A Lightweight Neural Network for Arabic Sign Language Recognition Using Mediapipe Landmarks

Raed Fidawi[a], Zein Chehabeddine[a], Rida Assaf[a,*]

[a]*Department of Computer Science, American University of Beirut, Beirut, Lebanon*

Abstract

Hearing impairment affects approximately 6% of the global population, creating significant barriers to communication and social integration for millions of people. Sign language offers a crucial means of communication for the deaf and hard-of-hearing community; however, its widespread adoption and understanding remain limited. This gap is particularly evident in Arabic-speaking regions, where multiple dialects further complicate sign language translation. Additionally, research on Arabic Sign Language (ArSL) has traditionally lagged behind more commonly studied sign languages such as American Sign Language (ASL).

In this work, we propose a lightweight neural network for ArSL recognition that requires only a fraction of the parameters used by state-of-the-art models. Our approach leverages Mediapipe to generate feature vectors using extracted hand landmarks. We evaluate our model on both static hand symbols (using the ArSL21L dataset) and dynamic hand gestures (using a newly collected Lebanese dialect dataset). Crucially, our new in-house dataset fills a gap in public datasets for dynamic ArSL gestures, and we make it available for future research upon request.

Experimental results show that our symbols model achieves 97% accuracy, precision, recall, and F1-score on ArSL21L, while the gestures model reaches 100% accuracy on our in-house dataset. These results are competitive with YOLO (You Only Look Once)-based architectures (YOLOv5 and YOLOv7), yet our model requires significantly fewer parameters—on the order of 0.078 M versus millions in YOLO networks. Our findings contribute to the expanding research on sign language recognition for underrepresented languages like ArSL.

Keywords: Arabic Sign Language, Lightweight Neural Network, Hand Gesture Recognition, Mediapipe, YOLO, Deep Learning

1. Introduction

Hearing impairment affects approximately 6% of the global population, creating significant barriers to communication and social integration. Sign language is a critical mode of communication for deaf and hard-of-hearing communities. Despite considerable progress in machine learning–based sign language recognition for widely used languages like ASL, Arabic Sign Language (ArSL) has received less attention. Moreover, multiple Arabic dialects add complexity to sign language usage, underscoring the need for robust technological solutions.

Recent advances in object detection frameworks (e.g., YOLO) have shown promise in real-time sign language recognition tasks [1, 2, 3, 4, 5]. However, these frameworks tend to involve millions of parameters, making them resource-intensive. Static sign classification has been a cornerstone in ArSL research. Batnasan [1] experimented with YOLOv5 on ArSL21L, demonstrating high accuracy for various YOLOv5 variants. Attia *et al.* [3] improved YOLOv5 by modifying activation functions, and similar enhancements have been proposed for YOLOv7 [5] and YOLOv8 [4]. While these models excel in accuracy, their parameter counts frequently range in the millions.

Dynamic gestures involve sequential hand motions. Cayme *et al.* [6] tackled Filipino Sign Language with a CNN-LSTM hybrid, whereas Caliwag *et al.* [7] applied CNN-based approaches to Argentinian and Chinese sign languages. For ArSL, limited availability of public gesture datasets has been a consistent bottleneck. Rady El Rwelli *et al.* [8] experimented with multiple classifiers, including SVM and CNN, and reported around 90% accuracy on their dataset (not publicly available).

Our study seeks to develop a lightweight alternative that maintains competitive accuracy while drastically reducing parameter count, and introduces a dataset that can be used for gesture prediction.

We target two main tasks for ArSL:

1. **Symbols Model:** Classifying static hand signs (alphabets) using the ArSL21L dataset [9].
2. **Gestures Model:** Recognizing dynamic hand gestures (words or short phrases) from video frames. Because publicly available ArSL gestures datasets are scarce, we introduce a novel Lebanese dialect dataset that can be extended or adapted for broader research.

