-style additive attention mechanism with temporal variability, in which each hidden state in Equation (11) is projected into a scalar alignment score using a fully connected attention layer and subsequently normalized via the Softmax function to obtain attention weights, as defined in Equation (14). These weights quantify the relevance of each keypoint within the overall sequence representation. The last context vector is calculated using Equation (15), which is weighted towards more informative frames with less weight assigned to irrelevant frames. This mechanism enhances interpretability and robustness, enabling the model to focus on critical temporal regions of a sign.

The BiLSTM generates a sequence of hidden states, as indicated in Equation (12).

$$H = \{h_1, h_2, \ldots, h_T\},\ h_t\ \epsilon R^{2H} \tag{12}$$

For a sequence of BiLSTM hidden states, h_t (where $t = 1$, the attention score for each time step) is calculated by applying Equation (13).

$$e_t = v^T \tanh(W_h h_t + b_h) \tag{13}$$

Attention weights are computed via Softmax:

$$a_t = \frac{\exp(e_t)}{\Sigma_{k=1}^{T} \exp(e_k)} \tag{14}$$

The context vector is:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{15}$$

where v, W_h, and b_h are learnable parameters.

The vector produced by the attention mechanism is passed through a fully connected layer of 256 units with ReLU activation, coupled with dropout regularization set to 0.5 to reduce the risk of overfitting. The dropout formulation is presented in Equation (16).

$$z = Dropout(W_{f_c} c + b_{f_c}) \tag{16}$$

Table 3 presents the architectural summary of the baseline ConvLSTM model used as an initial test on the curated dataset. The architecture consists of Time-Distributed layers designed to effectively capture sequential patterns representing each signed word, thereby enabling more accurate recognition of temporal dependencies.

Table 3: Architectural summary of ConvLSTM.

Stage	Layers	Output Shape
ConvLSTM Block 1	ConvLSTM2D(4) + MaxPool	(20, 31, 31, 4)
ConvLSTM Block 2	ConvLSTM2D(4) + MaxPool	(20, 15, 15, 4)
ConvLSTM Block 3	ConvLSTM2D(8)	(20, 13, 13, 8)
ConvLSTM Block 4	ConvLSTM2D(8) + MaxPool	(20, 6, 6, 8)
Final ConvLSTM	ConvLSTM2D(16) + MaxPool	(20, 2, 2, 16)
Feature Aggregation	GlobalAveragePooling3D	16
Dense Head	Dense(32) + Dropout	32
Output	Dense(115)	115

To improve generalization, the ConvLSTM architecture was trained using different optimizers, including RMSprop and Adam. Key training parameters were systematically tuned, with batch sizes of 16, 32, and 64, and dropout rates in the dense layers adjusted between 0.2 and 0.5.

The Softmax classifier layer, connected to the fully connected layer, has an output dimensionality corresponding to the number of sign classes, which is 115 in the dataset. It produces a probability distribution over all possible classes, enabling the model to predict the most likely sign corresponding to a given input sequence. The training objective is defined using categorical cross-entropy loss, which penalizes incorrect predictions while encouraging high confidence in the correct class.

3.5. Experimental Setup

The implementation was carried out in Python 3.12 in an Anaconda environment using Jupyter Notebook, with PyTorch for model training, Scikit-learn for evaluation metrics and data splitting, Pandas and NumPy for data han-

dling, and Matplotlib for visualization, on an HP Pavilion system equipped with a 10th Gen Intel Core i5 processor and 16 GB of RAM. MediaPipe Holistic was employed for keypoint extraction.

3.6. Training Procedure

Models were trained for 100 epochs with early stopping (patience = 10 based on validation loss). Batch sizes of 16, 32, and 64 were evaluated, with a batch size of 32 providing the best balance between training stability and computational efficiency. The Adam optimizer was employed with an initial learning rate of 0.001, incorporating exponential decay for sparse gradients. L2 regularization (λ = 0.0001) was applied to dense layers, while a dropout rate of 0.5 was introduced to mitigate overfitting. Categorical cross-entropy loss was utilized for multi-class classification.

