Electronic Letters on Computer Vision and Image Analysis 25(2):1-19, 2025

Arabic Sign Language Recognition with Deep Learning Models and Keypoint Landmarks

Farah Alshanik[1], Saif Ziad Aljunidi[2], and Ethar Qawasmeh[3,4]

[1] *Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan*
[2] *Department of Public Law, Yarmouk University, Irbid, Jordan*
[3] *Department of Computer Science and Engineering, Ohio State University, Ohio, United States*
[4] *Faculty of Information Technology and computer Science, Yarmouk University, Irbid, Jordan*

Received 13th of august, 2025; accepted 21th of december, 2025

Abstract

Communication is a fundamental aspect of human interaction, essential for expressing emotions and building relationships. While individuals with typical hearing rely on spoken language, the deaf and mute community communicates through visual gestures and facial expressions, commonly known as sign language. However, communication barriers persist between hearing and non-hearing individuals, especially in regions with limited assistive technologies. To address this gap, we developed a system designed for real time operation that converts Arabic sign gestures into textual output. Unlike most existing systems that are limited to individual alphabets or numbers, our model recognizes complete, meaningful words. It was trained on a curated subset of 112 Arabic sign language (ArSL) words extracted from the KArSL dataset. Using OpenCV and the MediaPipe Holistic framework, multimodal keypoints from the hands, face, and upper-body pose were extracted to form 225-dimensional feature vectors per frame. These features were used to train deep learning models: CNN, GRU, LSTM, and Bi-LSTM. Among these models, the Bi-LSTM model achieved the highest performance with a training accuracy of 99.89% and testing accuracy of 99.61%. These results are based on a signer-dependent evaluation conducted on a 112-word subset of the KArSL dataset, reflecting performance within the same group of signers. This finding emphasize the potential of MediaPipe-based landmark extraction combined with deep learning to support accessible communication for Arabic speaking deaf communities.

Key Words: Arabic Sign Language, KArSL dataset, MediaPipe, LSTM, GRU, Sign language recognition, Deep learning.

1 Introduction

Sign languages are natural ways for millions of deaf and hard-of-hearing people to communicate using hand gestures, movements, and facial expressions [1]. In recent years, there has been an increasing international focus on persons with disabilities. The deaf, in particular, are now seen as linguistic minorities rather than just

Correspondence to: fmalshanik@just.edu.jo

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2317
ELCVIA ISSN: 15108-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

a population with medical disabilities [2]. The CRPD aims to set up a worldwide language justice agenda by recognizing the sign language of deaf communities as a legitimate language, which may be considered a landmark policy change toward supporting their linguistic culture and achieving cultural justice in global diversity [3]. However, they generally face daily communication challenges since fewer qualified interpreters exist or conflicting information on technology are provided [4]. Many deaf communities find these services, educational opportunities, and employment largely inaccessible due to inadequate support, which constitutes a form of discrimination against persons with disabilities. To address these issues, modern techniques such as computer vision and machine learning have recently contributed to the build of different systems for sign language comprehension through analysis of visual information [5]. Tools like MediaPipe track hand movements in real time with accuracy, making gesture analysis more accurate by providing coordinates for up to 21 key points per frame [6]. Compared with image or depth-based techniques, MediaPipe Holistic provides a lightweight, cross-platform solution that extracts precise hand, face, and pose landmarks in real time [7]. Its reliance on coordinate-based keypoints rather than pixel intensities makes it robust to lighting variations, camera angle changes, and complex backgrounds, while significantly reducing computational cost [6, 8]. These advantages make it well-suited for large-scale Arabic Sign Language recognition tasks and for deployment on resource-constrained devices [9]. By combining these tools with powerful neural networks, an accuracy of over 99% has been achieved in recognizing individual letters in American Sign Language [10]. In fact, most recognition systems can recognize only single signs like letters or numbers and do not consider any flow of genuine sign language conversation [11]. Continuous sign recognition is extraordinarily complicated since in continuous sign recognition, signs get mixed in real-time. This means that models have to learn to recognize patterns over time and distinguish between the smooth transitions of gestures [12]. Special neural networks, such as LSTM and GRU, have shown strong potential in understanding these time based pattern. Research on Arabic sign languages is still in its early days, with a big gap in both available data and focused studies compared to American and other Western sign languages. Most work so far has focused on recognizing single signs, using letters or numbers. Very few studies examine continuous sign recognition, which requires monitoring sign movements over time. The absence of large, public datasets and standard benchmarks makes progress even more difficult. In this paper, we propose a system for recognizing words in Arabic sign language. We utilized a part of the KArSL dataset which includes 112 different dynamic word-level signs recorded in RGB video format [13]. To extract keypoints of sign gestures from the video stream, we employed MediaPipe, a free and open-source tool. The extracted keypoints were used to train and compare several deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Bidirectional Long Short-Term Memory (Bi-LSTM).

2 Related Work

The growing field of sign language recognition is maturing with the development in computer vision and deep learning. This field intends to ease communication for several millions of deaf and hard-of-hearing individuals [14] by recognizing sign language correctly and translating it into text or spoken language [15]. Early sign language recognition studies focused on single, static signs using theories of simple machine learning and features handcrafted by researchers [15, 16]. Recently, computer vision and deep learning models have been used to handle continuous, dynamic signing [17, 18]. The development of statistical methods for continuous sign language recognition with a large vocabulary by Koller et al. [19] marks an essential step towards natural communication. Accurately identifying hand movements is essential for Sign Language Recognition. Hand landmarks can be tracked in 2D or 3D using programs like MediaPipe Hands [6], OpenPose [20], and DeepHand [21], which can track hand landmarks in 2D or 3D in real time. Deep learning models, such as Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks, are often used for understanding the temporal dependencies in signs over time [22–24]. GRUs have demonstrated comparable accuracy with reduced computational power [25]. Subramanian et al. [4] combined GRUs with MediaPipe features and achieved faster and more accurate results compared to traditional LSTMs. Other models combine CNNs (for spatial feature) and RNNs

(for temporal feature) and perform well in continuous sign recognition [26–28]. Recently, the use of the Transformer architecture has shown promising results in handling long-range dependencies [14]. Continuous Sign Language Recognition (CSLR) is considered a more difficult task compared to isolated sign language recognition because the signs are variable in length, contain coarticulation, and have a larger vocabulary size [29, 30]. Some datasets have played a crucial role in promoting research in the domain of CSLR, including the RWTH-PHOENIX-Weather dataset for German Sign Language, the How2Sign dataset for American Sign Language, and the SIGNUM dataset [31–34]. However, most studies have focused on Western sign languages, which has resulted in the identification of a research gap in the domain of Arabic Sign Language (ArSL), as the availability of datasets, resources, and supporting tools is very low [35]. In the region of the Arab world, most studies have focused on the recognition of isolated gestures rather than continuous signing [36]. Research on Arabic Sign Language (ArSL) recognition has developed significantly in recent years, changing from just recognizing the individual letters of the alphabet to isolated words and, more recently, continuous signing. Early work focused on static alphabets, such as ArSL2018[37], which provides over 54,000 images of 32 letters and has become the most popular benchmark. Later versions, such as AArSLRS [38] addressed Quranic surahs with disjointed letters, while ArSL21L [39] introduced 14,202 annotated images, enabling modern detectors like YOLOv5l to achieve high performance. More recent works applied CNNs, ResNet, and transformers to further improve alphabet recognition [40–42]. Moving beyond alphabets, isolated-word datasets provide a richer linguistic coverage. The KArSL dataset [13] is still the largest resource with 502 words and 75,300 video samples. However, due to limited computational resources, many works limit training on smaller KArSL subsets (approximately 100 words) to evaluate the architectures using consistent three-signer data only. Recent studies also explored multimodality: [43] combined Leap Motion with deep learning (achieving 89% for one-hand, 96% for two-hand gestures), while [29] fused manual (hands) and non-manual (face) features using transformer-based models. Domain-specific corpora such as the Deaf Driver dataset [44] further expand isolated-word recognition into real-world communication settings. Continuous ArSL recognition is much more complex because of coarticulation and sentence-level variations. ArabSign [36] introduced a continuous large-scale dataset (9,335 sentences, 6 signers) for setting benchmarks on word error rate. Later works mainly focused on specific domain applications. For instance, [45] developed a multimodal dataset that connects 1,950 audio signals, 131 texts, and 262 ArSL videos in daily and Islamic communication, while [46] compared temporal models on 30 common sentences and showed the TCN model achieves about 99. 5% accuracy. Different resources like the Saudi Sign Language continuous dataset [47] strongly underline the existence of a large difference between signer-dependent and signer-independent (78%) performances, thus focusing on the generalization issue as the main problem.