*Address for correspondence: Rida Assaf, Department of Computer Science, American University of Beirut, P.O. Box 11-0236, Riad El-Solh, Beirut 1107 2020, Lebanon, Bliss Hall 221, Phone: +961 (1) 350000, Ext: 4236, email address: ra278@aub.edu.lb

Email addresses: rnf14@mail.aub.edu (Raed Fidawi), zms32@mail.aub.edu (Zein Chehabeddine), ra278@aub.edu.lb (Rida Assaf)

By extracting hand landmarks via Mediapipe [10] and converting these landmarks into normalized feature vectors, we reduce input dimensionality and training overhead. We attain high accuracy on both static (symbols) and dynamic (gestures) tasks, using models with a fraction of the parameters commonly found in YOLO-based methods.

2. Methodology

We design two models: (1) a symbols model for static alphabets and (2) a gestures model for dynamic signs. Both adopt Mediapipe's hand recognition to extract 21 landmarks per hand, each with (x, y, z) coordinates (Figure 1). If both hands are present, we obtain 42 total landmarks (i.e., 126 coordinates).

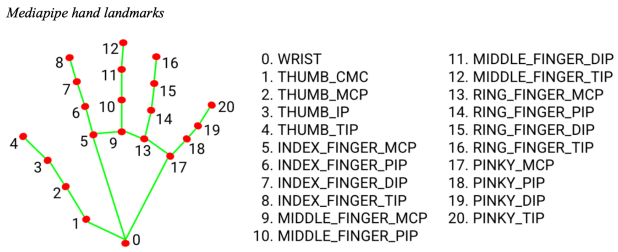

Figure 1: Visual representation of Mediapipe's 21 hand landmarks, including wrist, finger joints, and tips, used for tracking hand movements in three-dimensional space (X, Y, Z coordinates). Each point corresponds to a specific anatomical location on the hand.

2.1. Datasets

ArSL21L (Symbols). For static alphabets, we use the ArSL21L dataset [9] which contains images of ArSL letters in varied settings (lighting, angles, distances) to encourage robust generalization.

Lebanese Dialect Dataset (Gestures). We introduce a new in-house dataset focusing on 5 word-level gestures in the Lebanese dialect of ArSL. Each gesture is recorded via short video clips at 4 FPS, yielding 10 frames per clip. This rate was determined empirically to balance between capturing sufficient gesture information while avoiding redundant frames. We extract 126 landmarks per frame for both hands (if present) and concatenate them, forming a 1260-dimensional feature vector per gesture sequence.

Since collecting extensive videos is time-intensive, we employ data augmentation at the landmark level. By adding small offsets to the extracted coordinates, we efficiently increase dataset diversity without storing additional images or video files. Our dataset is publicly available and linked to at the end of the manuscript.

2.2. Symbols Model Development

Each image in ArSL21L is processed via Mediapipe's *static* mode, yielding 63 landmarks per hand. For consistency across images, we apply the following:

1. **Scaling**: Multiply x and y coordinates by image width and height, respectively.
2. **Shifting**: Translate landmarks so the wrist (landmark 0) is the origin $(0, 0)$.

This produces a 126-dimensional feature vector for two hands. A fully connected neural network (with layers $[256, 128, 64]$, ReLU activations, dropout layers, and a final softmax for multi-class classification) is trained using Adam (lr = 0.001, batch size = 128) for 1000 epochs. Training completes in about five minutes on an AMD Ryzen 5 3600X CPU with 16 GB RAM and a 1 TB HDD.

2.3. Gestures Model Development

For dynamic gestures, Mediapipe's *video* mode is applied to 10 frames per gesture. We again scale, shift, and concatenate the coordinates across frames, creating a single 1260-dimensional feature vector for classification. We use a similar network architecture but adjust the input dimension to match 1260. Training finishes in roughly one minute on the same hardware.

3. Results

3.1. Symbols Model on ArSL21L

Our symbols model achieves 97% accuracy, precision, recall, and F1-score on ArSL21L. Figure 2 shows a confusion matrix that demonstrates minimal misclassifications across the alphabet. Despite this strong performance, the model's parameter count is only about 78k, far below typical YOLO models.

