Dynamic Arabic Sign Language recognition using deep learning

(R.S.Abdul Ameer) rafalsaleh@tu.edu.iq

(M.A.Ahmed) mohamed.aktham@tu.edu.iq

University of Tikrit, College of Computer Science and Mathematics, Department of Computer Science

Abstract:

Sign language is a crucial means of communication for people with hearing impairments. However, some sign language interpreters are scarce. Due to this deficit, a considerable section of the hearing-impaired population is denied access to services, particularly in public settings. This research is dedicated to reducing the accessibility gap by using technology to develop systems capable of recognizing Arabic Sign Language (ArSL) using deep learning methods. Our model is designed to capture the unique features of sign language (words). It employs a Gated Recurrent Units (GRU) classifier to extract spatial and temporal properties for sequential data. To demonstrate the effectiveness of our approach, we created a dataset of 15 distinct phrases, resulting in 450 videos for 15 dynamic gesture words in ArSL. Our model performs impressively, with MediaPipe and GRU classifiers achieving an accuracy rate of 97%. These results highlight the potential of our strategy to significantly enhance communication accessibility for the hearing-impaired population. This research represents a significant step towards promoting inclusivity and, more importantly, improving the quality of life for the hearing impaired, offering a ray of hope for a better future.

I. Introduction

Sign language is not just a mode of communication for the hearing impaired; it is their lifeline. It is a crucial part of their community, yet it is often misunderstood. This is where the significance of research into automated sign language recognition algorithms becomes apparent. These algorithms have the potential to revolutionize the way hearing-impaired individuals

communicate, significantly improving their quality of life. Sign language conveys semantic information through hand forms, motion trajectory, facial expressions, lip motions, eye contact, and more. It often involves one or more gestures, motions, and transitions between them. Even a small change in one of these components can drastically alter the meaning, making sign language a complex yet fascinating mode of communication [1].

Gestures, the most common form of symbolic communication, are a natural and efficient way for people to express themselves. They span from basic to complex actions, enabling us to interact with others. With significant advancements in deep learning and computer vision technology, the focus has shifted to leveraging human biological traits to revolutionize human-computer interaction. The hand, being the most versatile part of the human body, can represent a wide range of human-machine communication. Hand gestures are not just a means of communication between people, but also with computers and other electronic devices like cellphones, robots, and automobile entertainment systems. Gesture recognition has the potential to replace traditional human-computer interaction methods using touch or wired-controlled input devices [2].

Figure 1 depicts two kinds of gestures: static gesture, in which there is no movement in body stance or arm; only the hand remains fixed in a given pose across time. In the second scenario, the arm and hand move, and a series of positions change depending on the time interval. This is referred to as a dynamic gesture. Hand gesture detection approaches typically include three major steps: (i) pre-processing, (ii) feature extraction, and (iii) gesture identification. Gesture recognition needs hand movement in a video feed. It is accomplished by dividing the video clip into frames, then executing feature extraction stages and lastly identifying the hand motions.

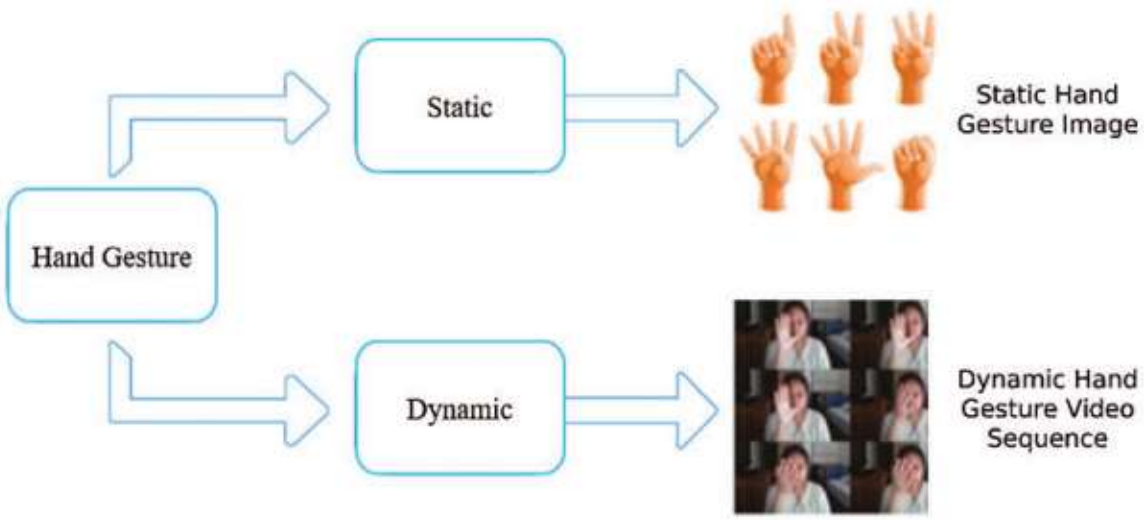

Figure 1: Type of hand gesture [3].

Deep Learning (DL) is a growing discipline and a subset of machine learning that is inspired by the human brain's function and structure. It employs several hidden neural network layers to improve model learning. It can correctly and quickly learn the characteristics of an item in complicated environments or backdrops. The Convolutional Neural Network (CNN) is a well-known deep learning model utilized in image-based applications. Nowadays, deep learning is employed for visual object identification and recognition, audio recognition, and a variety of other applications. Many hand gesture recognition algorithms have been presented in recent years that use deep learning techniques. Figure 2 depicts the fundamental processes for automated identification of hand movements [4].