Table 1 summarizes the main ArSL datasets and representative studies across alphabets, isolated words, and continuous signing.

Besides Arabic language datasets, recent AI technologies have enabled real-time sign detection. For example, Cerqun et al. [48] and Marquez et al. [49] developed systems based on LSTMs and other AI models that allow sign recognition to be conducted at high speeds and with high accuracy, which is a great benefit for deaf people. Despite progress, SLR still faces challenges, like recognizing signs from different people, combining hand shapes with facial expressions or body movements, dealing with messy backgrounds, or working on devices with limited power [50, 51]. Building on these insights, our study proposes a system specifically for Arabic Sign Language. It leverages MediaPipe for robust hand-tracking and deep learning models for sequential sign analysis. The research aims to address the underrepresentation of Arabic sign languages in SLR systems, advance inclusivity in communication technologies, and explicitly document signer dependence while outlining future directions toward signer-independent and cross-dataset validation.

3 Methodology

In this study, an isolated-sign recognition system in Arabic Sign Language (ArSL) was realized. We used a subset of 112 signs from the publicly available KArSL dataset, where each gesture is stored as RGB video

Table 1: Arabic Sign Language (ArSL) datasets and representative studies.

Category	Dataset / Paper (Year)	Size / Type	Key Notes
Static (Alphabets)	ArSL2018 (Latif et al., 2019) [37]	32 letters; 54k images; 40 signers	Widely used alphabet benchmark
	AArSLRS (Tharwat et al., 2021) [38]	9,240 images (alphabets); Quranic letters	First Quranic surahs with disjointed letters
	ArSL21L (Batnasan et al., 2022) [39]	14,202 images; 32 letters; 50 signers (23 female, 27 male)	The YOLOv5l model reported the best performance, achieving a COCO mAP score of 0.83
	Al-Barham et al. (2023) [40]	ArSL2018; 32 letters	CNN-based alphabet recognition
	Balat et al. (2024) [41]	ArSL2018 + AASL	$\sim$99.6% with transformers
	Alasmari et al. (2025) [42]	ArASL2018/ArASL2021	UrSL-CNN, ResNet/U-Net
Isolated Words	Alnahhas et al. (2020) [43]	440 videos; 10 words; 5 signers	Highest classification rate: 89% (one-hand), 96% (two-hand).
	KArSL (Sidig et al., 2021) [13]	502 words; 75,300 videos; 3 signers	Largest word-level dataset
	Deaf Driver Corpus (Abbas, 2021) [44]	860 videos; 215 words/phrases; 4 signers	Transportation domain (driver–passenger communication)
	Alyami et al. (2024) [52]	KArSL subset; $\sim$100 words	Transformer + landmarks, hands+face fusion
Continuous Sentences	ArabSign (Luqman et al., 2023) [36]	9,335 samples; 30 sentences; 6 signers	Large continuous dataset
	Algethami et al. (2025) [46]	3,046 videos; 30 sentences; 3 signers	TCN vs Bi-LSTM, $\sim$99.5%
	Abbas et al. (2024) [45]	Multimodal (audio, text, video); 1,950 audio signals; 131 texts; 262 ArSL videos; 2 signers	Domain-specific (daily + Islamic)
	Elhassen et al. (2025) [47]	5,810 samples; 24 signers	Shows generalization gap

frames. MediaPipe Holistic was used to extract hand, face, and body landmarks from each frame. The framework has four major phases: dataset description, data preprocessing, feature extraction, and sequence modeling (Figure 1). Each video is represented as a sequence of landmark vectors and used to train and evaluate deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Bidirectional Long Short-Term Memory (Bi-LSTM), to classify the gestures into their corresponding Arabic sign labels.

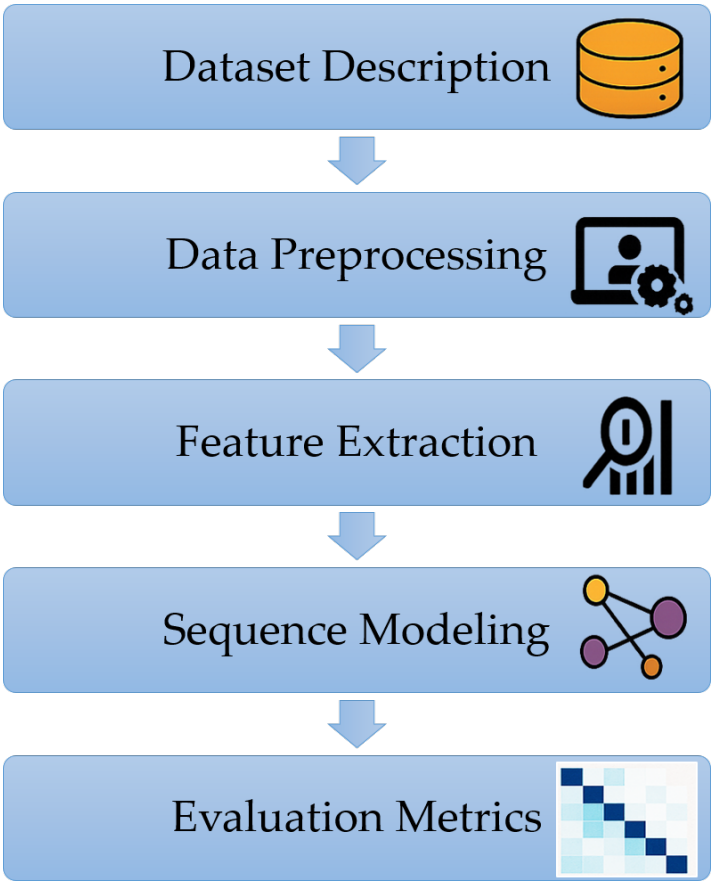


Figure 1: System architecture showing the overall pipeline: from dataset input and keypoint extraction using MediaPipe Holistic to deep learning-based classification

3.1 Preliminaries

1. MediaPipe: an open-source framework built by Google research, provides a cross-platform solution for building in-depth ML and computer vision learning applications. We leverage its Holistic module to estimate the real-time 3D landmarks output from the video frames, which include facial, hand, and full-body pose keypoints, as demonstrated in Figure 2. Such fine-grained spatial information is crucial for accurate gesture recognition [53].

2. Long Short-Term Memory (LSTM): LSTM is a particular kind of RNNs which can learn data sequences across time (such as videos or speech). It employs memory cells and three blocking mechanisms: input, forget, and output gates to manage the flow of information. These gates determine what to retain, what to discard, and what to pass along for the learned model to remember significant time-related sequence patterns, and forget irrelevant information [54]. This architecture helps solve the vanishing gradient

problem in most RNNs so that the LSTM network is able to store important information that is useful even several thousand time steps back.