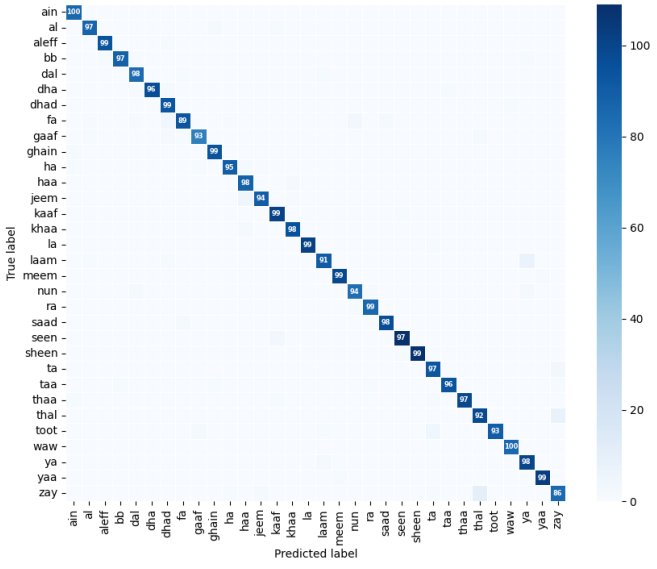

Figure 2: Confusion matrix of our symbols model on the ArSL21L dataset, illustrating high classification accuracy across classes.

Table 1 compares our approach to YOLO-based models, showing that while some YOLO variants have slightly higher accuracy (e.g., enhanced YOLOv5x, YOLOv7-medium, enhanced YOLOv8), they require significantly more parameters,

often multiple orders of magnitude greater than our model. Notably, our lightweight approach outperforms certain smaller YOLO variants (e.g., YOLOv5s, YOLOv5m) in precision and recall.

Table 1: Comparison of YOLO-based models vs. our lightweight model on the ArSL21L dataset. Our approach achieves competitive accuracy with a significantly smaller parameter count.

Model	Reference	Precision(%)	Recall(%)	Params (M)
YOLOv5s	[1]	95.3	94.0	7.2
YOLOv5m	[1]	96.8	94.6	21.2
YOLOv5I	[1]	97.8	97.6	46.5
Enhanced YOLOv5x	[3]	98.2	98.5	86.7
YOLOv7-tiny	[5]	96.4	96.4	6.0
YOLOv7-medium	[5]	98.2	98.3	36.9
Enhanced YOLOv8	[4]	99.0	99.0	8.0
Proposed model	–	97.0	97.0	0.078

3.2. Gestures Model on the Lebanese Dialect Dataset

Our gestures model achieves 100% accuracy, precision, recall, and F1-score on the 5-class Lebanese dialect dataset. While this high result is promising, it should be noted that the dataset remains small and covers only a handful of words. Future expansions will include additional classes to ensure broader coverage of ArSL gestures. Similar studies on other languages or dialects (e.g., Argentinian) have reported accuracies in the 90–98% range.

Because other ArSL gesture datasets are either unavailable or incompatible with our landmark-based approach, we could not perform a direct head-to-head comparison with existing gesture models. Attempts to adapt other publicly accessible gesture datasets (e.g., Argentinian, Filipino sign languages) were often thwarted by data format mismatches or poor landmark detection in non-Arabic sign datasets. However, the 100% result highlights the potential of our approach, and we expect to benchmark more extensively once additional ArSL gesture datasets become publicly available.

4. Discussion

Our study shows that feeding coordinate-based data from Mediapipe into a fully connected neural network can produce high-performing models for both static and dynamic ArSL recognition. Key advantages include:

- **Lightweight Architecture.** Fewer than 80k parameters drastically reduce memory usage relative to YOLO models.
- **Effective Normalization.** Scaling and shifting landmarks to a wrist-based origin helps reduce sensitivity to varying image sizes, backgrounds, or hand placements.
- **Custom Data Augmentation.** Altering numerical landmarks instead of raw images enables fast data expansion with minimal storage overhead.

Limitations. The main limitation is the relatively small size of our Lebanese dialect gestures dataset (5 classes). Consequently, we are unable to compare directly to many existing gesture-recognition systems, because their datasets are unpublished or not suitable for Mediapipe-based processing. We plan to expand our dataset to include more gestures and dialect variations, enabling richer comparisons and more robust conclusions.

