Computing&AI Connect
ISSN: 3104-4719
2026, Vol. 3, Article ID. x, Cite as: https://www.doi.org/10.69709/xxx

SCIFINITI PUBLISHING

OPEN ACCESS

Research Article

Attention-Based LSTM for Sign Language Recognition Leveraging Spatial-Temporal Keypoint

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

[1] Department of Computer Science, Faculty of Computational and Physical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Abstract

Sign language is a crucial means of communication for the Deaf and hard-of-hearing communities. Most individuals find it challenging to communicate with the deaf when they try to do so without an interpreter. The advancement in technology, computer vision, and deep learning approaches provides a different approach to tackling the problem. Literature indicates that the unique nature of Ghanaian Sign Language (GSL) has been understudied due to a lack of large and publicly available datasets, as well as limited research on the use of landmark keypoints for computational research on GSL. This study curated a large video-based dataset, AkwaabaSign, that reflects the indigenous nature of the GSL. The study employed two different baseline models to assess the dataset: an Attention-enhanced LSTM model and a ConvLSTM model, which extracted and specially normalized the keypoints using Mediapipe. With this approach, the attention-enhanced LSTM achieved a test accuracy of 94.69%, with balanced performance metrics of 93.32% precision, 92.70% recall, and 92.66% F1-score. The ConvLSTM achieved 90.28% accuracy, lagging behind the attention-enhanced LSTM. The study fulfils the aim of producing a large dataset for sign language recognition, provides a specialized normalization process for dataset processing, and establishes a base model for the practical use of the dataset. The proposed model also outperforms some other algorithms in the domain of sign language computational works in GSL. The study aims to expand the dataset to the sentence level and develop continuous GSL recognition.

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

1. Introduction

The medium of communication is a vital tool for sharing information. Done either verbally or non-verbally, it allows the exchange of information. Sign language (SL) is a visual, natural means by which Deaf and Hard of Hearing (DHH) people use for everyday communication. Approximately 5% of the global population, around 430 million people, have lost their hearing and require rehabilitation, according to the World Health Organisation (WHO) [1]. There are several different sign languages based on regional differences, such as American Sign Language (ASL) and German Sign Language (DGS). Hence, it is not a universal language. Although different in form, it serves the same functions as a spoken language [2]. In the world we live in, most people do not understand SL, making communication with the DHH more challenging in areas of health and education. Sign Language serves as a bridge to this barrier. Linguists have long misunderstood sign languages as mere gestural accompaniment to speech. Still, William Stokoe's pioneer-

ing analysis of American Sign Language (ASL) stated that SL possesses its own phonological and grammatical structures, marking the recognition of sign languages as fully fledged human languages [3]. Researchers have since studied the various forms of sign languages and the information they convey visually, using manual and non-manual means of expression.Manual parameters encompass hand shape, hand posture, hand location, and hand motion, while non-manual parameters include head and body posture, facial expressions, gaze, 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, motivating the need for dedicated dataset curation and model development efforts. Most available datasets are from high-income countries, such as ASL and DGS, which do not capture the native nature of GSL. The scarcity of data leads researchers to create their own datasets to train models [3].

Globally, research in sign language recognition has evolved significantly, with early systems relying heavily on handcrafted features and sensor-based gloves. These systems often use Hidden Markov Models (HMMs) to model temporal dynamics. [4], but were limited in scope and practicality [5]. Comprehensive surveys underlining recent developments in sign language have shown a drift to the use of concepts such as deep learning, especially convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and hybrid models, often achieving high accuracy on isolated datasets, and how the deep learning architecture is dominating the revolution of the field [6], [7], [8].

Despite these global advances, many sign languages, such as GSL, which are linguistically distinct in their own right, with unique phonology and morphological patterns, are under-researched. Studies have indicated that there is a severe shortage of publicly available datasets for GSL, which can be observed in the research by [3]. They created a self-created dataset for the model, which achieved an accuracy of around 96% using CNNs and transfer learning. 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 hinders the development of robust GSL recognition tools; second, research into native appropriate sign language technology for Ghana's Deaf community is set back by this scarcity. Addressing this shortfall requires the collection of datasets and technological implementations tailored to GSL.

At the same time, rapid advancements in computer vision and sequential modelling open new possibilities. MediaPipe and other modern pose-estimation frameworks can reliably and quickly extract human keypoints, enabling lightweight feature representation from video frames. Models that combine Long Short-Term Memory (LSTM) units with attention mechanisms or transformer-based architectures have demonstrated strong performance in sign and gesture recognition, effectively capturing temporal and spatial dependencies in sequential data [8], [9], [10].
With the growing interest in African sign language, Ghanaian Sign Language (GSL) remains significantly underrepresented in computational linguistics and machine learning research. Most existing Sign language recognition systems are based on American or British sign language datasets, and indigenous datasets that are collected are labour-intensive or small, falling short of meeting the robustness and scalability required for 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 utilises 2D keypoint coordinates (face, hands, and body pose) for gesture recognition
- Proposed a robust normalisation strategy that enhances recognition accuracy by aligning keypoints with the canonical representation
- Designed a deep learning model architecture that utilises spatial and temporal features to better recognise 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, the data preprocessing step, the normalization approach, and the model architecture. Section 4 provides a summary of the results, including an analysis of the base models, a comparison with existing research, and a description of the evaluation metrics used. Section 5 presents a detailed discussion of the findings. Section 6 concludes the paper and outlines recommendations for future research directions.

2. Literature Review

In this section, we highlight various research works related to sign language recognition. The review commences

with theoretical insights into the linguistic and visual attributes of sign languages, and then dissects the conceptual building blocks of SLR systems. It culminates in an empirical synthesis of key investigations across diverse sign languages, emphasizing advancements in machine learning, deep learning, and computer vision, while spotlighting GSL's marginalization in these domains, issues like dataset scarcity, signer variability, and generalization limitations that this study directly confronts through its custom AkwaabaSign dataset and efficient modelling techniques.