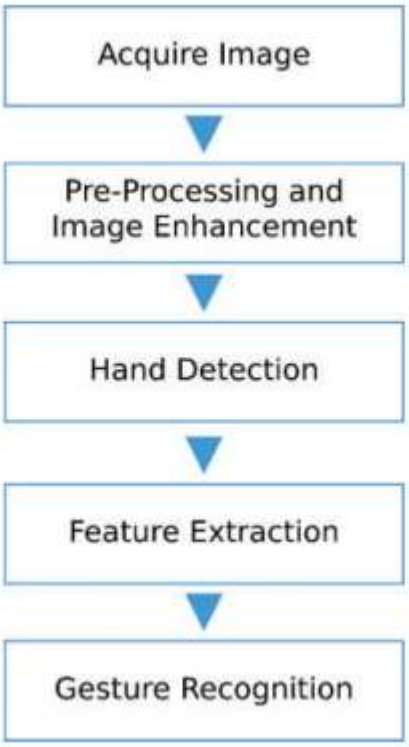

Figure 2: Basic steps for automatic gesture recognition[4].

Sign languages are not standard; they vary from spoken languages; even Arabic sign language differs by country and dialect. In general, deaf communities in Arab nations employ a variety of sign languages, including Egyptian, Jordanian, Tunisian, and Gulf. Although these languages may contribute to certain signals, there is a lack of education and communication between deaf and hearing individuals. These issues in the ArSL, as well as the discrepancies between it and spoken Arabic, highlight the necessity for machine translation, in addition to ArSL recognition systems. These devices may also assist the deaf in integrating into various stages of schooling and providing access to scientific information in their home language [5].

Two approaches adopted for sign language recognition can acquire data used in the academic literature ,namely ; Sensor-Based and Vision-Based [6] as shown in Figure 3.

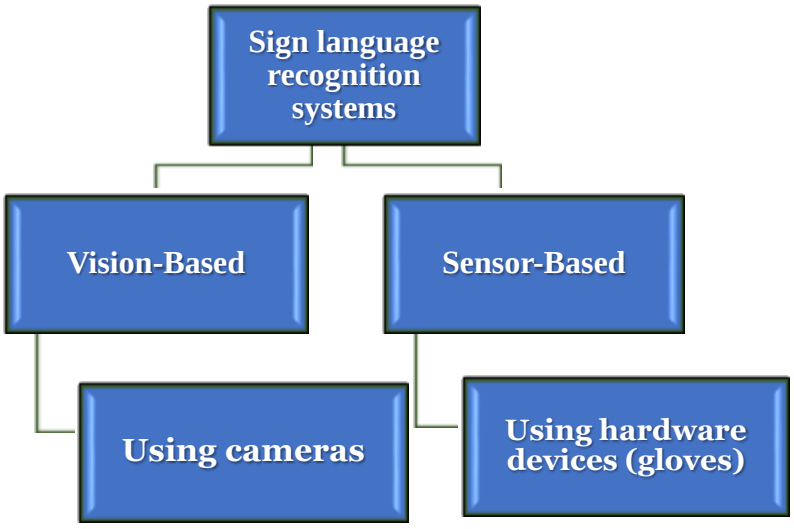

Figure 3 Sign language recognition approaches.

Sensor-based approach, n this approach, the location, hand motion, and velocity can be captured by the help of sensors and instruments. The main gestures capturing technology in the sensor-based approach, according to the study in [7] are:-Inertial measurement unit (IMU)—Measure the acceleration of the fingers ,This includes the use of gyroscope and accelerometer. Wi-Fi and Radar that senses the changes in the strength of the signals in the air by using electromagnetic signs. Electromyography (EMG) that detects the finger motion by taking the measurement of the electrical pulse in human muscle and reducing the bio-signal. and Others that utilize haptic technologies, mechanical, electromagnetic, ultrasonic, and flex sensors.

Sensor-based systems have a significant advantage over vision-based systems in that gloves may send the computer immediately data [8]. The device-based (Microsoft Kinect sensor, leap motion controller, and electronic gloves) can directly extract features without pre-processing, which means the Device-based can minimize the time for preparing sign language dataset , data can be obtained directly , and also can give a good accuracy rate in compared with vision-based [9]. But the disadvantage of this approach is that a physical

connection to the computer is required for the end user, which makes this method not preferred. In addition, it costs a lot due to the use of sensory gloves [8]. Despite the accuracy of the data that can be taken from devices, devices remain uncomfortable, whether they are wearing gloves or attached to a computer like a Leap motion or Microsoft Kinect [9]. Figure 4 shows the main stages of SL gesture data collection and recognition using the sensor-based system.