Model performance was evaluated using five metrics: overall accuracy, precision, sensitivity, F1-score, and specificity. The models were evaluated on the test set to assess their final performance in terms of accuracy, precision, recall, and F1-score. Results were also benchmarked against prior SLR studies for robustness and scalability.

4. Results and Discussion

4.1. Summary of Dataset

In this study, a lack of publicly available and up-to-date Ghanaian Sign Language (GSL) datasets was identified. To address this gap, the AkwaabaSign dataset, a multi-signer, video-based dataset was introduced, as highlighted in the literature [11]. The dataset is stratified to ensure signer diversity across all subsets, thereby reducing bias arising from individual signing styles, a common issue in small-scale SLR datasets [26]. Figure 6 illustrates the distribution of data used for training, validation, and testing.

4.2. Training Results

In this study, two baseline models were trained on the proposed dataset: an attention-based LSTM and a ConvLSTM. During initial training, the attention-based LSTM achieved 65.2% training accuracy and 52.14% validation accuracy. This was primarily due to the high-dimensional input space, which amplified noise from minor variations in facial expressions or body positioning unrelated to GSL semantics. A multifaceted optimization strategy was implemented to address this, resulting in a more robust normalization approach. Hyperparameter tuning significantly improved model performance, with the attention-based LSTM model achieving improved results.

The model achieved a training accuracy of 99.06% and a validation accuracy of 95.13%. The loss curves showed smooth convergence, indicating stable and consistent learning over 100 epochs. It achieved 94.69% accuracy on the test set, with weighted averages of 93.32% precision, 92.70% recall, and 92.66% F1-score. The ConvLSTM achieved a training accuracy of 94.49% and a validation accuracy of 90.28% over 100 epochs, with a validation loss of 0.854. While these results are respectable, they are lower than those obtained by the attention-based model.

The LSTM model, particularly in validation, exhibited limitations in capturing the temporal nuances of GSL without explicit attention weighting. Figure 7 illustrates the pre- and post-optimization accuracy and loss curves of the attention-based LSTM, highlighting the impact of these interventions on model convergence. Figures 8 and 9 illustrate the total training and validation loss, as well as the loss curve for the ConvLSTM model, highlighting a steep decline from approximately 4.5 to 1.5 within 10 epochs, indicating effective learning and a reduction in error. After the initial decline, the validation loss remained slightly above the training loss and stabilized within a comparable range, which is consistent with expected model behaviour and indicates reasonable generalization on unseen data. The evaluation metrics used in this study include accuracy, precision, recall, and F1-score. The proposed attention-based model achieved a test accuracy of 94.69%. Precision (93.32%) indicates a low rate of incorrect sign classifications, recall (92.70%) reflects the model's ability to identify actual instances, and the F1-score (92.66%) is also notable.

Per-class analysis indicated strong performance on visually distinct signs, such as "hospital," which achieved an F1-score of 98% owing to its distinctive handshapes and poses. In contrast, reduced accuracy (85–90%) was observed for signs with overlapping motion trajectories, such as "learn" and "study," primarily due to subtle variations in non-manual markers. Figure 10 presents the confusion matrix for the attention-based LSTM across all classes, demonstrating effective sign recognition, as indicated by the prominent diagonal pattern representing correct predictions. The distribution from 0 to 8, as shown in the legend color scale, further reflects the model's prediction performance.