3. Bidirectional Long Short-Term Memory (BiLSTM): BiLSTM adds two layers, one processing in a forward direction and the other backward, to a standard LSTM. It essentially learns input sequence in two opposite directions, from past to future and from future to past, so that it knows the context before and after a given event. The double view provides the network with the opportunity to learn more complex temporal patterns in continuous data, such as sign language gestures [55]. In gesture recognition, the BiLSTM excels because the meaning of a gesture depends on what happened before and after. For instance, BiLSTM improves accuracy a lot when interpreting sequences, coupled with spatial landmark detection systems (like those for hand and body keypoints) and temporal modeling [56].

4. Gated Recurrent Unit (GRU): The GRU represents the stripped-down version of an LSTM, with fewer components and faster speed. It contains only two gates, which are an update gate and a reset gate. This is unlike LSTM, which consists of three gates and a separate memory cell. Hence, there are fewer parameters during the learning process, consequently requiring less computation time for both training and running [57]. GRUs have been integrated with spatial feature extractors such as MediaPipe landmarks in sign language and gesture recognition systems. A GRU-based model for Indian sign languages, for instance, had the advantages of faster learning, higher accuracy, and better efficiency than other models [4].

5. Convolutional Neural Network (CNN): though CNNs are thought to be suitable for extracting spatial features from each video frame, in sign language, these systems take landmark vector inputs (like hand or body joint positions) for every frame to study different meaningful shapes and configurations. Observing these spatial relations (like finger positions or body posture), CNNs can help recognize the static form of each hand gesture [58]. If the CNN is concatenated with a temporal model such as LSTM or BiLSTM, the hybrid network is capable of learning both the spatial patterns of what a gesture looks like (through CNN) and sequence modeling of their evolution over time (through RNN), thereby making spatial–temporal fusion an effective sign language classification method, especially when a CNN LSTM with attention mechanism is employed [27].

3.2 Dataset Description

This study uses the publicly available KArSL dataset, a large-scale collection of Arabic sign language videos recorded using Microsoft Kinect V2 sensors. The dataset includes 502 distinct sign words, each performed by three male signers with 50 repetitions per sign, resulting in 75,300 video samples. The vocabulary covers a wide range of categories, including numbers, letters, health, verbs, family, and professions, and features both one-handed and two-handed gestures. All the samples provide color video recordings as well as depth information and skeletal joint data, thus giving a rich multimodal information for sign recognition research.

Within the scope of this paper, a group of 112 sign words labeled from 289 to 400 was chosen to create and test the recognition system, focusing on the most frequently used words in daily communication, to ensure the sign language representation was practical and effective. Two practical constraints motivated this choice: (i) the computational resources (GPU memory, as well as time for training) needed to train and optimize sequence models on the entire 502-class dataset are huge; and (ii) the subset serves as a conveniently manageable benchmark for model behavior analysis before proceeding to more complex cases. It is also worth mentioning that each sign is represented by a video sequence rather than a static image, and afterwards each instance consists of 30 frames with 225 keypoints on each frame. With 3 signers and 50 repetitions per sign, even a 100-word subset can produce about 15,000 video samples, or almost half a million frames, leading to the generation of millions of feature vectors that need to be processed. Previous research has similarly used KArSL subsets (100 words, 3 signers) to perform controlled evaluations, thus confirming that subset-based experiments are a widely accepted

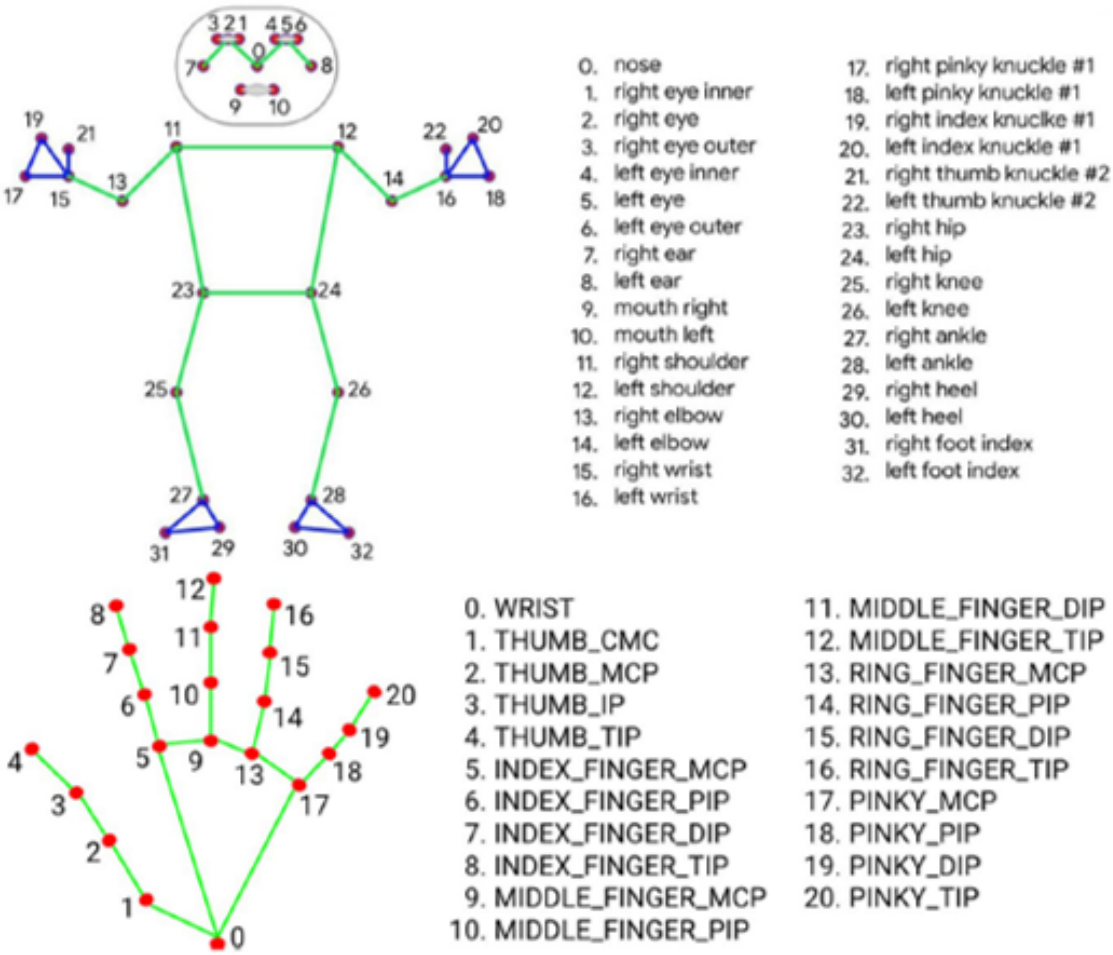


Figure 2: Example of pose, left-hand, and right-hand landmarks extracted by MediaPipe Holistic[53]

and practical approach in Arabic Sign Language research [52]. The videos that were chosen were segmented into frames, and from these frames, pose-based landmarks were extracted, which were then used for training the model and evaluation.