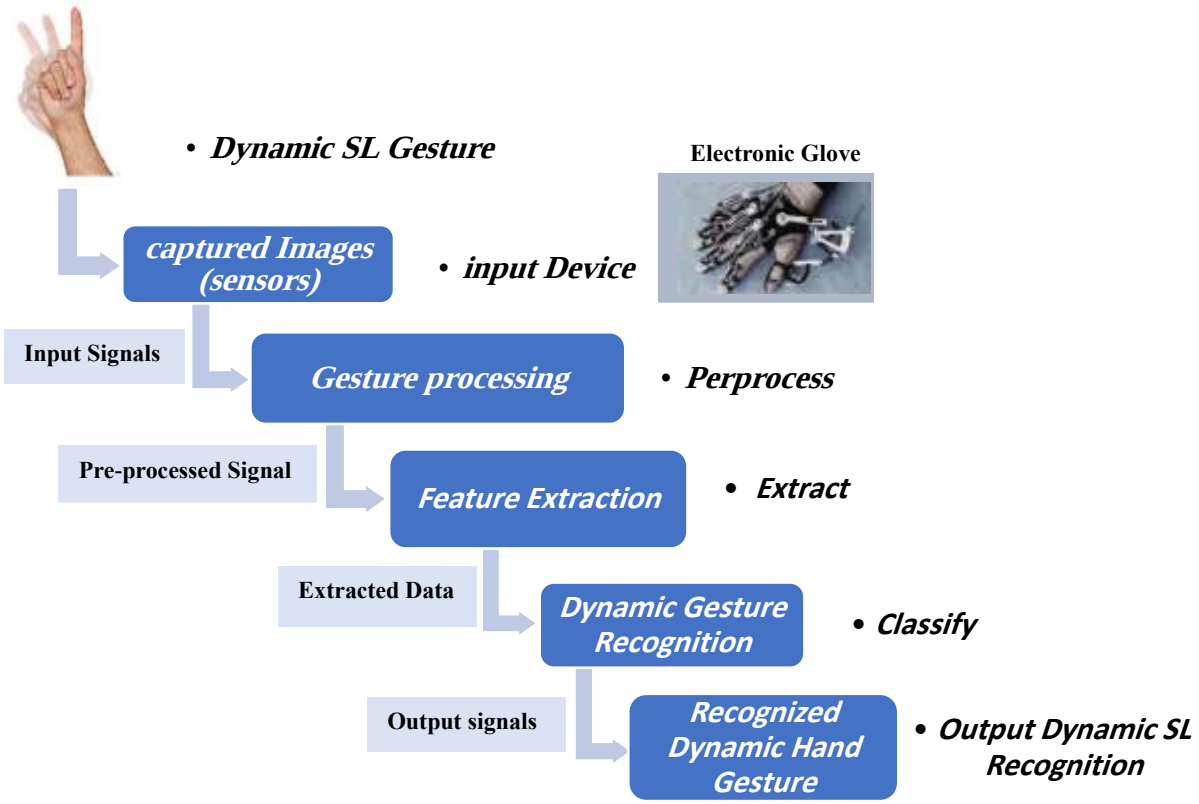

Figure 4 The main phases with regard to collecting and recognizing SL gesture data using the sensor-based system.

Another approaches the vision-based approach, in this approach, the hand gesture images can be acquired by employing a video camera. In gesture recognition, this approach is made up of appearance and 3D hand model approach. The main gestures capturing technology in vision-based approach found in [8] study are techniques depend on body markers including colored gloves, wristbands, and LED lights. Active techniques that use Kinect and LMC (Leap motion controller) for light projection. A single camera such as smartphone camera, webcam, and video camera .and Stereo camera, which provides extensive information by employing various monocular cameras. The main advantages of using a camera is that it eliminates the requirement for sensors in sensory gloves and reduced the system's build costs. Cameras are quite cheap , and most laptops use a high-specification camera because of the blur effect performed by a web camera [8].

Basically, vision-based sign language recognition can be roughly defined as the following four stages: acquiring gesture samples by Input Devices; preprocessing images to gain meaningful information; extracting features based on the design of gesture features and models; training models and achieving correct classification/recognition [10]. Figure 5 provides a detailed procedure of vision-based sign language recognition.