2026, Vol. 3, Article ID. 2026.0034
https://doi.org/10.69709/CAIC.2026.130891

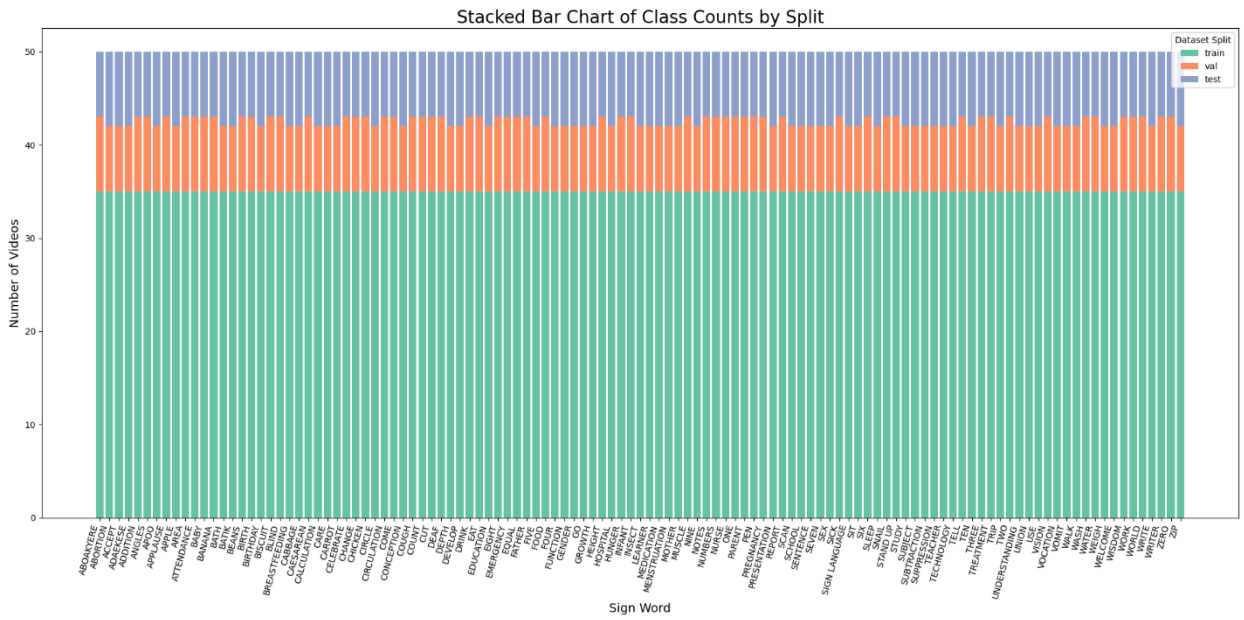

Figure 6: Illustrates the distribution of videos across the splits, confirming class and signer balance.