Historically, sign languages were often misunderstood as mere pantomime or simplified gestural systems. However, this perception was overturned by William Stokoe's pioneering work, which established American Sign Language (ASL) as a legitimate linguistic system. Just as any spoken language, sign languages are fully fledged natural languages with phonological, morphological, syntactic, and semantic systems, expressed through visual-manual [12]. While spoken language utilizes auditory signals for relaying information, sign Language uses manual parameters such as hand shape, hand location, hand orientation, and hand movement, with non-manual markers such as facial expressions, gaze direction, and body posture, which form the linguistic system of sign language recognition worldwide, although research coverage remains uneven [11]. GSL, although having a significant sociolinguistic benefit to the DHH community in Ghana, is underrepresented both as a language that stands on its own and in its computational exploration, as most research is carried out using datasets of ASL, BSL, or DGS, which are available compared to the absence of large-scale datasets and deep learning benchmarks [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 SLR system designs utilized data gloves and motion sensors to capture the position of the hands and fingers, aiding in gesture identification [13]. Coupled with statistical and mathematical models, as well as the development of machine learning and computer vision, research in sign language has taken several directions. Feature extraction techniques such as Histogram of Oriented Gradients (HOG) [14] and Scale-Invariant Feature Transform (SIFT) [15] have been used with classifiers, Support Vector Machines (SVMs) or Hidden Markov Models (HMMs), achieving accuracies of 78.85% on RGB images [16]. Although this approach yielded promising results, it is flawed when faced with variability in signers, lighting, and background.

Advancements in deep learning have presented modern architectures, such as Convolutional Neural Networks (CNNs), which are key components of spatial feature extraction, and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, have been used to address both the temporal dynamics and modelling of sequential data that provide more robust and reliable findings. Shin et al. applied a CNN model to the KSL dataset, obtaining excellent performance compared to the existing system [17]. Pigou et al. demonstrated in their work the end-to-end learning of raw video without the use of handcrafted features, which achieved a high recognition rate [18]. This enabled lightweight, language-agnostic SLR pipelines with the help of 2D or 3D keypoints of body, hand, and face landmarks, as well as the adaptation of Pose estimation frameworks like OpenPose [19] and MediaPipe [20] can be used to produce lightweight models, minimizing the effects of variation in background and lighting [21], [22].

In recent years, self-attention-based architectures [23], especially transformers, have shown impressive performance on sequence modelling tasks, often matching or surpassing the capabilities of recurrent neural networks (RNNs) in capturing long-range dependencies within sequential data [24]. These advancements have also been extended to sign language recognition (SLR), where transformer-based models have achieved state-of-the-art accuracy in recognizing [10], [25]. Their computational complexity, however, has hindered their application in resource-constrained settings.

The intersection of linguistics, computer vision, and machine learning contexts has been seen 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 modelling, and classification. The choice of methodology is determined by decisions made at each of these stages, which influence recognition performance, computational efficiency, and adaptation [9]. The process of data collection commonly employs RGB cameras for recording signs, rather than depth sensors or infrared sensors, as they are more affordable and deliver accurate results [5], [10]. Preprocessing involves transforming the acquired data through various means, including frame resizing, background subtraction, segmenting signing regions, and normalizing spatial coordinates, thereby placing the data in a form that facilitates easy manipulation [27]. In using pose-based approaches, preprocessing is closely linked to the output of landmark detection algorithms such as MediaPipe Holistic, which provide high-dimensional keypoint coordinates for the face, hands, and upper body. Early SLR research relied on handcrafted descriptors such as Histogram of Oriented Gradients,

Scale-Invariant Feature Transform, and Motion History Images for feature extraction; however, modern approaches integrate deep neural networks, where convolutional layers automatically derive hierarchical spatial features [28], [29]. Recently, skeletal keypoint-based representations have emerged as a lightweight alternative, particularly beneficial for datasets with limited samples, as they preserve essential motion and posture information while reducing computational cost [30].
Sign language is inherently temporal and therefore requires sequential modelling, playing a central role in recognition. HHMs have been utilized for real-time recognition by capturing state transitions; however, they were limited in their ability to handle long-range dependencies. RNNs, such as LSTM, address issues related to HHMs, including the vanishing gradient problem [31]. The incorporation of attention mechanisms and transformers enables models to dynamically weight temporal segments and focus on the most informative portions of a sequence [24].

Compared to traditional methods, deep learning has achieved impressive results in computer vision. They have been employed in sign language and gesture recognition, utilizing techniques such as CNN, LSTM, and RNN. Chen et al. proposed a 3D-CNN as a robust approach for human action recognition, which was later adapted for SLR, enabling the direct learning of spatial and temporal features from raw video frames [29]. It is more common to see a mix of these architectures, like a CNN-LSTM-based model proposed for the recognition of Arabic sign language, temporal convolution, and bidirectional RNNs [22], [32].

Instead of working with extensive and computationally intensive raw images, pose estimation with keypoints has become an adapted approach that relies on 2D or 3D skeletal landmarks of the face, hands, and body, significantly reducing input dimensionality while maintaining essential cues such as motion, which are core components for SLR [9]. When compared to CNN-based, RNN-based, and pose-based SLR methods, they are computationally efficient and more robust to background variation, an essential consideration in low-resource settings. Kumar et al. conducted a comparative analysis of various sign language recognition (SLR) techniques, traditional vision-based methods, the use of deep learning architectures, and, where possible, hybrid approaches [5]. The review demonstrated that keypoint-based representations of data are advantageous in scenarios where computational resources are limited and there is a need to mitigate limitations such as signer dependency and background.

There have been significant global advancements in SLR studies, but research on SLR in Africa remains relatively underexplored. The review of this study reveals that pose-based approaches combined with attention-enhanced recurrent architectures often present an effective solution for isolated sign recognition, particularly in resource-constrained environments. These approaches address common challenges, including signer variability, background clutter, and dataset limitations, in the GSL. Using MediaPipe for pose estimation and an attention-based LSTM for temporal modelling, we aim to leverage the strengths of the most effective empirical approaches directly and address the gaps identified in Ghanaian sign language research.

3. Methodology and Experiments

In this work, we propose a pose-based approach for recognizing Ghanaian Sign Language (GSL). Here, we outline the step-by-step process for developing the pose-based GSL recognition system.

3.1 Research Design

Using human keypoints extracted from video frames, rather than raw pixel data, to reduce dimensionality while preserving motion and structural cues essential for sign recognition [9]. This research proposes a methodology that enhances generalization across signers by minimizing signer dependency, reducing overfitting to background or clothing, ensuring computational efficiency for deployment on low-resource hardware, and 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 presents a flowchart of the pipeline, which encompasses data curation, keypoint extraction using MediaPipe Holistic, normalization, model training (with an attention-enhanced LSTM baseline and ConvLSTM comparison), and evaluation.

3.2 Data Description

The dataset used in this study is a novel, custom-compiled, publicly available dataset for GSL, designed to address the lack of resources for sign language research in Ghana. It comprises 5,750 videos of 115 words, each signed by five indigenous signers (with≥10 videos per word-signer pair), and each is ~5 seconds long (30 FPS, 150 frames) in .avi format. The videos were captured in a controlled environment under standardized lighting conditions with a Sony camera and an ikan teleprompter for capture. Words were selected in consultation with deaf educators and linguists, with a focus on education (71) and health (44). Figure 2 illustrates the distribution by category per signer, and Figure 3 displays the hierarchical directory structure, which consists of words and subfolders per signer, with each signer's folder containing 10 videos. A graphical representation of the total number of videos signed by each signer is shown in Figure 4.