3.3 Data Preprocessing

The KArSL dataset provides sign language samples in the form of recorded videos, where each video represents a single sign. However, these videos vary in duration due to differences in signer speed and sign complexity, resulting in a non-uniform number of frames per video. Since deep learning models such as LSTM, GRU, and CNN require fixed-length input sequences, a standardization process was applied to ensure consistency across samples. The solution was to downscale the longer video or pad the shorter video to a length of 30 frames that indicates the average video length in the dataset. For the videos that are shorter than 30 frames, the last available frame was continuously repeated until the total number of frames reached 30, thereby imitating a pause at the end of the gesture. After the selection of frames, each frame got its 3D keypoints for pose, left hand, and right hand extracted by the MediaPipe model. Since MediaPipe works on RGB images, all the frames were converted from BGR (the default OpenCV format) to RGB. Each landmark is shown through (x, y, z) coordinates. The keypoints that were taken out were then made into vectors and normalized using reference anchor points in order to reduce inter-signer variation: the nose was the reference for the pose keypoints and the wrist joints for

each hand. Non-detected landmarks were set to zero in order to keep the dimensionality consistent.

3.4 Feature Extraction

In order to capture the visual and spatial structure of sign language gestures accurately, the MediaPipe Holistic framework, which offers a combination of pose and hand tracking into a single pipeline, was utilized in this research. As a part of the preprocessing, all videos were normalized to a fixed length of 30 frames, and the MediaPipe Holistic model was then applied to each of these frames to extract three key sets of 3D landmarks: pose, left hand, and right hand. Pose landmarks: the model extracts 33 upper body anatomical points like shoulders, elbows, and wrists. Each of them is represented by (x, y, z) coordinates, which results in a 99-dimensional vector for each frame (33 points × 3 coordinates). Left-hand landmarks: the 21 keypoints of the left hand, which include the palm and all five fingers, are also captured using the same coordinate structure, thus creating a 63-dimensional feature vector for each frame. Right-hand landmarks: in the same way, 21 right-hand landmarks are identified and turned into a 63-dimensional vector. Thus, each video frame is encoded as a 225-dimensional feature vector composed of pose (99), left hand (63), and right hand (63) keypoints. To increase spatial consistency between frames and among different signers, a normalization procedure was performed. It involved changing the coordinates with respect to reference points: the nose landmark was used as a central anchor for pose keypoints, while the wrist joints served as local anchors for each hand. This adjustment reduces inter-signer variation and helps the model generalize better. The extracted feature vectors for all frames were stored as NumPy arrays. These sequences, each consisting of 30 frames of 225-dimensional vectors, form the structured input to the deep learning models evaluated in this study.

3.5 Training

To evaluate the performance of different deep learning architectures on Arabic Sign Language recognition, four models were implemented and trained: LSTM, GRU, Bi-LSTM, and 1D CNN. All models were trained using sequences of 30 frames, where each frame consisted of a 225-dimensional feature vector (representing pose, left-hand, and right-hand keypoints). The dataset was split into 80% training and 20% testing, and training was conducted for 200 epochs using a batch size of 32. The Adam optimizer and categorical cross-entropy loss function were used consistently across all models. For this experiment, both the training and testing samples were drawn from the same group of signers, making the evaluation signer-dependent. However, we recognize that such an evaluation may slightly overestimate accuracy compared to signer-independent protocols. Future work will adopt leave-one-signer-out cross-validation to assess generalization across unseen signers. Dropout layers were used within the recurrent models to minimize overfitting, and because the training and testing accuracies were closely aligned (99.89% and 99.61%, respectively), the results demonstrated stable generalization, validating the reliability of the adopted split.

4 Results and Discussions

4.1 Model Training

We trained our model using keypoints extracted with MediaPipe Holistic, which provides 3D landmarks for the pose, left hand, and right hand. After preprocessing, each video was standardized to 30 frames, with every frame represented by a 225-dimensional feature vector (pose: 99, left hand: 63, right hand: 63). This results in 30×225 keypoints for each action sequence used in training. The lightweight nature of the extracted landmark features and the compact model architecture make the proposed system computationally efficient and suitable for real-time operation on standard hardware.

4.2 Evaluation Metrics

In order to assess the effectiveness of our sign language recognition system, four deep learning models were compared: CNN, LSTM, GRU, and Bi-LSTM. The evaluation was based on several standard metrics that included overall accuracy, confusion matrix, precision, recall, and F1-score for each class. The performance comparison across these models on the 112-class Arabic Sign Language recognition task is summarized in Table 2.

Table 2: Performance Comparison of Deep Learning Models on 112 Arabic Sign Language Words from the KArSL Dataset.

Algorithm	**Precision**	**Recall**	**F1-Score**	**Accuracy (%)**
LSTM	0.99	0.99	0.99	99.33
CNN	0.99	0.99	0.99	98.60
GRU	1.00	0.99	0.99	99.50
Bi-LSTM	1.00	1.00	1.00	99.61

Among the models, Bi-LSTM achieved the highest accuracy of 99.61%, closely followed by GRU (99.50%), LSTM (99.33%), and CNN (98.60%). Precision, recall, and F1-score values were consistently high across all classes and models, with Bi-LSTM reaching perfect scores (1.00) in most classes. The Bi-LSTM model achieved the best performance because it processes input sequences in both forward and backward directions, allowing it to capture long-range temporal dependencies and contextual information from the entire sequence. This bidirectional structure enables a deeper understanding of motion dynamics and coarticulation effects in sign gestures, which single-directional models like LSTM or GRU often miss. Similar findings have been reported in previous research, where Bi-LSTM consistently outperformed CNN, LSTM, and GRU models in continuous sign-language and action-recognition tasks [59].

The confusion matrix shown in Figure 3 highlights the model's performance across 20 random sign classes. It demonstrates near-perfect classification performance, with only minor misclassifications. For example, the "Mattress" class showed a small drop in recall (0.94), indicating a few incorrect predictions, but overall, most classes were classified with perfect precision and recall.

The individual class precision, recall, and F1-Score values are calculated and shown in Table 3.

Since Bi-LSTM achieved the highest accuracy of 99.61%, it was selected for further analysis of learning behavior across epochs.Figure 4 illustrates the training and validation accuracy, Figure 5 presents the corresponding training and validation loss, and table 4 shows the Bi-LSTM model's learning progress over epochs. As shown, the categorical accuracy steadily increased, reaching 99.89% at 200 epochs, while the loss decreased to a minimal 0.004, confirming effective training. These results highlight the strength of sequence models, particularly Bi-LSTM, in capturing temporal dependencies in sign language sequences and accurately recognizing isolated signs from pose-based landmarks. Overall, the models achieved strong performance, with the highest training and testing accuracies reaching 99.89% and 99.61%, respectively, demonstrating the effectiveness of the approach for isolated Arabic sign word recognition.

While the Bi-LSTM achieved 99.61% accuracy within our signer-dependent split, we note that such settings can overestimate performance due to signer overlap. Learning curves figure 4 suggest stable optimization, yet signer-independent evaluation (e.g., leave-one-signer-out) typically yields more conservative estimates. We therefore treat the present results as within-subset, upper-bound indicators pending signer-independent and cross-dataset validation.