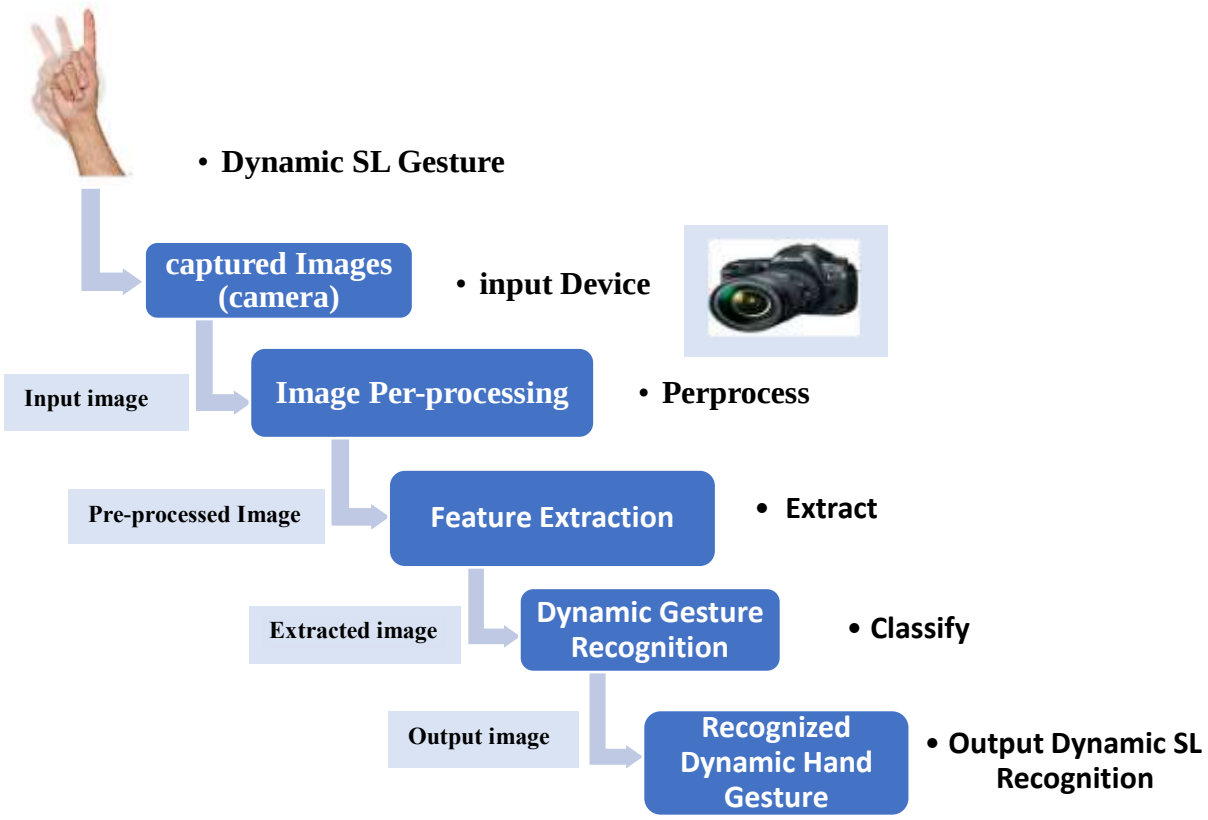

Figure 5 Procedure of vision-based sign language recognition.

there are a lot of studies related to Sign Language Recognition (SLR) that aim to reduce this gap between deaf and normal people, as it can replace the need for an interpreter. However, there are a lot of challenges facing the sign recognition systems, such as low accuracy, complicated gestures, the inexistence of large and complete datasets with different signs, and the inability of the models to process them adequately. Also, there are unique signs for each language [11, 12].

In this study, we suggested a deep learning (DL)-based model and employed MediaPipe in combination with RNN models to handle dynamic sign

language identification challenges while also automatically extracting robust temporal and spatial information. MediaPipe extracts keypoints from the hands, torso, and face to assess their position, shape, and orientation. RNN models, such as GRU, handle the problem of frame reliance in sign movement. These characteristics are challenging to extract using traditional machine learning approaches.

We relied on a vision-based approach. In this approach, the system relies on image processing and computer computations to process images and videos, as well as machine learning and deep learning techniques to classify and predict the processing data. In contrast, the sensor-based approach, such as using gloves, is expensive and impractical in everyday life situations because it requires the user to wear such sensors that rely on a continuous power supply, wires, and other requirements. This argument allowed the authors to discontinue this strategy for experiments and focus on the second approach. The second option has the benefit of using less expensive hardware. Smartphone cameras may be used as the capturing medium.

Due to a dearth of accessible datasets, we propose a new Arabic sign language dataset (words) emphasizing dynamic gestures. Most researchers deal with English-language datasets mostly concerned with static motions of letters and digits.

The rest of this paper is organized as follows: Section 2 shows the related work. Next, Section 3 provides the methodology and information on the Dataset, nominally DArSL 15. Then, in Section 4, the experimental results are discussed. In Section 5, the results are discussed. In Section 6, a conclusion and future work are presented.

II. Literature Review

As previously stated, sign language recognition is classified into two groups depending on how data is collected and processed prior to categorization. The

first kind is sensor-based; we employ hand-worn equipment, such as colored or customized gloves, to record key parts of the hand and sensor-based devices to capture motion. This approach is more dependable and accurate. However, it is unsuitable for real-world applications since it requires specialized equipment. The second kind is vision-based, which uses a standard camera to capture data.

This approach has several problems while reaching high precision, yet it is more time-efficient. This research will focus on a vision-based strategy that employs deep learning techniques and a review of other cutting-edge sign language recognition studies that utilize a variety of methodologies and datasets. A survey of sign language recognition techniques.

In [13] designed as ArSL based on the Hidden Markov Model (HMM). They collected a large dataset to recognize 20 isolated words from real videos taken of deaf people in different clothes and skin colors, and they achieved a recognition rate near 82.22%.

In [14] This study's scope includes the recognition of 50 dynamic word gestures. It proposes a novel technique to deal with pose variations in 3D object recognition. This technique uses a pulse-coupled neural network (PCNN) for image feature generation from two different viewing angles. The proposed technique obtained a recognition accuracy of 96%.

In [15] presents an automatic visual SLRS that translates isolated Arabic words signs into text. The proposed system has four main stages: hand segmentation, tracking, feature extraction, and classification. A dataset of 30 isolated words used in the daily school life of hearing-impaired children was developed to evaluate the proposed system, considering that 83% of the words have different occlusion states. Experimental results indicate that the proposed system has a recognition rate of 97% in signer-independent mode.

In [16] work in the Arabic Sign Language Recognition field. features extractor with deep behavior was used to deal with the minor details of Arabic Sign Language. A 3D Convolutional Neural Network (CNN) was used to recognize 25 gestures from the Arabic sign language dictionary. The recognition system was fed with data from depth maps. The system achieved 98% accuracy for observed data and 85%

average accuracy for new data. The results could be improved as more data from more different signers are included.

In [17] develop a computational structure for an intelligent translator to recognize the isolated dynamic gestures of the ArSL. We used 100-sign vocabulary from ArSL and applied 1500 video files for these signs. Experiments are performed on our own ArSL dataset, and the matching between the ArSL and Arabic text is tested using Euclidian distance. The evaluation of the proposed system for the automatic recognition and translation of isolated dynamic ArSL gestures has proven to be effective and highly accurate. The experimental results show that the proposed system recognizes signs with a precision of 95.8%.