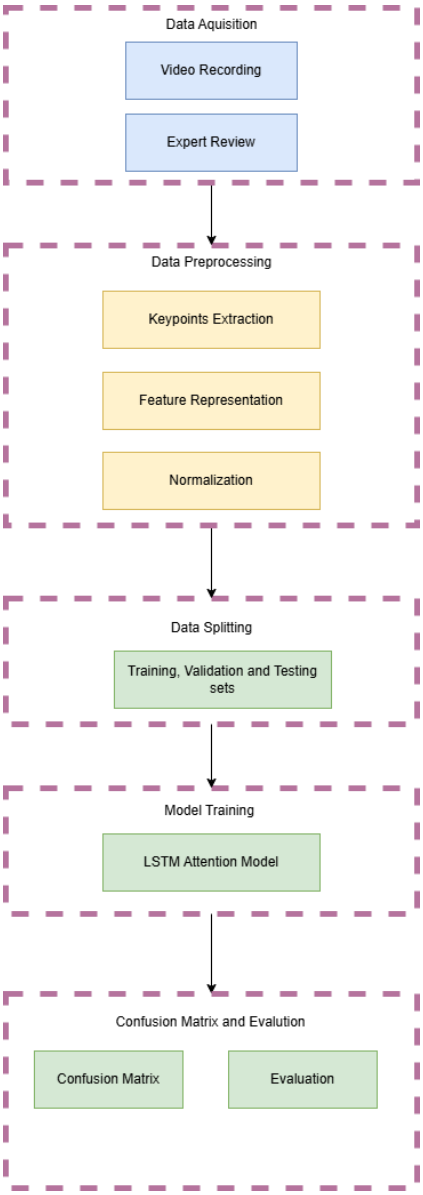

Figure 1: Flowchart of the Proposed Methodology Pipeline for GSL Recognition

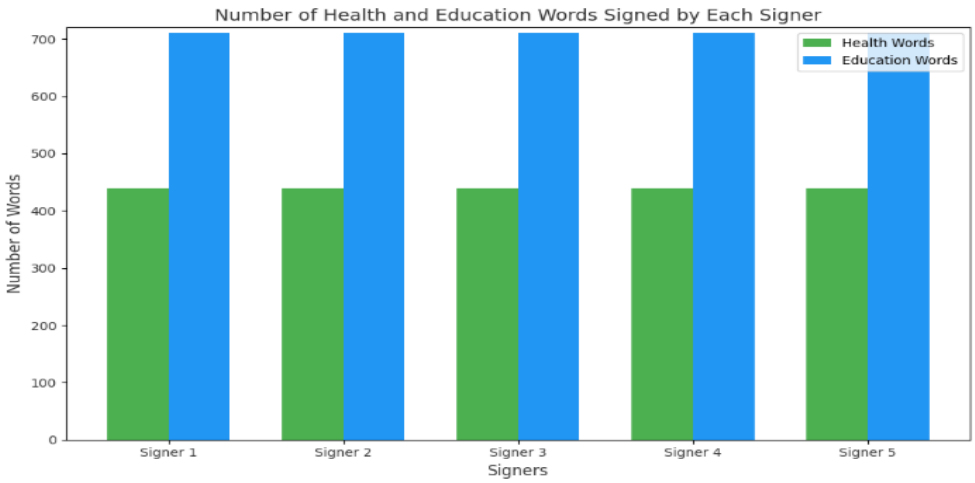

Figure 2: Plot of Videos signed by signer per category

3.3 Data Preprocessing and Feature Engineering

In this study, the raw video data set is processed into a form that makes it effective for use in training by scaling, augmenting, and normalizing, ensuring that it simulates real-world variations and enhances generalization.

3.3.1 Keypoint Extraction

Using the MediaPipe Holistic framework created by Google Research as the extraction tool, the videos are processed frame by frame, outputting 543 keypoints that cover the face, hands, and whole-body pose [20]. For the proposed study, a subset of landmarks was selected to capture the most relevant features, enabling robust sign language detection. The landmarks used include 33 pose landmarks for the upper body, 21 for each hand, and a subset of the 468 facial landmarks. Using the keypoints provides two main advantages over traditional approaches: reducing the input size compared to raw frames and lowering computational cost.

3.3.2 Pose-relative Normalisation

Normalisation 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 normalisation, models risk overfitting to signer-specific geometries, thus reducing generalisation performance [26]. The data is normalised by rescaling the coordinates into a fixed range of [–1, 1] to standardise across signers. Equations 1 and 2 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 normalization coordinates $p_i' = (x_i', y_i')$ are computed as:

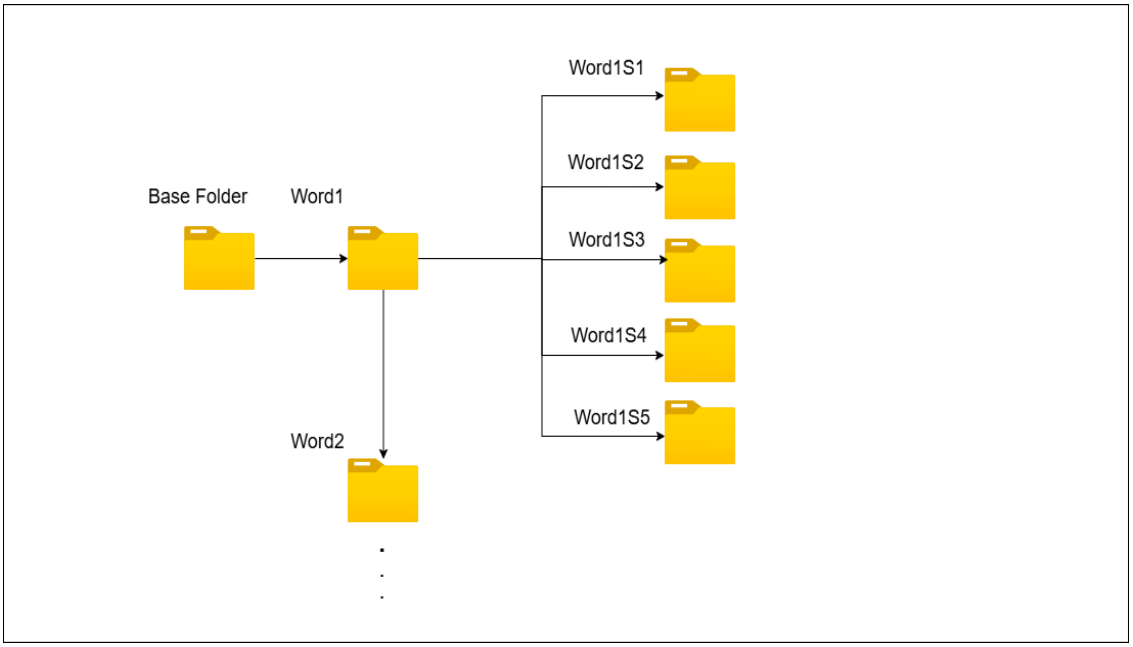

Figure 3: Hierarchical representation of dataset structure

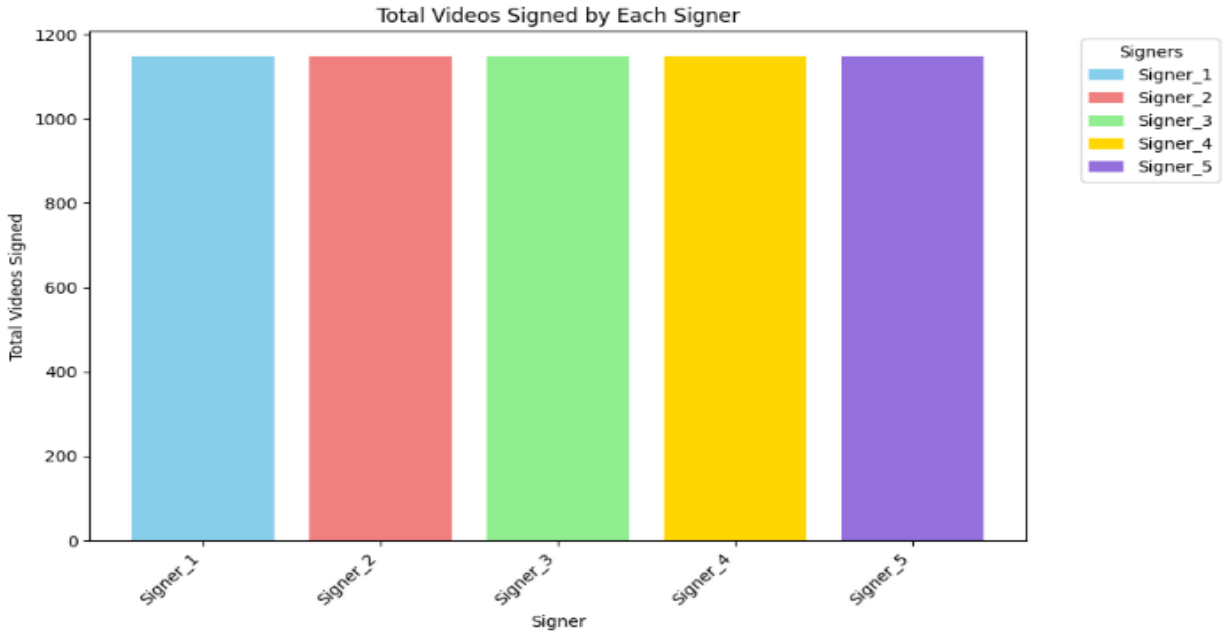

Figure 4: Total videos signed by each Signer

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

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

$$x_i^{"} = \frac{x_i' - \min(x')}{\max(x') - \min(x')} \cdot 2 - 1\,,$$
$$y_i'' = \frac{y_i' - \min(y')}{\max(y') - \min(y')} \cdot 2 - 1 \quad (2)$$

3.3.3 Feature Representation

To reduce computational overhead, once the frame-level landmarks were extracted from the 5 second video, approximately 30 frames per second to ensure uniformity they are padded into a fixed length of 244 with a zero vector to avoid introducing artificial movement and normalised, keypoints across frames from each video are compiled into a NumPy array to form the aggregated temporal sequence creating a lightweight, serialized and well-structured input for training the model.

3.4 Architecture Design

Figure 5 illustrates a visual representation of the proposed model architecture, and a summary of the architecture layer's functions and parameters is presented in Table 3. The input layer has a size of 244, which is compatible with the landmarks extracted from the sequence of frames t, represented as a feature vector in Equation 3.

$$X_t \epsilon R^d, \qquad d = 244 \quad (3)$$

Where d is the number of extracted landmarks (including body, hands, and facial keypoints). For a video with T Frames, the whole sequence can be represented in Equation 4.

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

The input layer is then stacked with LSTMs designed for effectively learning long-range dependencies in sequences [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, each with hidden units that model temporal dependencies across units. Mathematically, the LSTM can be represented by Equations 5-11. Each LSTM cell maintains memory using input, forget, and output gates:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (6)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (7)$$
$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (8)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (9)$$
$$h_t = o_t \odot \tanh(c_t) \quad (10)$$

Where i_t, f_t, o_t are the gates, c_t is the cell state, and h_t is represent the hidden state.

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

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (11)$$

This is followed by a Bahdanau-style additive attention layer with time variance, where each hidden state in Equation 11 is projected to a scalar score by a fully connected attention layer and normalized by a SoftMax function to obtain an attention weight as implemented in Equation 14, which is used to estimate the relevance of a keypoint to the complete representation of a sequence. 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.

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

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 and very sensitive to pose (reduces background noise)	High (Product 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)
Images Based	Static Images	Single frame capture	Low (Mostly sensitive to signer gestures)	Low to Moderate as it uses dingle-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)

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 these are learnable parameters.
The vector produced by the attention 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 formula for the dropout is shown in Equation 16.

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

Table 2 shows the architectural summary of the baseline ConvLSTM that was used as the initial test on the accreted dataset. The structure consists of time distributed layers to facilitate proper recognition for the sequential patterns that represent a signed word.

To improve generalization, the ConvLSTM architecture, different optimizers such as RMSprop and Adam were employed with key training parameters systematically tuned, including batch size of 16, 32, and 64, and dense head adjustment between 0.2 to 0.5.

Table 2: *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

The softmax classifier layer connected to the fully connected layer has a dimensionality corresponding to the number of sign classes, 115 in the dataset. It outputs the probability distribution across all the possible courses, enabling the model to predict the most likely sign represented by an 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

We used Python 3.12 in an Anaconda environment with Jupyter Notebook, PyTorch (for model training), Scikit-learn (for metrics and splitting), Pandas and NumPy (for data handling), Matplotlib for visualization on an HP Pavilion Core i5 10th Gen 16 gig RAM. MediaPipe Holistic was employed for keypoint extraction.