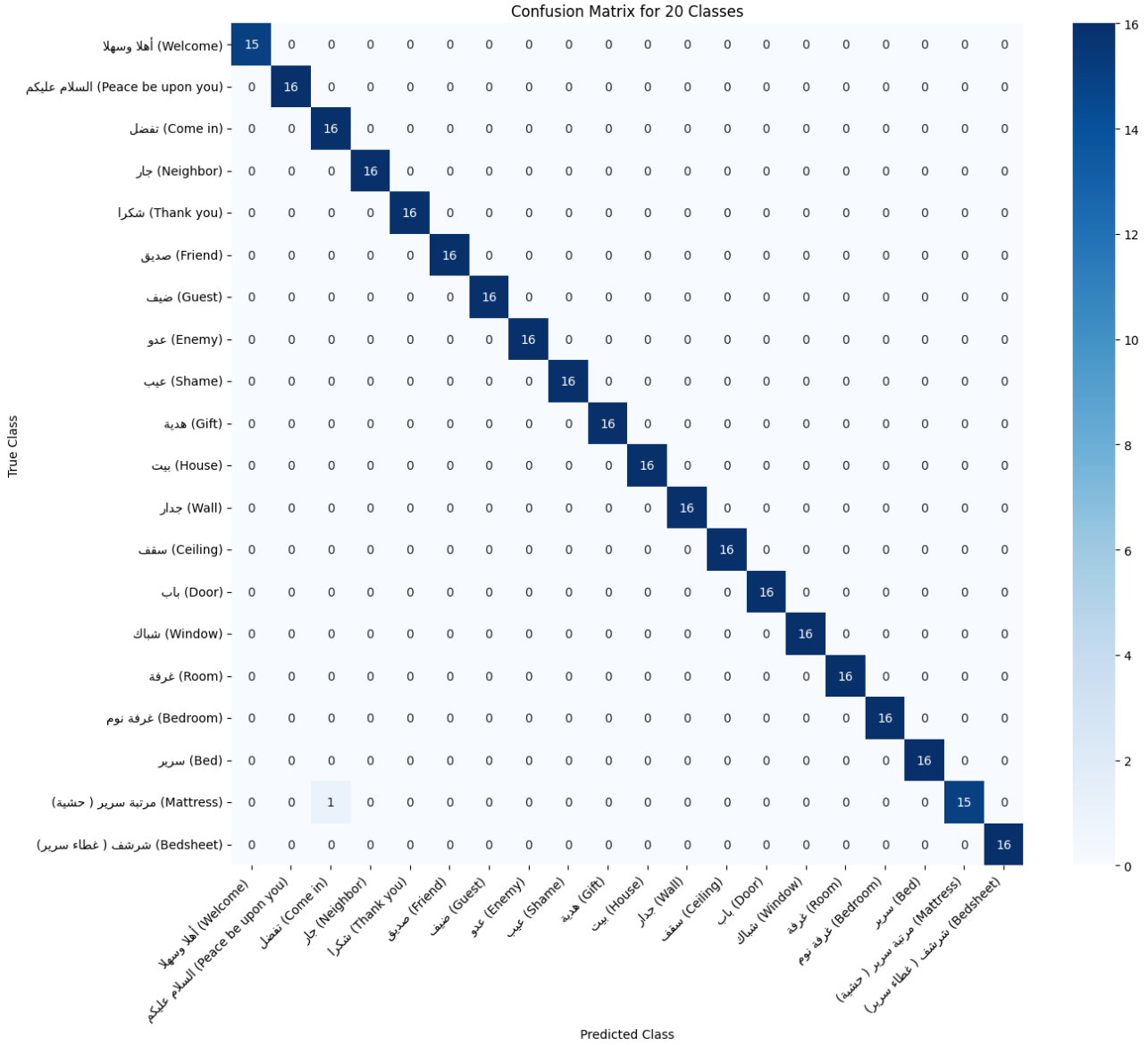


Figure 3: Confusion matrix of the Bi-LSTM model for 20 randomly selected sign classes (subset of the total 112).

Table 3: Precision, Recall, and F1-Score for 20 Randomly Selected Sign Classes (Subset of 112).

Label Class (English)	**Precision**	**Recall**	**F1-Score**
Hello	1.00	0.94	0.97
Peace be upon you	1.00	1.00	1.00
Come in	0.94	1.00	0.97
Neighbor	1.00	1.00	1.00
Thank you	1.00	1.00	1.00
Friend	1.00	1.00	1.00
Guest	1.00	1.00	1.00
Enemy	1.00	1.00	1.00
Shame	1.00	1.00	1.00
Gift	1.00	1.00	1.00
House	1.00	1.00	1.00
Wall	1.00	1.00	1.00
Ceiling	1.00	1.00	1.00
Door	1.00	1.00	1.00
Window	1.00	1.00	1.00
Room	1.00	1.00	1.00
Bedroom	1.00	1.00	1.00
Bed	1.00	1.00	1.00
Mattress	1.00	0.94	0.97
Bedsheet	1.00	1.00	1.00

Table 4: Comparison of evaluation metrics across different training epochs.

Epochs	**Loss**	**Categorical Accuracy (%)**
50	0.0444	0.9875
100	0.0054	0.9982
150	0.0161	0.9957
200	0.0040	0.9989

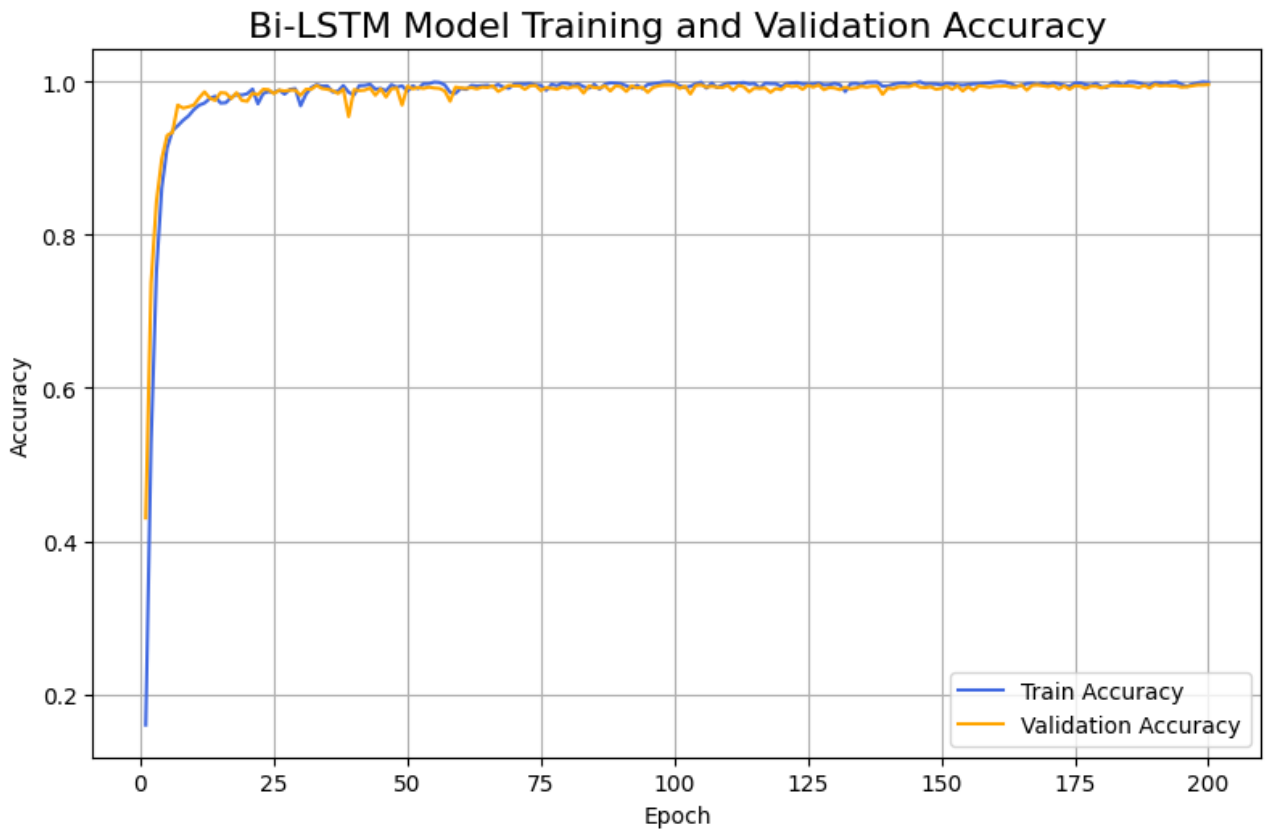


Figure 4: Training and validation accuracy curves of the Bi-LSTM model over 200 epochs

5 Discussion

The findings of this research demonstrate that temporal deep learning models combined with MediaPipe-based landmark extraction can effectively be used for Arabic Sign Language (ArSL) recognition. Our Bi-LSTM model achieved the highest performance, with 99.61% testing accuracy and perfect precision, recall, and F1-score on most gesture classes. These results align with previous studies that emphasize the advantage of bidirectional architectures in capturing temporal dependencies in gesture data [55, 56].