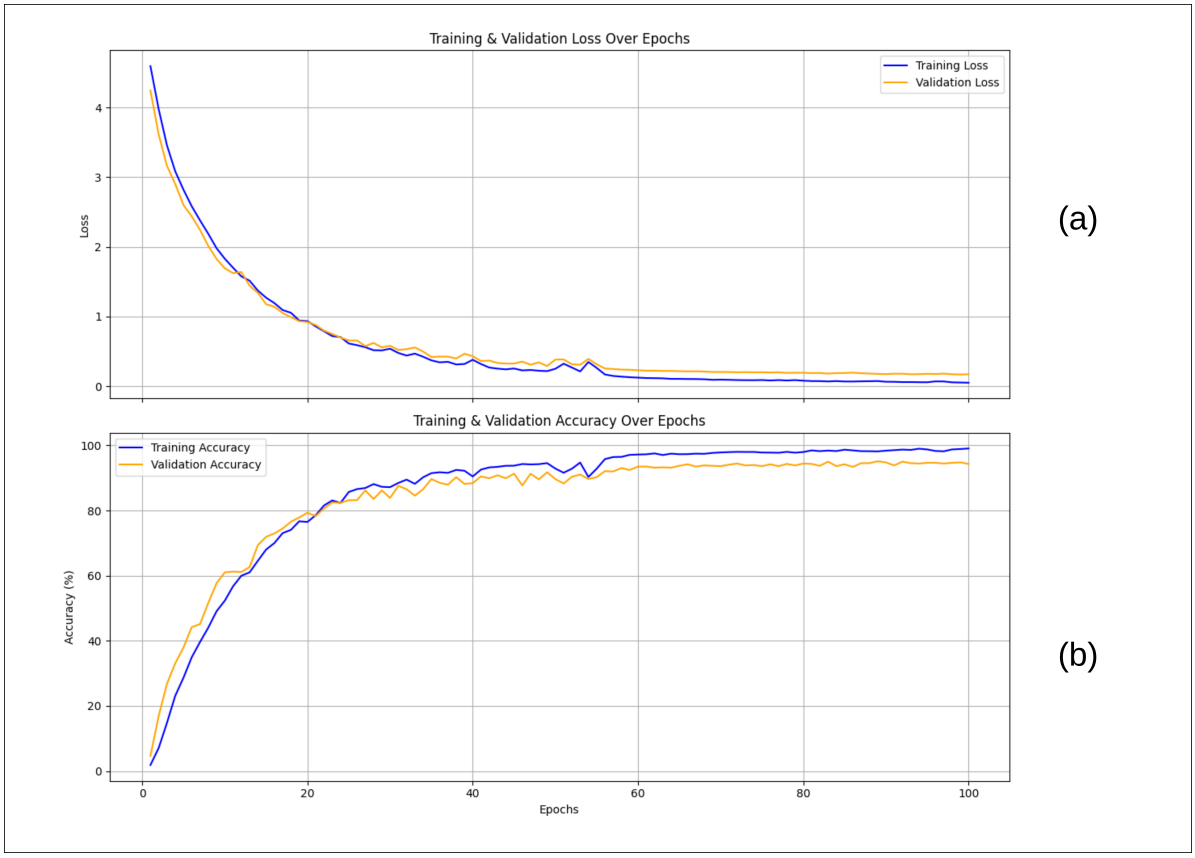

Figure 7: Training and validation accuracy and loss curves for the proposed attention-based LSTM model (**a**) Loss and (**b**) Accuracy.

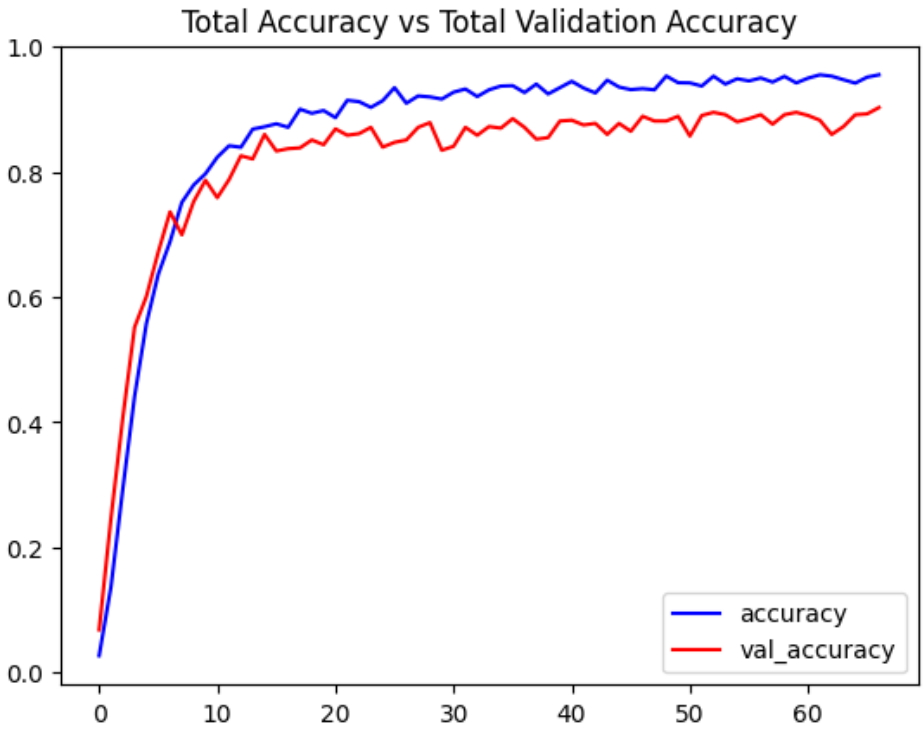

Figure 8: Training and validation accuracy curves for the ConvLSTM baseline model.