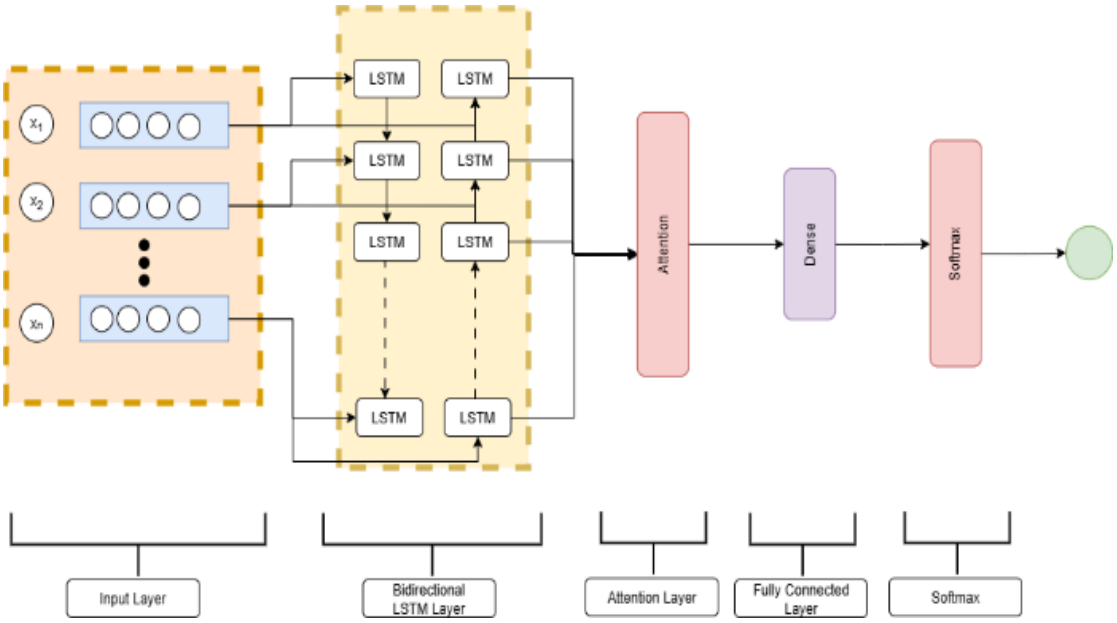

Figure 5: Architectural Design of Model

3.6 Training Procedure

Models were trained for 100 epochs with early stopping (patience = 10 on the validation loss). This study used batch sizes of 16, 32, and 64, and the batch size of 32 achieved a balance between stability and efficiency. Adam optimizer with an initial learning rate of 0.001, exponential decay applied to sparse gradients. L2 regularization with λ = 0.0001 was applied to dense layers, and dropout (0.5) was used to prevent overfitting. Categorical cross-entropy loss was used for multi-class classification.

The performance of the model is evaluated and validated using five evaluation metrics: overall accuracy, precision, sensitivity, F1-score, and specificity. The models were evaluated using 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, we found that there are no publicly available, updated GSL datasets. By introducing the AkwaabaSign dataset, a multi-signer video-based dataset, the study addresses this gap, as reviewed in the literature [11]. The dataset is stratified to maintain signer diversity in all subsets, thereby avoiding bias from signer styles a significant problem in small-scale SLR datasets [26]. Figure 6 illustrates the split of the data used for training, validation, and testing.

Table 3: Components of the Proposed Model Architecture

Layer	Role	Size/Parameters	Hyperparameters
Input	Receives normalized keypoint sequences	244 frames × 126 keypoints	None
BiLSTM Layer 1	Capture temporal dependencies in both forward and backward directions.	256 units x 2	batch_first=True, bidirectional=True
BiLSTM Layer 2	Captures higher-level temporal patterns.	256 units x 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 generalisation	Dropout = 0.5 after attention	None
Output Layer	Classifies the context vector into sign probabilities	115 units, Categorical Cross-Entropy	Softmax activation

4.2 The Results of the Training

In this study, two base models were trained on the novel datasets, an attention-based LSTM model and a ConvLSTM. The initial training of the attention-based LSTM yielded a training accuracy of 65.2% and a validation accuracy of 52.14%.

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 multi-faceted optimization strategy was implemented to address this, aligning to propose a robust normalization approach. Hyperparameter tuning of the model parameters significantly improved its performance, as the attention-based LSTM model achieved.

99.06% training accuracy and 95.13% validation accuracy. The loss curves converged smoothly, indicating that the model was learning steadily after 100 epochs. It achieved 94.69% accuracy in the test set, with weighted averages of 93.32% precision, 92.70% recall, and 92.66% F1-score. The ConvLSTM achieved 94.49% training accuracy and 90.28% validation accuracy over 100 epochs, with a validation loss of 0.854. While respectable, these figures lag behind the attention-based

LSTM model, particularly in validation, suggesting limitations in handling GSL's temporal nuances without explicit attention to weighting. Figure 7 illustrates the pre- and post-optimisation accuracy and loss curves of the attention-based LSTM, highlighting the impact of these interventions on convergence. Figures 8 and 9 illustrate the total training

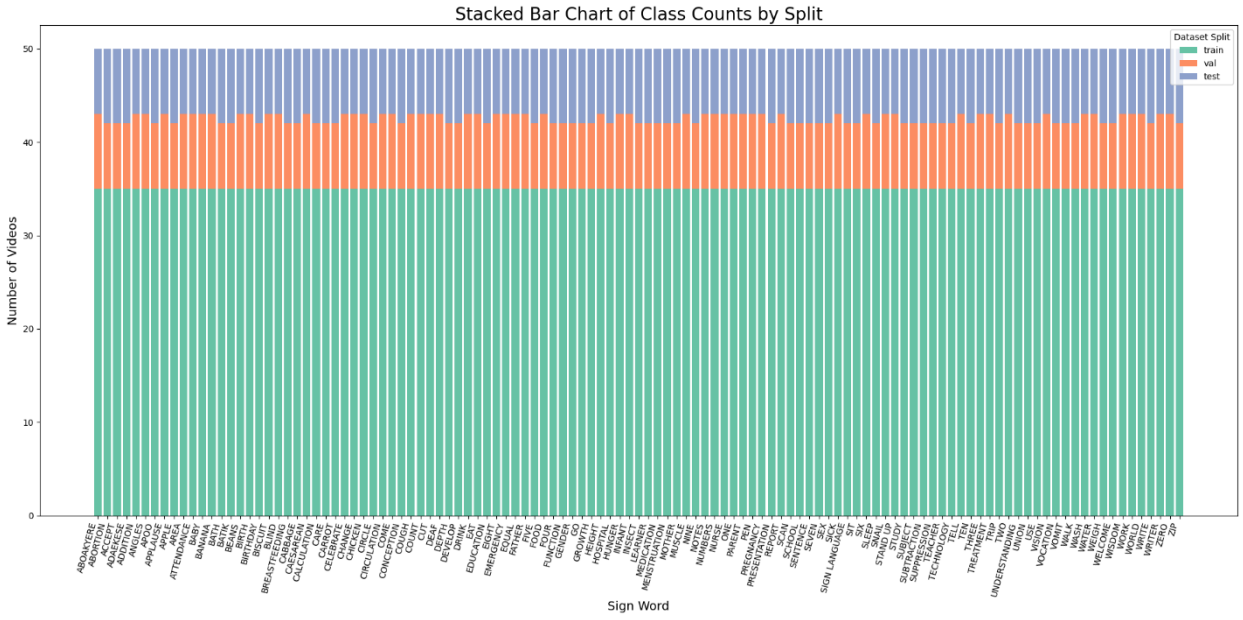

Figure 6: Illustrates the distribution of videos across the splits, confirming balance among classes and signers