In [12], a video-based Arabic sign language dataset containing 20 signs, performed by 72 signers, was created, and a deep learning architecture was proposed by combining CNN and RNN models. The authors have divided the data pre-processing into 3 stages. In the first stage, they reduced the dimensions of each frame in order to achieve less overall complexity. In the second, they passed the output to a function subtracting every two consecutive frames to find the motion between them. Finally, in the third stage, they unify each class's features and have 30 frames as output, where each unified frame combines (3x3) frames. The purpose of stage 3 is to reduce redundancy without losing any information. The main idea of the proposed architecture is to train two different CNNs independently for feature extraction, then concatenate the output into one single vector and pass the vector to an RNN for classification. The proposed model achieved 98% and 92% on the validation and testing subsets, respectively, on the suggested dataset. Moreover, they achieved promising accuracies of 93.40% and 98.80% on top-1 and top-5, respectively, on the UFC-101 dataset.

In [18] presents a computer program that can translate Iraqi sign language into Arabic (text). First, the translation starts with capturing videos to make up the dataset (41 words); the proposed system uses a convolutional neural network (CNN) to classify sign language based on its features to impute the sign's meaning. The accuracy of the part of the proposed system that translates sign language into Arabic text is 99% for words sign.

III. Methodology

The use of AI and deep learning in the area of sign language recognition has the potential to transform communication with hearing-impaired people. Deep learning techniques allow AI to understand complicated sign language gestures reliably. As such, this research aims to develop a deep learning model capable of detecting sign language video-based data utilizing feature extraction methodologies. utilizes the MediaPipe framework to extract characteristics based on crucial points in each signer's stance and hand. Finally, a classification operation was carried out in which the output features of each strategy were input into the GRU model, which falls within the category of RNN models.
Most DSL techniques do not detect motions. To address this issue, we divided our strategy into two components. The first step is feature extraction, which uses the MediaPipe framework to extract keypoints. The second is the DSL recognition module, which analyzes sign movement and generates the sign label.

Dataset

The main purpose of this study is to recognize dynamic gestures of Arabic sign language. Therefore, we create a new dataset containing a wide range of dynamic gestures of Arabic sign language, namely DarSL15-Dataset.
The DArSL15-Dataset contains fifteen Arabic gestures representing words, which were selected as the daily common adynamic gestures for ArSL. One volunteer collected each gesture. The collected gestures are chosen from two dictionaries: ("قاموس لغة الاشارة للأطفال الصم ") (see Figure 5) and("قاموس الاشاري العربي") (see Figure 6).

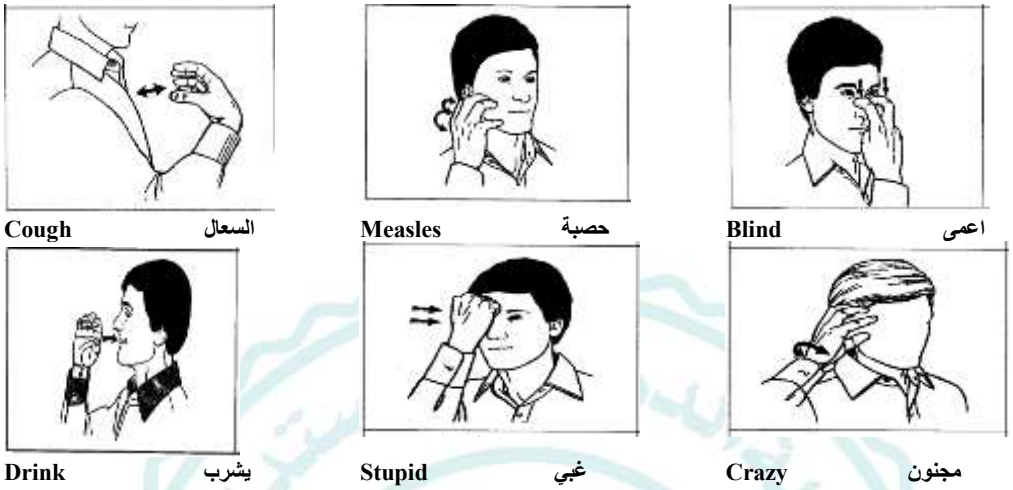

Figure 5. images of ArSL words dataset (dictionary قاموس لغة الاشارة للأطفال الصم) .

Figure 6 images of ArSL words dataset (dictionary قاموس الاشاري العربي).

One volunteer captured the videos, each recording 30 videos for each word gesture; a laptop camera was used to acquire the gestures, leading to a total number of videos = 15 × 1 × 30 = 450. The data were recorded using the Video Capture function in the OpenCV library. The video is saved in the NumPy format to be analyzed later.

Features Extraction Using MediaPipe

Google created MediaPipe, an open-source framework that allows developers to create multi-modal (video, audio, or any time series data) cross-platform applied machine learning pipelines. MediaPipe provides many human body identification and tracking algorithms trained using Google's enormous and varied dataset. They monitor critical spots on various body regions as a skeleton of nodes, edges, and landmarks. All coordinate points are three-dimensionally normalized. Google engineers use TensorFlow Lite to create models that are readily adaptable and customizable using graphs [19].

Sign language relies on hand gestures and stance estimation, yet DSL confronts several challenges due to continual mobility. The obstacles include detecting the hands, establishing their form, and determining orientation. MediaPipe was employed to address these issues. It extracts the important points for both hands' three dimensions (X, Y, and Z) and estimates the stance for each frame.

The pose estimation technique predicted and tracked the hand location relative to the body. The output of the MediaPipe framework is a list of keypoints for hands and pose estimation. For each hand, MediaPipe extracts 21 keypoints [20], as shown in Figure 7. The keypoints are calculated in the three-dimension space: X, Y, and Z for both hands. Thus, the number of extracted keypoints of hands is calculated as follows:

keypoints in hand × Three dimensions × No. of hands = (21 × 3 × 2) = 126 keypoints.