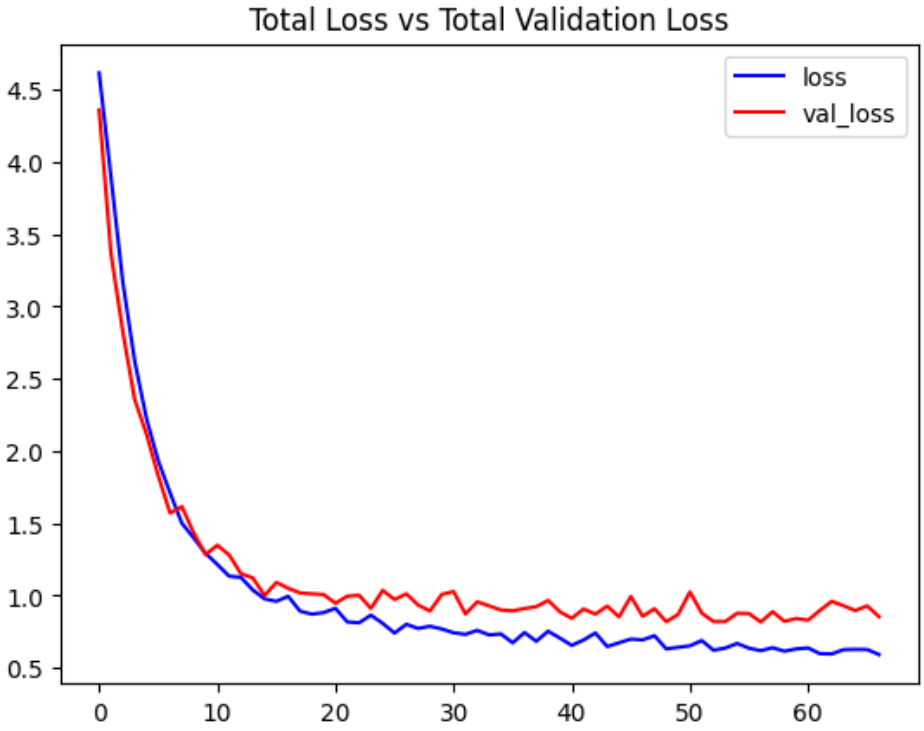

Figure 9: Loss and validation loss curves for the ConvLSTM baseline model.