loss 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. The stable and consistent lower validation loss compared to the training loss, between 1.0 and 0.5, suggests that the model performs well on unseen data. The evaluation metrics for this study were accuracy, precision, recall, and F1-score. The proposed attention model achieves 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 revealed exceptional performance on distinct signs, such as "hospital" (98% F1-score), which involves unique handshapes and poses, but lower accuracy (85-90%) on signs with overlapping trajectories, like "learn"

and "study," due to subtle differences in non-manual markers. Figure 10 is the confusion matrix of the attention-based LSTM across all classes, which shows that the model effectively recognizes signs through the log indication of the concise diagonal line, representing the potential for correct prediction from 0 to 8 based on the legend color scale.

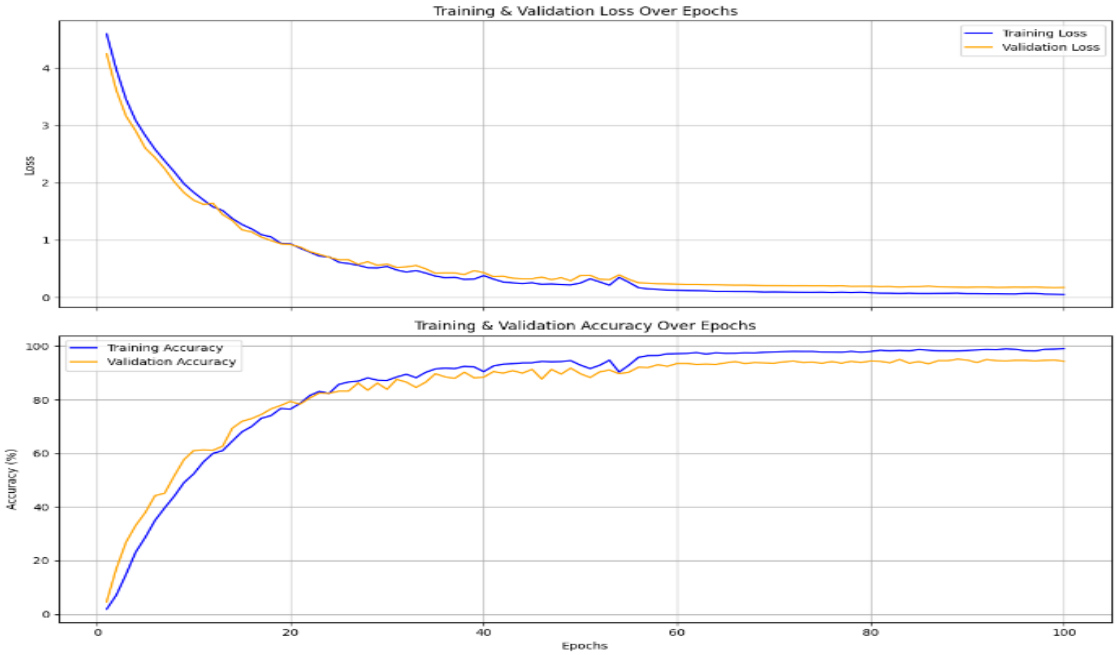

Figure 7: Training and Validation Accuracy and Loss Curves for the Proposed Attention-based LSTM Model

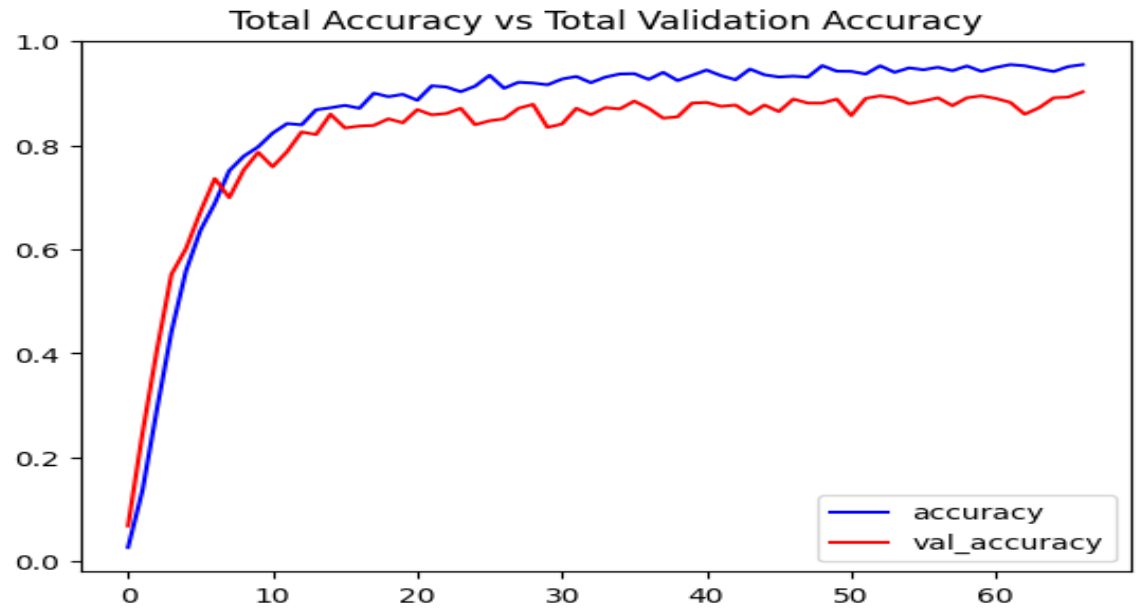

Figure 8: Training and Validation Accuracy Curves for the ConvLSTM Baseline Model