These high accuracy values are consistent with previous sign language recognition studies that employed signer-dependent evaluation protocols. For example, [4] achieved 99.2 % accuracy for Indian Sign Language using a GRU-based model with MediaPipe keypoints, [60] reported over 99 % accuracy on a dynamic sign-language dataset using MediaPipe landmarks with RNNs. In Arabic Sign Language, [52] achieved approximately 99.7 % accuracy on a subset of the KArSL dataset using transformer-based fusion of landmark features. Our result of 99.61 % thus aligns well with existing benchmarks for word-level recognition under controlled signer-dependent conditions.

The strong performance of GRU (99.50%) further supports earlier claims that GRUs provide a good trade-off between accuracy and computational efficiency [25]. The accuracy improvements over CNN-only approaches (98.60%) validate the significance of temporal modeling over static feature extraction, confirming the need for sequence-aware architectures in sign language recognition [22, 26]. However, it is important to note that our current evaluation was conducted under a signer-dependent split, where the same signers appear in both training and testing sets. This setting can lead to optimistic performance estimates, as models may partially

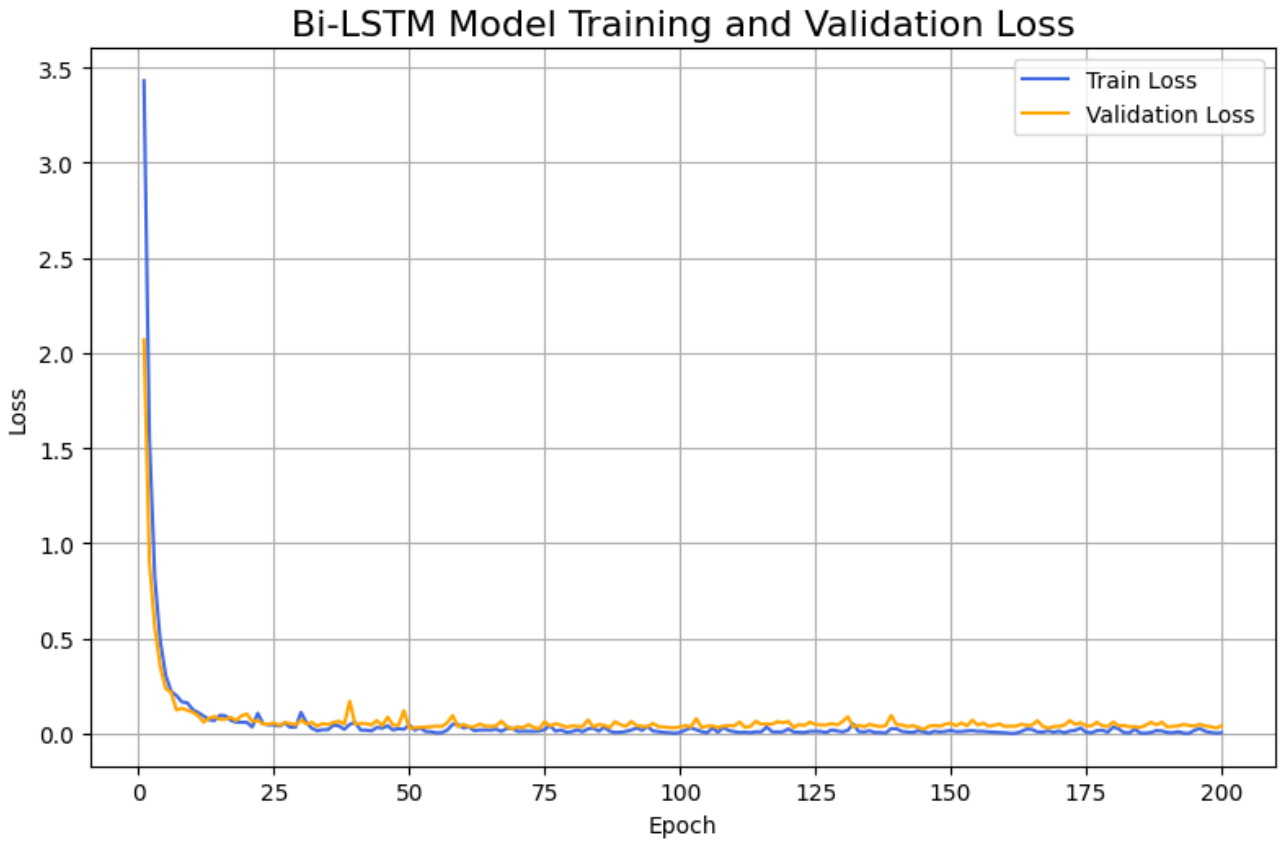


Figure 5: Training and validation loss curves of the Bi-LSTM model over 200 epochs.

learn signer-specific motion patterns in addition to intrinsic sign-language features.

Our work builds upon and extends the approaches of Subramanian et al. [4] and Verma et al. [10], who used MediaPipe features with GRUs and CNNs, respectively, to achieve high accuracy in Indian and American Sign Languages. Unlike prior Arabic-based systems that primarily focused on alphabets or isolated signs [35, 36], our system was trained on 112 commonly used ArSL words, contributing a significant step toward practical, word-level sign recognition in Arabic.

The implications of these results are important for developing accessible technologies in underrepresented languages. In the broader context, this work addresses the lack of robust systems for ArSL and supports inclusive communication tools for Arabic-speaking deaf communities. Future research may explore continuous sign recognition, signer-independent systems, and real-time mobile deployment to improve accessibility in diverse settings.

5.1 Limitations and Future Work

This study has several limitations that provide directions for future research. First, our use of a 112-word subset from the KArSL dataset and reliance on only three signers restrict the generalizability of our results. While subset-based evaluations are common in ArSL research due to their feasibility and have been employed in earlier KArSL studies, future work will aim to scale toward the full 502-word dataset (75,300 samples). Another important limitation concerns the evaluation protocol. The current results were obtained under signer-dependent splits, which may inflate accuracy due to signer overlap between training and testing sets. In our

future work, we therefore plan to adopt leave-one-signer-out cross validation and to report macro-averaged metrics under signer independence in order to better capture real world generalization. In addition, future work will explore cross-dataset evaluation, particularly with ArabSign, a large-scale continuous ArSL corpus [36]. This would allow us to investigate whether models trained on isolated word signs from KArSL can transfer to continuous signing tasks. Such validation will focus on overlapping vocabulary between the datasets, since cross-dataset comparisons require a shared set of sign classes. For non-overlapping cases, transfer learning and fine-tuning strategies will be necessary to bridge the gap between isolated-word and continuous-sentence recognition. Finally, we plan to extend our models toward practical deployment scenarios, including on-device and mobile applications. By integrating efficient sequence architectures and model quantization techniques, we aim to preserve accuracy while enabling real-time operation in unconstrained environments, where background noise, occlusion, and device limitations become significant factors.

