International Student Research Review
Volume 3 Issue 1 April 2026
DOI: https://doi.org/10.53378/isrr.212

Enhancement of Recurrent Neural Networks (RNN) applied in hand gesture recognition for American Sign Language (ASL) alphabet recognition

[1]Jenny R. Jimenez & [2]Augustin Brain C. Sabordio

Abstract

This study focused on improving Recurrent Neural Network (RNN) algorithms to recognize American Sign Language (ASL) alphabets to better hand segmentation, feature extraction, and time modeling to obtain better accuracy and robustness in the real-time recognition. An experimental research design was employed using 260 video samples representing all 26 ASL letters under both ideal and challenging environmental conditions. The enhanced model integrates MediaPipe-based hand detection with adaptive preprocessing, multimodal feature extraction combining 3D landmarks and engineered articulation features (99-dimensional vectors), and adaptive temporal modeling using extended sequence buffering and prediction smoothing. Dual-stream neural architecture takes visual and numerical data of landmarks and processes them before being classified through LSTM layers and softmax output. The improved system achieved an overall accuracy of 97.70% and a mean confidence of 91.45%, which is 38.85 percentage points higher than the baseline model. Accuracy in challenging conditions was significantly improved, with a degradation rate of only 1.60% compared to 8.50% in the baseline. The recognition of visually similar letters reached 98% accuracy, while dynamic letters J and Z achieved relative improvements of 160% and 126%, respectively. The current research study is limited to the recognition of ASL alphabets (A–Z) in controlled experimental conditions. Future research may extend the system to full-word recognition and real-world deployment scenarios.

Keywords: *3D landmark feature extraction, temporal modeling, human–computer interaction, MediaPipe*

Article History:

***Received**: March 8, 2026* ***Revised**: April 10, 2026*
***Accepted**: April 18, 2026* ***Published online**: April 30, 2026*

Suggested Citation:
Jimenez, J.R. & Sabordio, A.B.C. (2026). Enhancement of Recurrent Neural Networks (RNN) applied in hand gesture recognition for American Sign Language (ASL) alphabet recognition. *International Student Research Review*, 3(1), 22-40. https://doi.org/10.53378/isrr.212

About the authors:
[1]Corresponding author. Bachelor of Science in Computer Science student, Pamantasan ng Lungsod ng Maynila. Email: jimenezjenny599@gmail.com
[2]Bachelor of Science in Computer Science Student, Pamantasan ng Lungsod ng Maynila. Email: austinbrain25@gmail.com

1. Introduction

The recent intensive development of artificial intelligence (AI) and deep learning has significantly changed the field of Human-Computer Interaction (HCI), specifically the field of gesture-based communication systems. Modern studies show that such technologies can enable more sophisticated and attentive interfaces, which further increase their engagement and interaction fidelity (Goodfellow et al., 2016; LeCun et al., 2015). Recurrent Neural Networks (RNNs) are one of the varied deep learning structures that demonstrate significant effectiveness during the process of sequential data modelling and it can be explained through the fact that RNNs have the inherent ability to maintain temporal associations among the input streams. This feature makes RNN-modeled systems particularly best adapted to gesture recognition systems in which the spatiotemporal dynamics of movement are critical. These models are very important in the particular field of sign language recognition where visual movement of hands are translated into semantically meaningful linguistic representations. Human-Computer Interaction research emphasizes that natural user interfaces, such as hand gestures, provide intuitive and efficient communication between humans and machines. As such, integrating deep learning with gesture recognition technologies presents promising opportunities for accessibility-focused innovations.

The Sign-language recognition systems represent a critical line of improving the accessibility of communication to deaf or hearing-impaired persons. The World Health Organization (WHO, 2021) reports that over 1.5 billion people worldwide experience some degree of hearing loss, underscoring the critical need for inclusive and accessible technological interventions. One of the most widespread sign languages, the American Sign Language (ASL), represents a system of meaning encoding via advanced arrangements of the hand, fingers movements and motion in space. However, automatic recognition of ASL is still a daunting task, which can be explained by inter-, and intra-subject variation that comes due to the situation with the lighting conditions, background clutter, hand position, and the speed of a signer. These environmental and behavioral factors directly affect key research variables such as hand segmentation accuracy, feature extraction granularity, and temporal modeling effectiveness.

Existing RNN-based ASL recognition systems often depend on 2D image features or limited keypoint representations, which restrict their ability to distinguish visually similar letters such as A, E, M, N, S, and T. Furthermore, rigid temporal modeling techniques using

fixed-length frame sampling frequently fail to capture the natural timing of dynamic letters such as J and Z. Many prior studies report satisfactory accuracy under controlled laboratory conditions but demonstrate significant performance degradation in real-world environments (Abiyev et al., 2020; Saleh & Issa, 2020; Prakash et al., 2020). The observation defines an acute gap in the literature on the creation of hearty multi-modal feature representations and adaptive temporal modelling schemes capable of enduring high recognition accuracy under problematic conditions. Addressing this gap is essential to improve the reliability and practical deployment of sign language recognition systems.

The purpose of this study is to enhance RNN algorithms for ASL alphabet recognition by improving three critical components: robust hand detection and segmentation, high-resolution 3D feature extraction, and adaptive temporal modeling. This paper used MediaPipe-based three-dimensional landmark detection, designed articulation features, and a long sequence buffering scheme with prediction smoothing to improve the overall classification accuracy and confidence of the real-time applications. The significance of this research is supported by the fact that it helped in the creation of assistive capabilities, the designing of education tools to be used by the American Sign Language learners and the evolution of inclusive communication systems. The polished model sets performance standards that are higher in the alphabet-level recognition of ASL and provides a scaffold, which can be scaled in future to a word-level recognition and sentence-level recognition tasks.