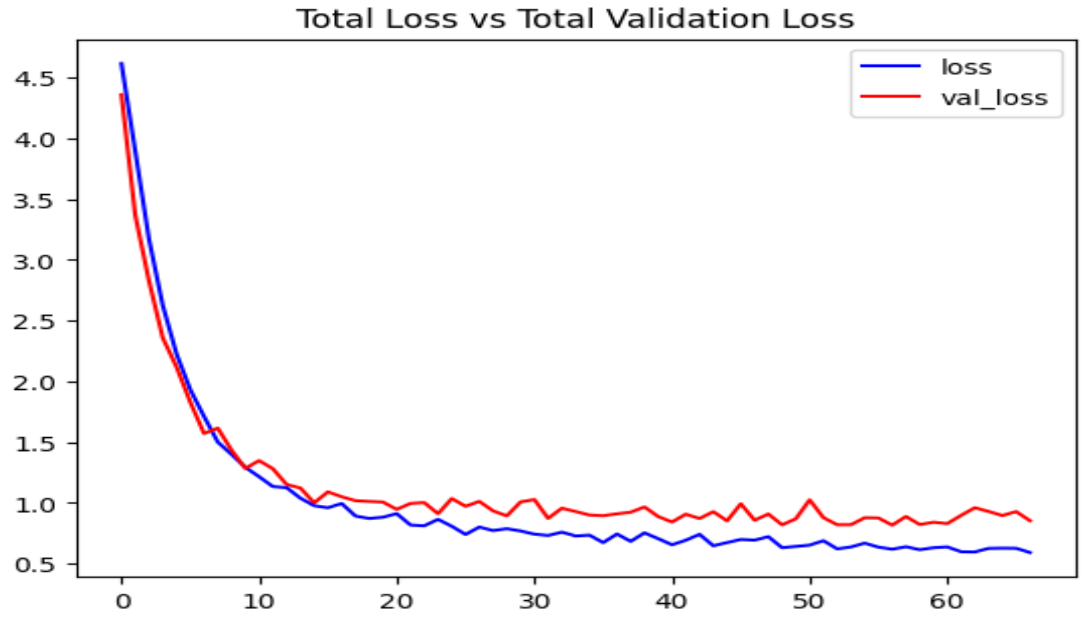

Figure 9: Loss and validation loss Curves for the ConvLSTM Baseline Model

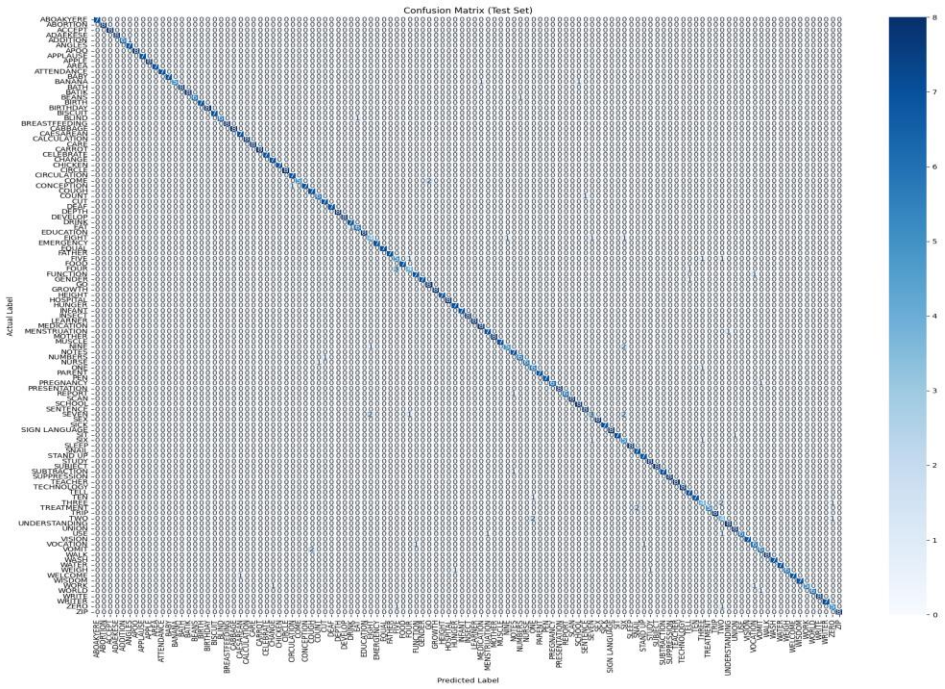

Figure 10: Confusion Matrix for Sign Language Recognition proposed model

5. Conclusions

This research achieved its goals by making the AkwaabaSign dataset available a multi-signer video collection that helps address the dire lack of a GSL dataset and by creating an optimized pipeline to process 2D keypoint coordinates. The study also substantially reduces 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 joint use of keypoint-based representation, normalization, and attention approaches significantly increases recognition accuracy, with 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 also marks the first attempt at using keypoint-based input for GSL recognition.

Looking to the future, the AkwaabaSign dataset can be expanded to include sentence-based expressions, interdisciplinary training programs in computational linguistics and sign language studies are recommended to build local capacity in low-context areas of Africa, such as Ghana, primarily focusing on transitioning the model to continuous GSL recognition, integrating transformer architectures alongside the LSTM-attention framework, signer-specific differentiation, comprehensive ablation studies and deploying the model into mobile devices for further analysis. Exploring multimodal systems that combine keypoints with audio cues for hearing interpreters or text outputs for broader accessibility could create versatile tools for bidirectional communication.

List of abbreviations

DHH Deaf and Hard-of-Hearing

SL Sign Language

SLR Sign Language Recognition

GSL Ghanaian Sign Language

CNN Convolutional Neural Network

RNNs: Recurrent Neural Networks

LSTM Long Short-Term Memory

ConvLSTM Convolutional Long Short-Term Memory

HMMs Hidden Markov Models

Author Contributions

Conceptualization, K. T.; methodology, K. T., L.T., and R-M. O. M. G.; software, K. T., 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.; visualisation, K. T., and L.T.; supervision, K. T., and R-M. O. M. G.; project administration, 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, can be accessed at https://zenodo.org/records/15730487

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. The details of participants would be kept secret. All participants were adults aged 18 and above.

Human Rights Statement

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

Consent for Publication

Participants were informed that the data collected would be used for this study and would only be made publicly available after approval. Consent has been confirmed for all individuals in the graphical abstract, consistent with our ethical clearance.

Conflict of Interest

The authors declare that they have 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 would like to thank the participants from Kwame Nkrumah University of Science and Technology who volunteered their time for data collection and testing of our findings.

AI-Declaration

The authors would like to declare that 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] "Deafness and hearing loss." Accessed: Jun. 07, 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," *Journal of Linguistic Geography*, vol. 12, no. 2, pp. 84–104, Oct. 2024, doi: 10.1017/jlg.2024.12.