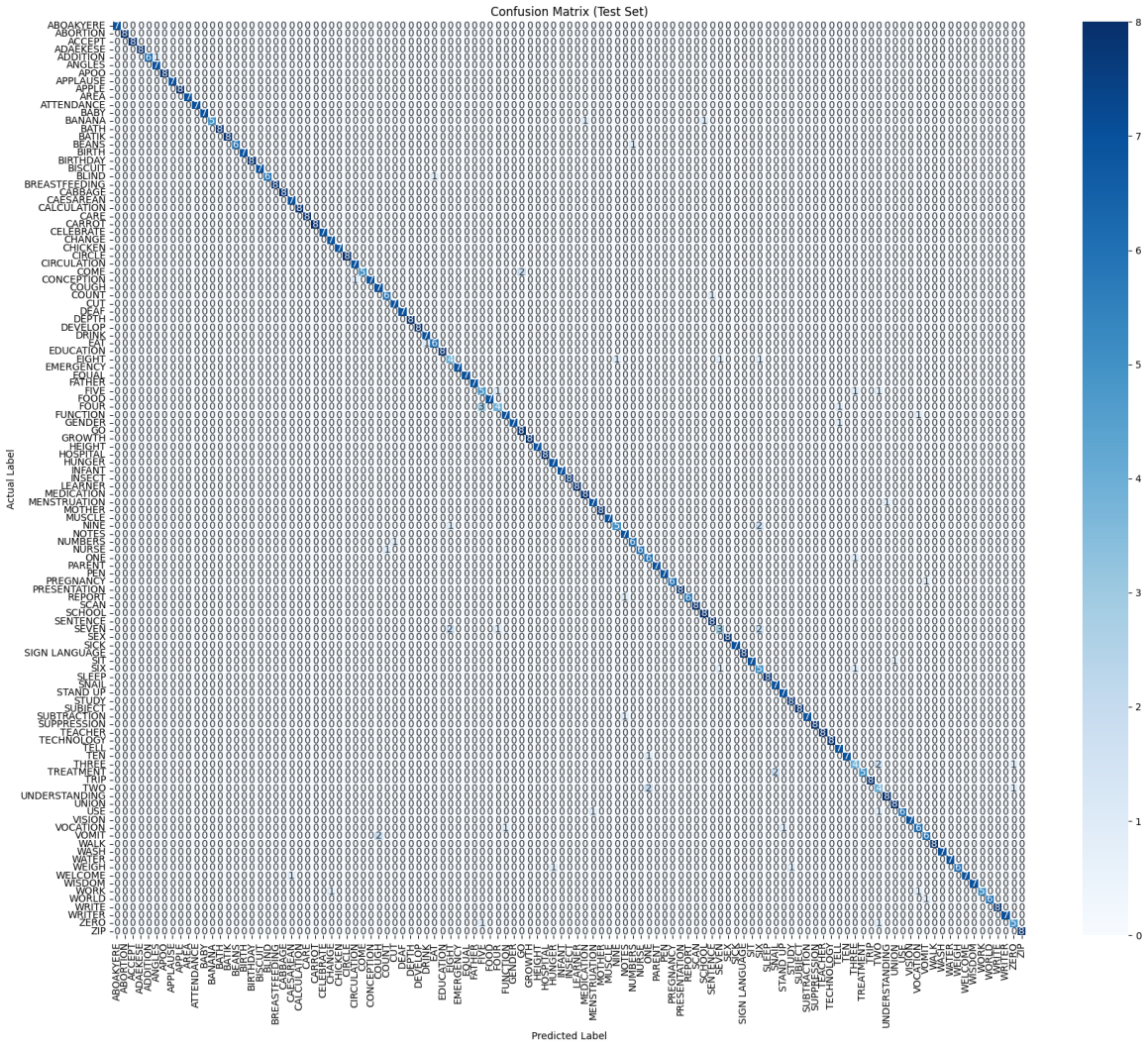

Figure 10: Confusion matrix for the proposed sign language recognition model.

5. Conclusions

This research achieved its objectives by introducing the AkwaabaSign dataset, a multi-signer video collection that addresses the significant lack of publicly available GSL datasets, and by developing an optimized pipeline for processing 2D keypoint coordinates. The study also helps reduce the gap in GSL research in computational linguistics and machine learning, addressing the underrepresentation of GSL by proposing an attention-based LSTM for SLR. The results indicate that the combined use of keypoint-based representation, normalization, and attention mechanisms significantly improves recognition accuracy, achieving a test accuracy of 94.69%. Balanced measures include 93.32% precision, 92.70% recall, and 92.66% F1-score. The proposed approach, therefore, provides a straightforward method for recognizing isolated words in Ghanaian Sign Language. This study represents one of the first attempts to use keypoint-based input for GSL recognition.

Looking ahead, the AkwaabaSign dataset can be expanded to include sentence-level expressions. Interdisciplinary training programs in computational linguistics and sign language studies are also recommended to strengthen local capacity in under-resourced regions of Africa, particularly Ghana. Future work will primarily focus on extending the model to continuous GSL recognition, integrating transformer architectures with the existing LSTM–attention framework, improving signer-specific differentiation, conducting comprehensive ablation studies, and deploying the model on mobile devices for further evaluation. Future exploration of multimodal systems that integrate skeletal keypoints with textual outputs or other assis-

tive modalities for broader accessibility may enable the development of versatile frameworks for bidirectional communication.

List of Abbreviations

CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long Short-Term Memory
DHH	Deaf and Hard-of-Hearing
GSL	Ghanaian Sign Language
HMMs	Hidden Markov Models
LSTM	Long Short-Term Memory
RNNs	Recurrent Neural Networks
SL	Sign Language
SLR	Sign Language Recognition

Author Contributions

Conceptualization: K.T.; Methodology: K.T., L.T., and R.-M.O.M.G.; Software: K.T. and L.T.; Validation: K.T., R.-M.O.M.G., and L.T.; Formal analysis: K.T. and L.T.; Investigation: K.T. and L.T.; Resources: K.T.; Data curation: K.T. and L.T.; Writing—original draft preparation: L.T. and K.T.; Writing—review and editing: K.T., L.T., and R.-M.O.M.G.; Visualization: K.T. and L.T.; Supervision: K.T. and R.-M.O.M.G.; Project administration: K.T.; Funding Acquisition: K.T. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials

The AkwaabaSign dataset, curated as part of this study, is available at https://zenodo.org/records/15730487, accessed on 24 June 2025.