Figure 7. 21 Key-points for hand [21].

For Pose Estimation MediaPipe Extracts 33 keypoints [22] ,as shown in Figure 8. They are calculated in the three-dimension space: X, Y, and Z in addition to the visibility. Visibility is a value indicating whether the point is visible or hidden (occluded by another body part) on a frame. Thus, the number of extracted keypoints from the pose estimation is calculated as follows:

keypoints in pose × (Three dimensions + Visibility) = (33 × (3 + 1)) = 132 keypoints.

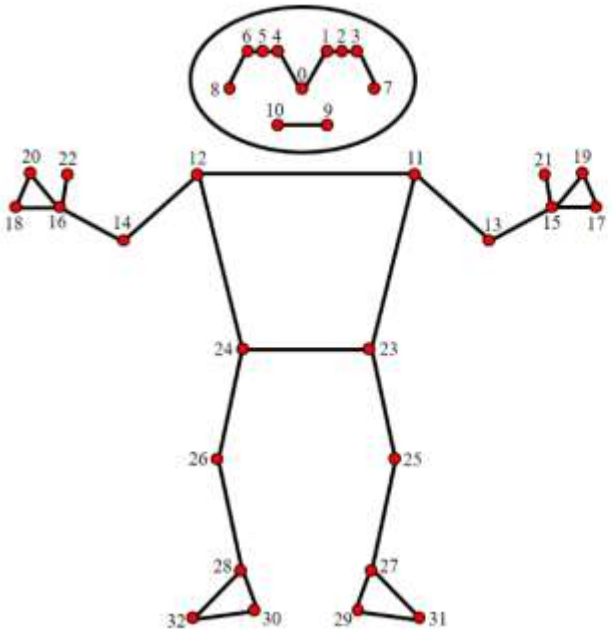

Figure 8. 33 Key-points for pose [23].

For the face, MediaPipe extracts 468 keypoints [24], as shown in Figure 9. Contours around the face, eyes, lips, and brows are represented by lines connecting landmarks, while the 468 landmarks are represented by dots. They are calculated in the three-dimension space: X, Y, and Z. Thus, the number of extracted key points from the face is calculated as follows:

Key points in face × Three dimensions = (468 × 3) = 1404 key points.

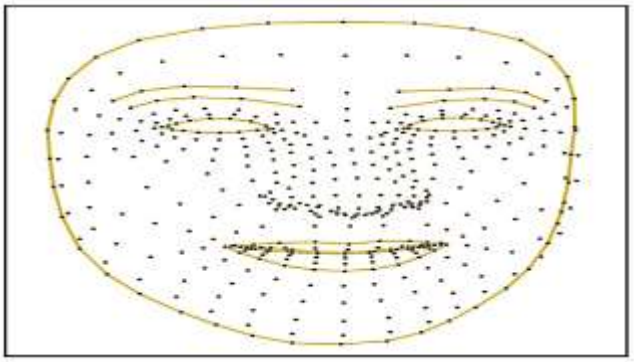

Figure 9. 468 Key-points for Face [11].

The total number of keypoints for each frame is calculated as follows: keypoints in hands + pose + face = (126 + 132 + 1404) = 1662 keypoints, as shown in Figure 10.

Volunteer 

Figure 10. Sample Frames from each extracted keypoint.

The Model:

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks, introduced in 2014 by Kyunghyang Cho et al. The GRU is like a long short-term memory (LSTM) with a gating mechanism to input or forget certain features but lacks a context vector or output gate, resulting in fewer parameters than LSTM [25].

Long-short-term memory (LSTM) is a type of Recurrent neural network (RNN) that can be used as a classifier. LSTM is known to be a suitable classifier for temporal data. It can determine the relations between different types of steps and find an internal representation of the whole series, which can be used as input to an artificial neural network that can classify this representation and find the suitable class.

Table 1. Model layer's parameters.

Parameters	value
RNN Model	GRU
Number of Nodes	Between (64,1662)

Activation	' Rule ' , ' SoftMax '
Optimizer	' Adam '

The model is now ready to receive the dataset and start the training phase based on the sequence of keypoints extracted from videos. Thus, the sign movement is analyzed, and the hand gesture label can be predicted. Therefore, the DArSL-15 can be recognized effectively.

Training and Testing Split:

Data Size: The dataset comprises data from volunteers, contributing 450 data points.
Training Set: For the training set, 337 data points were selected, representing 75% of the total data, ensuring a comprehensive representation of the variability within the dataset.
Testing Set: The remaining 113 data points were allocated to the testing set, representing 25% of the total data. This subset was reserved for evaluating the performance and generalization of the trained models, as shown in Table 2.

Table 2. Data size, training set, and test set for each volunteer.

Number of volunteers	Dataset size	Train	Test	Size test
One volunteer	450	337	113	0.25

Evaluation metrics

Although there are several ways to evaluate the performance of our classification model, the confusion matrix is the most widely used. It allows us to assess its effectiveness, pinpoint areas in which it failed, and get advice on adjusting our strategy.

A-Confusion Matrix:

An organized explanation of the predictive performance of a classifier on a dataset is used to compute a variety of assessment metrics (including accuracy, precision, and recall). This provides us with a full understanding of the effectiveness of our classification model and the kinds of errors it generates.

	Predicted positive	**Predicted negative**	
Actual positive	**TP (True positive)**	**FN (False negative)**	$Recall = \frac{TP}{TP + FN}$
Actual negative	**FP (False positive)**	**TN (True negative)**	
	$Precision = \frac{TP}{TP + FP}$		

Figure 11 Confusion matrix metrics.