6 Conclusion

This paper introduces a deep learning-based framework to recognize isolated words in Arabic Sign Language (ArSL) by utilizing pose and hand keypoints extracted through the MediaPipe Holistic model. By selecting 112 frequent signs from the KArSL dataset and employing advanced temporal sequence models such as LSTM, GRU, and Bi-LSTM, the proposed system achieved high accuracy in word-level gesture recognition. Among all, the Bi-LSTM model was the most successful with a training accuracy of 99. 89%, a testing accuracy of 99. 61%, and the significant ability to capture the temporal relationships that exist in sequential sign data. The system's reliability is also significantly demonstrated by the stability of the precision, recall, and F1-score measures across all gesture classes. This work also shows that spatial features based on landmarks, when combined with temporal deep learning architectures, can be very effective in Arabic sign recognition tasks. Besides, the proposed approach adds to the literature on ArSL recognition, an area significantly less represented compared to Western ones. Future work will explore the expansion toward continuous sign language recognition, signer-independent models, and mobile deployment to promote real-world accessibility and interaction for the Arabic-speaking deaf and hard-of-hearing communities.

7 Acknowledgment

The authors would like to acknowledge the deanship of research at the Jordan University of Science and Technology for supporting this research (grant 552/2025).

References

[1] Diane Lillo-Martin and Jonathan Henner. Acquisition of sign languages. *Annual review of linguistics*, 7 (1):395–419, 2021. doi:10.1146/annurev-linguistics-043020-092357.

[2] Michael Higgins and Amy M Lieberman. Deaf students as a linguistic and cultural minority: Shifting perspectives and implications for teaching and learning. *Journal of Education*, 196(1):9–18, 2016. doi:10.1177/002205741619600103.

[3] World Federation of the Deaf. Crpd & deaf rights. `https://wfdeaf.org/crpd/`, 2025. World Federation of the Deaf – Convention on the Rights of Persons with Disabilities (CRPD).

[4] Barathi Subramanian, Bekhzod Olimov, Shraddha M Naik, Sangchul Kim, Kil-Houm Park, and Jeonghong Kim. An integrated mediapipe-optimized gru model for indian sign language recognition. *Scientific Reports*, 12(1):11964, 2022. doi:10.1038/s41598-022-15998-7.

[5] Taewan Kim and Bongjae Kim. Techniques for detecting the start and end points of sign language utterances to enhance recognition performance in mobile environments. *Applied Sciences*, 14(20):9199, 2024. doi:https://doi.org/10.3390/app14209199.

[6] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. doi:https://doi.org/10.48550/arXiv.2006.10214.

[7] Google Research. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. `https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html`, December 2020. (Accessed: October 22, 2025).

[8] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020. doi:https://doi.org/10.48550/arXiv.2006.10204.

[9] Kanchon Kanti Podder, Maymouna Ezeddin, Muhammad EH Chowdhury, Md Shaheenur Islam Sumon, Anas M Tahir, Mohamed Arselene Ayari, Proma Dutta, Amith Khandakar, Zaid Bin Mahbub, and Muhammad Abdul Kadir. Signer-independent arabic sign language recognition system using deep learning model. *Sensors*, 23(16):7156, 2023. doi:https://doi.org/10.3390/s23167156.

[10] Aditya Raj Verma, Gagandeep Singh, Karnim Meghwal, Banawath Ramji, and Praveen Kumar Dadheech. Enhancing sign language detection through mediapipe and convolutional neural networks (cnn). *arXiv preprint arXiv:2406.03729*, 2024. doi:https://doi.org/10.48550/arXiv.2406.03729.

[11] Bader Alsharif, Easa Alalwany, Ali Ibrahim, Imad Mahgoub, and Mohammad Ilyas. Real-time american sign language interpretation using deep learning and keypoint tracking. *Sensors*, 25(7):2138, 2025. doi:https://doi.org/10.3390/s25072138.

[12] Din Ezra, Shai Mastitz, and Irina Rabaev. Signsability: Enhancing communication through a sign language app. *Software*, 3(3):368–379, 2024. doi:https://doi.org/10.3390/software3030019.

[13] Ala Addin I Sidig, Hamzah Luqman, Sabri Mahmoud, and Mohamed Mohandes. Karsl: Arabic sign language database. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–19, 2021. doi:https://doi.org/10.1145/3423420.

[14] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. doi:https://doi.org/10.48550/arXiv.2003.13830.

[15] Sylvie CW Ong and Surendra Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(06):873–891, 2005. doi:10.1109/TPAMI.2005.112.

[16] Mohammed Waleed Kadous et al. Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language. In *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, volume 165, pages 165–174. DE Wilmington, 1996.

[17] Brandon Garcia, Sigberto Alarcon Viesca, et al. Real-time american sign language recognition with convolutional neural networks. *Convolutional Neural Networks for Visual Recognition*, 2(225-232):8, 2016.

[18] Prabhakara Uyyala. Sign language recognition using convolutional neural networks. *J. Interdiscip. Cycle Res*, 14:1198–1207, 2022. doi:https://doi.org/10.17613/47ga-zw60.

[19] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. doi:https://doi.org/10.1016/j.cviu.2015.09.013.

[20] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. doi:10.1109/TPAMI.2019.2929257.

[21] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. doi:https://doi.org/10.48550/arXiv.1502.06807.

[22] Pranav Sheth, Sanju Rajora, and Yogeshwari Makwana. Sign language recognition application using lstm and gru (rnn). *Int. J. Sci. Res. Comput. Sci.*, 8(2), 2023. doi:10.13140/RG.2.2.18635.87846.

[23] Kenneth Mejía-Peréz, Diana-Margarita Córdova-Esparza, Juan Terven, Ana-Marcela Herrera-Navarro, Teresa García-Ramírez, and Alfonso Ramírez-Pedraza. Automatic recognition of mexican sign language using a depth camera and recurrent neural networks. *Applied Sciences*, 12(11):5523, 2022. doi:https://doi.org/10.3390/app12115523.

[24] Suguna Mariappan, Ponmalar Murugesan, and Hemapriya Muthamil Selvan. Real-time interpreter for short sentences in indian sign language using mediapipe and deep learning. *Information Technology and Control*, 53(3):888–898, 2024. doi:https://doi.org/10.5755/j01.itc.53.3.33935.

[25] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. doi:https://doi.org/10.48550/arXiv.1412.3555.

[26] Su Yang and Qing Zhu. Continuous chinese sign language recognition with cnn-lstm. In *Ninth international conference on digital image processing (ICDIP 2017)*, volume 10420, pages 83–89. SPIE, 2017. doi:https://doi.org/10.1117/12.2281671.

[27] Diksha Kumari and Radhey Shyam Anand. Isolated video-based sign language recognition using a hybrid cnn-lstm framework based on attention mechanism. *Electronics*, 13(7):1229, 2024. doi:https://doi.org/10.3390/electronics13071229.

[28] Qidan Zhu, Jing Li, Fei Yuan, and Quan Gan. Multiscale temporal network for continuous sign language recognition. *Journal of Electronic Imaging*, 33(2):023059–023059, 2024. doi:https://doi.org/10.1117/1.JEI.33.2.023059.

[29] Sarah Alyami and Hamzah Luqman. A comparative study of continuous sign language recognition techniques. *arXiv preprint arXiv:2406.12369*, 2024. doi:https://doi.org/10.48550/arXiv.2406.12369.

[30] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019. doi:10.1109/TMM.2018.2889563.

[31] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference 2016*, 2016. doi:https://doi.org/10.5244/C.30.136.

[32] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744, 2021. doi:https://doi.org/10.48550/arXiv.2008.08143.

[33] Ulrich von Agris and Karl-Friedrich Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, 2010. URL https://www.sign-lang.uni-hamburg.de/lrec/pub/10006.pdf.

[34] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, 2012.