Ethical Approval and Consent to Participate

The study was conducted with ethical approval from the Committee on Human Research, Publications, and Ethics (CHRPE) at Kwame Nkrumah University of Science and Technology (KNUST), reference number CHRPE/AP/456/25, dated 28th May 2025. All signers provided informed consent, and their identities were anonymized to ensure confidentiality. Participant information was handled with strict confidentiality, and all individuals were adults aged 18 years or older. Participants were informed that the collected data would be used exclusively for this study and would be made publicly available only after getting their approval. Informed consent was also obtained from all individuals depicted in the graphical abstract, in accordance with the approved ethical clearance.

Human Rights Statement

The research was carried out in accordance with the Declaration of Helsinki, ensuring that no harm was caused to participants.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work is funded by the KNUST Research Fund (KREF) with award reference number KREF8/23/153/M4.

Acknowledgments

The authors gratefully acknowledge the participants from Kwame Nkrumah University of Science and Technology who generously volunteered their time for data collection and for supporting the evaluation of the proposed methods.

AI Declaration

The authors would like to declare that the AI-assisted tool (ChatGPT) was used to refine sentence structure and clarity of the manuscript. All descriptions of the AkwaabaSign dataset, data analyses, and scientific interpretations were independently authored, verified, and approved by the research team.

References

[1] WHO, "Deafness and hearing loss." Accessed: Jun. 7, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] J. Webster and J. Hosemann, "Sign language vitality through the lens of a pioneering interactive atlas: a first look at the sociolinguistic profile data collected by the sign hub project," *J. Linguist. Geogr.*, vol. 12, no. 2, pp. 84–104, Oct. 2024. [CrossRef]

[3] L. K. Odartey, Y. Huang, E. E. Asantewaa, and P. R. Agbedanu, "Ghanaian sign language recognition using deep learning," in *ACM Int. Conf. Proc. Ser.*, Aug. 2019, pp. 81–86. [CrossRef]

[4] I. Sandjaja, A. Alsharoa, D. Wunsch, and J. Liu, "Survey of hidden Markov models (HMMs) for sign language recognition (SLR)," *Ind. Cyber-Phys. Syst.*, 2024, pp. 1-6. [CrossRef]

[5] R. Kumar, A. Sinha, A. Bajpai, and S. K. Singh, "A comparative analysis of techniques and algorithms for recognising sign language," 2023, *arXiv:2305.13941*. [CrossRef]

[6] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art no. 113794. [CrossRef]

[7] S. Subburaj and S. Murugavalli, "Survey on sign language recognition in context of vision-based and deep learning," *Meas. Sens.*, vol. 23, Oct. 2022, Art no. 100385. [CrossRef]

[8] S. Tan, N. Khan, Z. An, Y. Ando, R. Kawakami, and K. Nakadai, "A review of deep learning-based approaches to sign language processing," *Adv. Robot.*, vol. 38, no. 23, pp. 1649–1667, 2024. [CrossRef]

[9] N. Adaloglou et al., "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. Multimed.*, vol. 24, pp. 1750–1762, Mar. 2021. [CrossRef]

[10] M. Bohacek and M. Hruz, "Sign pose-based transformer for word-level sign language recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, 2022, pp. 182–191. [CrossRef]

[11] M. Edward and G. Akanlig-Pare, "Sign language research in Ghana: an overview of indigenous and foreign-based sign languages," *J. Afr. Lang. Lit.*, vol. 2, pp. 114–137, 2021. [CrossRef]

[12] I. Murtagh, V. U. Nogales, and J. Blat, "Sign language machine translation and the sign language lexicon: a linguistically informed approach," 2022. Accessed: Jan. 29, 2026. [Online]. Available: https: //aclanthology.org/2022.amta-research.18/

[13] M. S. Amin, S. T. H. Rizvi, and M. M. Hossain, "A comparative review on applications of different sensors for sign language recognition," *J. Imaging*, vol. 8, no. 4, Apr. 2022, Art. no. 98. [CrossRef]

[14] M. A. Khan et al., "Human action recognition using fusion of multiview and deep features: an application to video surveillance," *Multimed. Tools Appl.*, vol. 83, no. 5, pp. 14885–14911, Mar. 2024. [CrossRef]