[3] L. K. Odartey, Y. Huang, E. E. Asantewaa, and P. R. Agbedanu, "Ghanaian sign language recognition using deep learning," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2019, pp. 81–86. doi: 10.1145/3357777.3357784.

[4] I. Sandjaja, A. Alsharoa, D. Wunsch, and J. Liu, "Survey of Hidden Markov Models (HMMs) for Sign Language Recognition (SLR)," *Industrial Cyber-Physical Systems*, 2024, doi: 10.1109/ICPS59941.2024.10640040.

[5] R. Kumar, A. Sinha, A. Bajpai, and S. K. Singh, "A Comparative Analysis of Techniques and Algorithms for Recognising Sign Language," May 2023, Accessed: Aug. 02, 2025. [Online]. Available: https://arxiv.org/pdf/2305.13941

[6] R. Rastgoo, K. Kiani, and S. Escalera, "Sign Language Recognition: A Deep Survey," *Expert Syst. Appl.*, vol. 164, p. 113794, Feb. 2021, doi: 10.1016/J.ESWA.2020.113794.

[7] S. Subburaj and S. Murugavalli, "Survey on sign language recognition in context of vision-based and deep learning," *Measurement: Sensors*, vol. 23, p. 100385, Oct. 2022, doi: 10.1016/J.MEASEN.2022.100385.

[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," *Advanced Robotics*, vol. 38, no. 23, pp. 1649–1667, 2024, doi: 10.1080/01691864.2024.2442721/ASSET/8AA04971-BE42-43FF-B02C-4476203A4EF9/ASSETS/GRAPHIC/TADR_A_2442721_ILG0003.GIF.

[9] N. Adaloglou *et al.*, "A Comprehensive Study on Deep Learning-based Methods for Sign Language Recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1750–1762, Mar. 2021, doi: 10.1109/TMM.2021.3070438.

[10] M. Bohacek and M. Hruz, "Sign Pose-based Transformer for Word-level Sign Language Recognition," Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2022, pp. 182–191, 2022, doi: 10.1109/WACVW54805.2022.00024.

[11] M. Edward and / George Akanlig-Pare, "Sign language research in Ghana: An overview of indigenous and foreign-based sign languages," *JOURNAL OF AFRICAN LANGUAGES AND LITERATURES*, vol. 2, pp. 114–137, 2021, doi: 10.6092/jalalit.v2i2.8039.

[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," Journal of Imaging 2022, Vol. 8, vol. 8, no. 4, Apr. 2022, doi: 10.3390/JIMAGING8040098.

[14] M. A. Khan et al., "Human action recognition using fusion of multiview and deep features: an application to video surveillance," Multimedia Tools and Applications 2020 83:5, vol. 83, no. 5, pp. 14885–14911, Mar. 2020, doi: 10.1007/S11042-020-08806-9.

[15] W. Burger and M. J. Burge, "Scale-Invariant Feature Transform (SIFT)," pp. 709–763, 2022, doi: 10.1007/978-3-031-05744-1_25

[16] T. Raghuveera, R. Deepthi, R. Mangalashri, and R. Akshaya, "A depth-based Indian Sign Language recognition using Microsoft Kinect," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 45, no. 1, pp. 1–13, Dec. 2020, doi: 10.1007/S12046-019-1250-6/METRICS.

[17] J. Shin *et al.*, "Korean Sign Language Recognition Using Transformer-Based Deep Neural Network," *Applied Sciences (Switzerland)*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/APP13053029.

[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, doi: 10.1007/S11263-016-0957-7/METRICS

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *," 2018, Accessed: Aug. 12, 2025. [Online]. Available: https://youtu.be/pW6nZXeWlGM

[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 2024, Vol. 12, vol. 12, no. 8, Jul. 2024, doi: 10.3390/TECHNOLOGIES12080124.

[21] Y. Kim, H. Baek, and J. M. Corchado, "Preprocessing for keypoint-based sign language translation without glosses," Sensors, vol. 23, no. 6, p. 3231, Mar. 2023, doi: 10.3390/s23063231.

[22] F. Shafizadegan, A. R. Naghsh-Nilchi, and E. Shabaninia, "Multimodal vision-based human action recognition using deep learning: a review," Artificial Intelligence Review 2024 57:7, vol. 57, no. 7, pp. 178-, Jun. 2024, doi: 10.1007/S10462-024-10730-5.

[23] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 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 2021, Vol. 21, vol. 21, no. 4, pp. 1–22, Feb. 2021, doi: 10.3390/S21041120.

[25] A. Núñez-Marcos, O. Perez-de-Viñaspre, and G. Labaka, "A survey on Sign Language machine translation," Expert Syst. Appl., vol. 213, p. 118993, Mar. 2023, doi: 10.1016/J.ESWA.2022.118993.

[26] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous Sign Language Recognition With Correlation Network," 2023.

[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, p. e7653, Apr. 2023, doi: 10.1002/CPE.7653;JOURNAL:JOURNAL:10969128;REQUESTEDJOURNAL.

[28] Q. Wu, Q. Huang, and X. Li, "Multimodal human action recognition based on spatio-temporal action representation recognition model," Multimedia Tools and Applications 2022 82:11, vol. 82, no. 11, pp. 16409–16430, Nov. 2022, doi: 10.1007/S11042-022-14193-0.

[29] C.-F. Chen et al., "Deep analysis of CNN-based spatio-temporal representations for action recognition," in Proc. IEEE CVPR Workshops, 2020.

[30] B. Zhang, M. Müller, and R. Sennrich, "SLTUNET: A Simple Unified Model for Sign Language Translation," May 2023, [Online]. Available: http://arxiv.org/abs/2305.01778
[31] S. H. Noh, "Analysis of Gradient Vanishing of RNNs and Performance Comparison," *Information 2021, Vol. 12, Page 442*, vol. 12, no. 11, p. 442, Oct. 2021, doi: 10.3390/INFO12110442.
[32] B. Dabwan and M. Jadhav, "A CNN-LSTM Model for Arabic Sign Language Recognition," 2023, pp. 459–470. doi: 10.2991/978-94-6463-196-8_35.
[33] K. K. Podder *et al.*, "Signer-Independent Arabic Sign Language Recognition System Using Deep Learning Model," *Sensors 2023, Vol. 23, Page 7156*, vol. 23, no. 16, p. 7156, Aug. 2023, doi: 10.3390/S23167156.
[34] S. B. Abdullahi and K. Chamnongthai, "American Sign Language Words Recognition of Skeletal Videos Using Processed Video Driven Multi-Stacked Deep LSTM," Sensors 2022, Vol. 22, vol. 22, no. 4, Feb. 2022, doi: 10.3390/S22041406.