[35] Ahmad Sami Al-Shamayleh, Rodina Ahmad, Nazean Jomhari, and Mohammad AM Abushariah. Automatic arabic sign language recognition: A review, taxonomy, open challenges, research roadmap and future directions. *Malaysian Journal of Computer Science*, 33(4):306–343, 2020. doi:https://doi.org/10.22452/mjcs.vol33no4.5.

[36] Hamzah Luqman. Arabsign: A multi-modality dataset and benchmark for continuous arabic sign language recognition. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. doi:10.1109/FG57933.2023.10042720.

[37] Ghazanfar Latif, Nazeeruddin Mohammad, Jaafar Alghazo, Roaa AlKhalaf, and Rawan AlKhalaf. Arasl: Arabic alphabets sign language dataset. *Data in brief*, 23:103777, 2019. doi:https://doi.org/10.1016/j.dib.2019.103777.

[38] Gamal Tharwat, Abdelmoty M Ahmed, and Belgacem Bouallegue. Arabic sign language recognition system for alphabets using machine learning techniques. *Journal of Electrical and Computer Engineering*, 2021(1):2995851, 2021. doi:https://doi.org/10.1155/2021/2995851.

[39] Ganzorig Batnasan, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Fady Alnajjar, and Timothy K Shih. Arsl21l: Arabic sign language letter dataset benchmarking and an educational avatar for metaverse applications. In *2022 IEEE global engineering education conference (educon)*, pages 1814–1821. IEEE, 2022. doi:10.1109/EDUCON52537.2022.9766497.

[40] Muhammad Al-Barham, Ahmad Jamal, and Musa Al-Yaman. Design of arabic sign language recognition model. *arXiv preprint arXiv:2301.02693*, 2023. doi:https://doi.org/10.48550/arXiv.2301.02693.

[41] Mazen Balat, Rewaa Awaad, Hend Adel, Ahmed B Zaky, and Salah A Aly. Advanced arabic alphabet sign language recognition using transfer learning and transformer models. In *2024 International Conference on Computer and Applications (ICCA)*, pages 1–6. IEEE, 2024. doi:10.1109/ICCA62237.2024.10927914.

[42] Naif Alasmari and Sultan Asiri. Asldetect: Arabic sign language detection using resnet and u-net like component. *Scientific Reports*, 15(1):18012, 2025. doi:https://doi.org/10.1038/s41598-025-01588-w.

[43] Ammar Alnahhas, Bassel Alkhatib, Nazeer Al-Boukaee, Noor Alhakim, Ola Alzabibi, and Noor Ajalyakeen. Enhancing the recognition of arabic sign language by using deep learning and leap motion controller. *Int. J. Sci. Technol. Res*, 9(4): 1865–1870, 2020. URL https://www.ijstr.org/final-print/apr2020/Enhancing-The-Recognition-Of-Arabic-Sign-Language-By-Using-Deep-Learning-And-I pdf.

[44] Samah Abbas, Hassanin Al-Barhamtoshy, and Fahad Alotaibi. Towards an arabic sign language (arsl) corpus for deaf drivers. *PeerJ Computer Science*, 7:e741, 2021. doi:https://doi.org/10.7717/peerj-cs.741.

[45] Samah Abbas, Dimah Alahmadi, and Hassanin Al-Barhamtoshy. Establishing a multimodal dataset for arabic sign language (arsl) production. *Journal of King Saud University-Computer and Information Sciences*, 36(8):102165, 2024. doi:https://doi.org/10.1016/j.jksuci.2024.102165.

[46] Nahlah Algethami, Raghad Farhud, Manal Alghamdi, Huda Almutairi, Maha Sorani, and Noura Aleisa. Continuous arabic sign language recognition models. *Sensors*, 25(9):2916, 2025. doi:https://doi.org/10.3390/s25092916.

[47] Soukeina Elhassen, Lama Al Khuzayem, Areej Alhothali, Ohoud Alzamzami, and Nahed Alowaidi. Continuous saudi sign language recognition: A vision transformer approach. *arXiv preprint arXiv:2509.03467*, 2025. doi:https://doi.org/10.48550/arXiv.2509.03467.

[48] Angel Diego Briones Cerquín, Johan Alonso Tumay Guevara, and Christian Ovalle. Mobile application for continuous recognition and classification of sign language images through deep learning. *International Journal of Interactive Mobile Technologies*, 19(7), 2025. doi:10.3991/ijim.v19i07.52853.

[49] Bogart Yail Marquez, Arnulfo Alanis, Angeles Quezada, José Sergio, et al. Development of a mobile application with artificial intelligence for mexican sign language recognition. *International Journal of Interactive Mobile Technologies*, 19(9), 2025. doi:10.3991/ijim.v19i09.54205.

[50] Ali Akdag and Omer Kaan Baykan. Enhancing signer-independent recognition of isolated sign language through advanced deep learning techniques and feature fusion. *Electronics*, 13(7):1188, 2024. doi:https://doi.org/10.3390/electronics13071188.

[51] Nada B Ibrahim, Hala H Zayed, and Mazen M Selim. Advances, challenges and opportunities in continuous sign language recognition. *Journal of Engineering and Applied Sciences*, 15(5):1205–1227, 2020. doi:10.36478/jeasci.2020.1205.1227.

[52] Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–19, 2024. doi:https://doi.org/10.1145/3584984.

[53] Google AI Edge. Mediapipe solutions guide. `https://ai.google.dev/edge/mediapipe/solutions/guide`, 2025. MediaPipe Solutions provides cross-platform ML tools and pre-trained task libraries.

[54] Alejandro Toro-Ossaba, Juan Jaramillo-Tigreros, Juan C Tejada, Alejandro Peña, Alexandro López-González, and Rui Alexandre Castanho. Lstm recurrent neural network for hand gesture recognition using emg signals. *Applied Sciences*, 12(19):9700, 2022. doi:https://doi.org/10.3390/app12199700.

[55] Kok Seang Tan, Kian Ming Lim, Chin Poo Lee, and Lee Chung Kwek. Bidirectional long short-term memory with temporal dense sampling for human action recognition. *Expert Systems with Applications*, 210:118484, 2022. doi:https://doi.org/10.1016/j.eswa.2022.118484.

[56] Yang Song, Mengru Liu, Feilu Wang, Jinggen Zhu, Anyang Hu, and Niuping Sun. Gesture recognition based on a convolutional neural network–bidirectional long short-term memory network for a wearable wrist sensor with multi-walled carbon nanotube/cotton fabric material. *Micromachines*, 15(2):185, 2024. doi:https://doi.org/10.3390/mi15020185.

[57] Yanqiong Zhang and Xianwei Jiang. Recent advances on deep learning for sign language recognition. *Computer Modeling in Engineering & Sciences (CMES)*, 139(3), 2024. doi:10.32604/cmes.2023.045731.

[58] Katerina Papadimitriou, Gerasimos Potamianos, Galini Sapountzaki, Theodoros Goulas, Eleni Efthimiou, Stavroula-Evita Fotinea, and Petros Maragos. Greek sign language recognition for an education platform. *Universal Access in the Information Society*, 24(1):51–68, 2025. doi:https://doi.org/10.1007/s10209-023-01017-7.

[59] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4297–4305, 2017. doi:10.1109/CVPR.2017.364.

[60] Gerges H Samaan, Abanoub R Wadie, Abanoub K Attia, Abanoub M Asaad, Andrew E Kamel, Salwa O Slim, Mohamed S Abdallah, and Young-Im Cho. Mediapipe's landmarks with rnn for dynamic sign language recognition. *Electronics*, 11(19):3228, 2022. doi:https://doi.org/10.3390/electronics11193228.