[15] W. Burger and M. J. Burge, "Scale-invariant feature transform (SIFT)," in *Digital Image Processing: An Algorithmic Introduction*. Cham, Switzerland: Springer International Publishing, 2022, pp. 709–763. [CrossRef]

[16] T. Raghuveera, R. Deepthi, R. Mangalashri, and R. Akshaya, "A depth-based Indian sign language recognition using Microsoft Kinect," *Sadhana Acad. Proc. Eng. Sci.*, vol. 45, no. 1, pp. 1–13, Dec. 2020. [CrossRef]

[17] J. Shin et al., "Korean Sign Language recognition using transformer-based deep neural network," *Appl. Sci.*, vol. 13, no. 5, Mar. 2023, Art. no. 3029. [CrossRef]

[18] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video," *Int. J. Comput. Vis.*, vol. 126, no. 2–4, pp. 430–439, Apr. 2018. [CrossRef]

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," 2017, *arXiv:1611.08050*. [CrossRef]

[20] T. J. Sánchez-Vicinaiz, E. Camacho-Pérez, A. A. Castillo-Atoche, M. Cruz-Fernandez, J. R. García-Martínez, and J. Rodríguez-Reséndiz, "MediaPipe frame and convolutional neural networks-based fingerspelling detection in mexican sign language," *Technologies*, vol. 12, no. 8, Jul. 2024, Art. no. 124. [CrossRef]

[21] Y. Kim, H. Baek, and J. M. Corchado, "Preprocessing for keypoint-based sign language translation without glosses," *Sensors*, vol. 23, no. 6, Mar. 2023, Art. no. 3231. [CrossRef]

[22] F. Shafizadegan, A. R. Naghsh-Nilchi, and E. Shabaninia, "Multimodal vision-based human action recognition using deep learning: a review," *Artif. Intell. Rev.*, vol. 57, no. 7, Jun. 2024, Art. no. 178. [CrossRef]

[23] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd Int. Conf. Learn. Represent.*, Sep. 2014. Accessed: Aug. 12, 2025. [Online]. Available: https://arxiv.org/pdf/1409.0473

[24] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, vol. 21, no. 4, Feb. 2021, Art. no. 1120. [CrossRef]

[25] A. Núñez-Marcos, O. Perez-de-Viñaspre, and G. Labaka, "A survey on sign language machine translation," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118993. [CrossRef]

[26] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 2529–2539. [CrossRef]

[27] E. J. Robert and H. J. Duraisamy, "A review on computational methods based automated sign language recognition system for hearing and speech impaired community," *Concurr. Comput.*, vol. 35, no. 9, Apr. 2023, Art. no. e7653. [CrossRef]

[28] Q. Wu, Q. Huang, and X. Li, "Multimodal human action recognition based on spatio-temporal action representation recognition model," *Multimed. Tools Appl.*, vol. 82, no. 11, pp. 16409–16430, Nov. 2022. [CrossRef]

[29] C.-F. Chen et al., "Deep analysis of cnn-based spatio-temporal representations for action recognition," presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 13–19, 2020, pp. 6161–6171. [CrossRef]

[30] B. Zhang, M. Müller, and R. Sennrich, "SLTUNET: a simple unified model for sign language translation," 2023, *arXiv:2305.01778*. [CrossRef]

[31] S. H. Noh, "Analysis of gradient vanishing of RNNS and performance comparison," *Information*, vol. 12, no. 11, Oct. 2021, Art. no. 442. [CrossRef]

[32] B. Dabwan and M. Jadhav, "A CNN-LSTM model for Arabic sign language recognition," in *First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (AC-*

VAIT 2022), Paris, France: Atlantis Press, 2023, pp. 459–470. [CrossRef]

[33] K. K. Podder et al., "Signer-independent Arabic sign language recognition system using deep learning model," *Sensors*, vol. 23, no. 16, Aug. 2023, Art. no. 7156. [CrossRef]

[34] S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition of skeletal videos using processed video driven multi-stacked deep LSTM," *Sensors*, vol. 22, no. 4, Feb. 2022, Art. no. 1406. [CrossRef]