Figure 11. illustrates the details of confusion matrix metrics and positive /negative production:

- **True Positives (TP)** are the number of positive class samples correctly classified by a model.
- **True Negatives (TN)** are the number of negative class samples correctly classified by a model.
- **False Positives (FP)** are the number of negative class samples that the model predicted (incorrectly) to be of the positive class.
- **False Negatives (FN)** are the number of positive class samples that the model predicted (incorrectly) to be of the negative class.

Confusion Matrix Calculation

Accuracy:

Definition: Accuracy is the most commonly used simple metric for classification. It represents the ratio of the correctly classified predictions out of the total number of predictions.
Interpretation: A high accuracy indicates that the model makes correct predictions overall.
Formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision:

Definition: Precision measures the proportion of true positive predictions among all positive predictions.
Interpretation: A high precision indicates that the model is likely to be correct when it predicts a positive class.
Formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

Definition: Recall measures the proportion of true positive predictions among all actual positive instances.
Interpretation: A high recall indicates that the model can identify most of the positive instances.
Formula:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

Definition: F1 score is the harmonic mean of precision and recall, providing a balanced measure between the two metrics.
Interpretation: F1 score considers precision and recall, making it suitable for imbalanced datasets where one class dominates.
Formula:

$$F1\text{-}Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

classification Report

The classification report for the model's performance on a specific dataset is presented. The report includes precision, recall, and F1-score for each class, along with overall metrics.

Utilization in Evaluation:

These evaluation metrics were utilized to assess the performance of the trained models on the testing datasets. By examining accuracy, precision, recall, and F1 score, we gained insights into the models' overall effectiveness, correctness, and robustness in recognizing Arabic Sign Language gestures.

IV. Results

An experiment was carried out using the DarSL15 dataset to assess the proposed system's functionality. To conduct the experiments, the DarSL15 Dataset was randomly divided into 75% for training and 25% for testing. The table below shows the performance metrics (Accuracy, Precision, Recall, and F1 Score) for various scenarios to test our proposed system.

Table 3 Experimental Results.

volunteer	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Volunteer	**0.97**	**0.98**	**0.97**	**0.97**

In Table 3, the volunteer contributed 450 data points, yielding 337 data points for training and 113 data points for testing.

Table 4 Classification Report of Experimental Results

The classification report offers more information about the model's performance for each class during the classification job. Here are the results of the classification report:

#	Arabic class name	English class name	Precision	Recall	F1-score
0	سعال	Cough	1.00	1.00	1.00
1	حصبة	Measles	0.83	1.00	0.91
2	اعمى	Blind	0.88	0.88	0.88
3	يشرب	Drink	1.00	1.00	1.00

4	غبي	Stupid	1.00	1.00	1.00
5	مجنون	Crazy	0.91	1.00	0.95
6	مع السلامة	Good bye	1.00	1.00	1.00
7	مهم	Important	1.00	1.00	1.00
8	نمو	To grow	1.00	1.00	1.00
9	اسكت(صمت)	Shut up(silence)	1.00	1.00	1.00
10	حضور	Presence (coming)	1.00	1.00	1.00
11	اهلا	Hello (congratulation)	1.00	1.00	1.00
12	توقف	To stop	0.71	1.00	0.83
13	امانة	Honesty	1.00	1.00	1.00

In Table 4. These results provide a thorough summary of the model's performance across multiple situations and detailed insights into its performance in each Classification Report of Experimental Results class.
Based on the categorization report results, we discovered that classes (1, 2, 5, and 12 performed relatively poorly compared to the other courses. This is due to the nature of movement in these classes, where the distinction between individual movements may be unclear. For example, the movement could be a slight hand gesture with no substantial variations in motion, or the difference between one movement and another may not be obvious enough, making classification more difficult for these classes.

V. Discussion

- **Interpretation of results**

High values of accuracy, precision, recall, and F1 score indicate successful model performance, while lower values may signify areas for improvement in the model's predictive capabilities.

The "macro" averages provide a basic average of the metrics produced for each class, whereas the "weighted" averages correct for class imbalance by weighting the average according to the number of instances in each class.
Overall, both "macro" and "weighted" averages show comparable trends across situations, with "weighted" averages indicating the impact of class distribution on model performance.

- **Discussion of Classification Report Results:**

Examining the categorization report reveals information about the model's performance across classes. Classes 1, 2, 5, and 12 notably had worse precision, recall, and F1-scores than the other classes. This finding shows difficulties in appropriately identifying these specific classes.

Several factors contribute to these classes' inferior performance. First, the nature of the movements within these classes may provide complexity that is difficult to determine fully. For example, these movements may include subtle gestures or minor differences between different signs, making it difficult for the model to distinguish between them efficiently.

Furthermore, the classification model may have problems capturing the intricacies of these movements, particularly if they include small fluctuations or sophisticated hand movements that are difficult to identify precisely.

Moreover, the minimal size and diversity of the dataset for these classes may have contributed to the poor performance. A larger and more diversified dataset would give the model a broader set of instances, improving its capacity to generalize and identify these complex movements.

To summarize, while the model's overall performance is acceptable, further modification and augmentation of the dataset and the model architecture are required to enhance classification accuracy for these hard classes. This highlights the need for ongoing research and development efforts in sign language recognition to solve these unique issues while improving the accessibility and effectiveness of sign language recognition technology.