Future Work. Beyond expanding the dataset, we aim to evaluate our model's real-time performance on mobile or embedded platforms, given its lightweight architecture. Incorporating additional body or facial cues could further improve recognition of complex signs that involve facial expressions or head movements.

Funding Sources

This work was fully supported by the University Research Board (Grant number: 104518) at the American University of Beirut (AUB).

Research Data

The dataset used in this study is publicly available and can be accessed at the following link: `https://drive.google.com/drive/u/2/folders/1YcptbU7E5TSQ-QyOS-IoTOnKRzQ4aO1C`

Competing interests

All authors declare no financial or non-financial competing interests.

Author contributions

R.A conceived and outlined the overall research goals and aims, shaped the experimental approach and technical design, supervised the development of the project, and contributed to writing and editing the manuscript. Z.C. and R.F. collected and processed the datasets, implemented the model and codebase, conducted the evaluations, and contributed to the writing of the manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the readibility of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Batnasan, G., Gochoo, M., Otgonbold, M. E., Alnajjar, F., & Shih, T. K. (2022). Arsl21l: Arabic sign language letter dataset benchmarking and an educational avatar for metaverse applications. In 2022 IEEE global engineering education conference (educon) (pp. 1814-1821). IEEE. https://doi.org/10.1109/EDUCON52537.2022.9766497

[2] Alamri, F. S., Rehman, A., Abdullahi, S. B., & Saba, T. (2024). Intelligent real-life key-pixel image detection system for early Arabic sign language learners. *PeerJ Computer Science*, 10, e2063. https://doi.org/10.7717/peerj-cs.2063

[3] Attia, N. F., Saidahmed, M., Ahmed, S., & Alshewimy, M. A. (2024). *Improved Deep Learning Model based Real-Time Recognition of Arabic Sign Language*. https://digitalcommons.aaru.edu.jo/erjeng/vol8/iss1/31?utm_source=digitalcommons.aaru.edu.jo%2Ferjeng%2Fvol8%2Fiss1%2F31&utm_medium=PDF&utm_campaign=PDFCoverPages

[4] Ahmadi, S. A., Mohammad, F., & Dawsari, H. A. (2024). Efficient YOLO-Based Deep Learning Model for Arabic Sign Language Recognition. *Deleted Journal*, 3(4). https://doi.org/10.57197/jdr-2024-0051

[5] Mazen, F., & Ezz-Eldin, M. (2024). A Novel Image-Based Arabic Hand Gestures Recognition Approach Using YOLOv7 and ArSL21L. *Fayoum University Journal of Engineering*, 7(1), 40-48. https://doi.org/10.21608/fuje.2023.216182.1050

[6] Cayme, K. J., Retutal, V. A., Salubre, M. E., Astillo, P. V., Cañete, L. G., & Choudhary, G. (2024). Gesture Recognition of Filipino Sign Language Using Convolutional and Long Short-Term Memory Deep Neural Networks. *Knowledge*, 4(3), 358–381. https://doi.org/10.3390/knowledge4030020

[7] Caliwag, A. C., Hwang, H., Kim, S. & Lim, W. (2022). Movement-in-a-Video Detection Scheme for Sign Language Gesture Recognition Using Neural Network. *Applied Sciences*, 12(20), 10542. https://doi.org/10.3390/app122010542

[8] El Rweily, R., Shahin, O. R., Taloba, A. I. (2021). Gesture based Arabic Sign Language Recognition for Impaired People based on Convolution Neural Network. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021.

[9] [dataset] Arabic Sign Language Dataset. Ammar Sayed Taha, Arabic Sign Language Dataset 2022 [dataset], Kaggle, 2022. https://www.kaggle.com/datasets/ammarsayedtaha/arabic-sign-language-dataset-2022 (accessed 10 October 2024).

[10] Google, Mediapipe: A cross-platform framework for building multimodal ML pipelines [software], v0.9.1.0, GitHub, 2023. https://github.com/google/mediapipe (accessed 10 January 2024).