The observed influence of increasing dataset size emphasizes the need of data augmentation and the establishment of larger, more diverse datasets in sign language recognition research. As part of the study's objectives, the goal was to create a comprehensive dataset exclusively for Arabic sign language recognition. By expanding the dataset, the model can be trained on a broader collection of instances, boosting its capacity to generalize and reliably identify sign language movements, especially in difficult categories. This is consistent with the overall goal of improving the accessibility and effectiveness of sign language recognition systems, ultimately leading to greater inclusivity and accessibility for people with hearing impairments.

VI. Conclusions and future work

This paper aimed to apply deep learning methods to provide better communication between the deaf community and the hearing majority. According to the World Health Organization (WHO), almost 446 million people worldwide suffer from impaired hearing loss; studies reveal that this number may continue to rise. Therefore, it is urgent to develop an effective sign language translator to facilitate the lives of the hearing-impaired community.

This research is divided into two approaches to feature extraction. The MediaPipe framework was used to extract the key points from the videos of the DArSL15_Dataset. The features were then passed through the proposed GRU model to identify the relationship between the sequences and produce the overall prediction. From the experimental results, it can be observed that the GRU model achieved an average performance of (97)%.

Creating a large dataset with various signers will be considered in the future. In addition, more advanced algorithms for data preprocessing and learning can be used to recognize real-time videos with more complex environments and different durations.

Reference

1. Meng, L. and R. Li, *An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network.* Sensors, 2021. **21**(4): p. 1120.
2. Ur Rehman, M., et al., *Dynamic hand gesture recognition using 3D-CNN and LSTM networks.* Computers, Materials & Continua, 2021. **70**(3).
3. Zhang, Y., et al., *A real-time recognition method of static gesture based on DSSD.* Multimedia Tools and Applications, 2020. **79**(25): p. 17445-17461.
4. Materzynska, J., et al. *The jester dataset: A large-scale video dataset of human gestures*. in *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019.
5. Tharwat, G., A.M. Ahmed, and B. Bouallegue, *Arabic sign language recognition system for alphabets using machine learning techniques.* Journal of Electrical and Computer Engineering, 2021. **2021**(1): p. 2995851.

6. Al-Shamayleh, A.S., et al., *Automatic Arabic sign language recognition: A review, taxonomy, open challenges, research roadmap and future directions.* Malaysian Journal of Computer Science, 2020. **33**(4): p. 306-343.
7. Cheok, M.J., Z. Omar, and M.H. Jaward, *A review of hand gesture and sign language recognition techniques.* International Journal of Machine Learning and Cybernetics, 2019. **10**: p. 131-153.
8. Ahmed, M.A., et al., *A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017.* Sensors, 2018. **18**(7): p. 2208.
9. Mohammed, R. and S. Kadhem. *A review on arabic sign language translator systems*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
10. Jiang, X., et al., *A survey on artificial intelligence in Chinese sign language recognition.* Arabian Journal for Science and Engineering, 2020. **45**: p. 9859-9894.
11. Samaan, G.H., et al., *Mediapipe's landmarks with rnn for dynamic sign language recognition.* Electronics, 2022. **11**(19): p. 3228.
12. Balaha, M.M., et al., *A vision-based deep learning approach for independent-users Arabic sign language interpretation.* Multimedia Tools and Applications, 2023. **82**(5): p. 6807-6826.
13. Youssif, A.A., A.E. Aboutabl, and H.H. Ali, *Arabic sign language (arsl) recognition system using hmm.* International Journal of Advanced Computer Science and Applications, 2011. **2**(11).
14. Elons, A.S., M. Abull-Ela, and M.F. Tolba, *A proposed PCNN features quality optimization technique for pose-invariant 3D Arabic sign language recognition.* Applied Soft Computing, 2013. **13**(4): p. 1646-1660.
15. Ibrahim, N.B., M.M. Selim, and H.H. Zayed, *An automatic Arabic sign language recognition system (ArSLRS).* Journal of King Saud University-Computer and Information Sciences, 2018. **30**(4): p. 470-477.
16. ElBadawy, M., et al. *Arabic sign language recognition with 3d convolutional neural networks*. in *2017 Eighth international conference on intelligent computing and information systems (ICICIS)*. 2017. IEEE.
17. Ahmed, A., et al., *Arabic sign language translator.* Journal of Computer Science, 2019. **15**(10): p. 1522-1537.
18. Kadhem, S.M., *Iraqi sign language translator system using deep learning.* Al-Salam Journal for Engineering and Technology, 2023. **2**(1): p. 109-116.
19. Halder, A. and A. Tayade, *Real-time vernacular sign language recognition using mediapipe and machine learning.* Journal homepage: www. ijrpr. com ISSN, 2021. **2582**: p. 7421.
20. Zhang, F., et al., *Mediapipe hands: On-device real-time hand tracking.* arXiv preprint arXiv:2006.10214, 2020.
21. Wu, T.-L. and T. Senda, *Pen Spinning Hand Movement Analysis Using MediaPipe Hands.* arXiv preprint arXiv:2108.10716, 2021.
22. Bazarevsky, V., I. Grishchenko, and K. Raveendran, *BlazePose: On-device Real-time Body Pose tracking. 2020. DOI: 10.48550.* ARXIV, 2006.

23. Chen, K.-Y., et al., *Fitness Movement Types and Completeness Detection Using a Transfer-Learning-Based Deep Neural Network.* Sensors, 2022. **22**(15): p. 5700.
24. Kartynnik, Y., et al., *Real-time facial surface geometry from monocular video on mobile GPUs.* arXiv preprint arXiv:1907.06724, 2019.
25. Dey, R. and F.M. Salem. *Gate-variants of gated recurrent unit (GRU) neural networks*. in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. 2017. IEEE.