This study aims to create strong hand segmentation schemes maintaining a high degree of accuracy in cluttered environments and changing light conditions, get improved multimodal feature extraction schemes, which can be used to discriminate visually similar ASL characters through 3D spatial encoding and the design of articulation descriptors, and design adaptive temporal modelling schemes allowing maintenance of the natural cadence of gestures and improve the recognition of motion-based characters.

2. Literature Review

2.1 Recurrent Neural Networks in Gesture Recognition

New developments in the field of deep learning have greatly enhanced gesture and sign language recognition by utilizing sequential structures, in terms of RNNs, Long Short-Term Memory (LSTM) units, and Gated Recurrent Units (GRUs) and so on. Prakash et al.

(2020) have provided evidence of how the combination of Convolutional Neural Networks (CNNs) and RNN models can be utilized with the aim of improving the recognition of gestures due to the simultaneous consideration of both spatial and temporal cues. On the same note, Rivera-Acosta et al. (2021) developed a real-time ASL-to-translation system that uses a YOLO network with LSTM layers with better temporal consistency in the predictions. More recently, Miah et al. (2024) provided a more detailed focus on pointing out that large scale datasets using graph-based deep neural networks perform better than traditional RNN-only systems, especially when it comes to the modeling of motion dynamics. However, even with these developments, a significant number of RNN-based models still do not demonstrate good results with visually related hand configurations and environmental changes, hence highlighting the necessity to develop improved multimodal feature encoding and flexible modelling controls.

2.2 Hand Detection and Multimodal Feature Extraction

Proper hand recognition and effective feature extractor are the main directions of the ASL recognition systems. Shin et al. (2021) derived the coordinates of the hand joints in three dimensions, which have better classification of ASL alphabets, showing that discrimination of similar signs is better with the help of spatial encoding of depth information. Saleh and Issa (2020) then demonstrated that overfitting and model generalization of deep neural networks are suppressed and enhanced by data augmentation and fine-tuning of the models when applied to sign language data. In another study, Akdag and Baykan (2024) investigated the use of the feature fusion to improve the performance in signer-independent recognition, and they found that a combination of the geometric and visual features enhances the recognition ability in different environmental factors. Further, Cayme et al. (2024) used multimodal feature extraction by combining CNN with LSTM networks to recognize Filipino Sign Language and claim that multimodal feature extraction is much more effective than single-stream architectures. These results also highlight the importance of using 3D landmark-based engineered characteristics in conjunction with visual data in distinguishing the minor articulation variations in ASL alphabets.

2.3 Temporal Modeling and Dynamic Gesture Recognition

Dynamic ASL characters like J and Z require specific time modeling other than the frame-based analysis. In their dataset, Abiyev et al. (2020) did not include motion-based

letters because they were limited to modeling the static postures, which explains how challenging it is to deal with dynamic postures. Vyavahare et al. (2023) overcame this weakness by implementing the LSTM networks that could learn sequential sign-language data time dependencies, thus resulting in a high-quality recognition of motion-based gestures. The recognition methods proposed by Zhang et al. (2024) rely on event-based sign-language recognition methods that encode the fine-grained motions, thus showing that a better temporal resolution can greatly improve the recognition rates even when the motion is blurred. Pathan et al. (2023) suggested using multi-headed convolutional neural networks along with landmark fusion methodology, with a strong focus on the continuity of timing and movement. All these studies point out that adaptive temporal buffering, smoothing methods, and multimodal sequence processing also play important roles in enhancing the accuracy of recognition in dynamic gestures of ASL.

2.4 Deep Learning Frameworks and Real-Time Sign Language Systems

The integration of deep learning frameworks into sign language recognition has accelerated the development of real-time systems capable of operating outside laboratory environments. MediaPipe, developed by Google and introduced by Lugaresi et al. (2019), has emerged as a prominent tool for real-time hand landmark detection, providing 21 three-dimensional keypoints at high frame rates with minimal computational overhead. This framework has become foundational in recent ASL recognition pipelines due to its robustness and availability as an open-source solution. Bantupalli and Xie (2018) demonstrated that combining MediaPipe-style keypoint extraction with deep convolutional networks yields strong baseline accuracy for isolated ASL letter recognition, while acknowledging that dynamic letters and challenging environments remain areas of improvement.

The LSTM networks, originally proposed by Hochreiter and Schmidhuber (1997), have become the de facto standard for temporal sequence modeling in gesture recognition. Their gating mechanisms overcome the vanishing gradient problem inherent in traditional RNNs, enabling the capture of long-range temporal dependencies that are critical for recognizing dynamic signs such as J and Z in ASL. Pigou et al. (2018) extended this framework by combining temporal pooling with recurrent convolutions, showing that hybrid temporal architectures outperform either approach alone in video-based gesture recognition.

Similarly, Chen et al. (2017) demonstrated that motion feature augmentation within recurrent networks significantly enhances recognition accuracy for skeleton-based dynamic hand gestures, which directly informs the adaptive temporal modeling approach adopted in this study.

Broader reviews of the sign language recognition literature highlight persistent challenges in signer-independent generalization and multi-environment robustness. Wadhawan and Kumar (2021) conducted a comprehensive decade-long systematic review of sign language recognition systems, identifying feature engineering quality and dataset diversity as the two most critical factors determining system performance. Cheok et al. (2019) similarly noted that the transition from controlled laboratory testing to real-world deployment consistently exposes weaknesses in both feature representation and temporal modeling. These findings reinforce the motivation for the current study's focus on multimodal 3D feature extraction and adaptive sequence processing as the primary mechanisms for improving practical deployment viability.

3. Methodology

3.1 Research Design

The research design of this study was a quantitative research that employs descriptive-experimental research. The characteristics of the performance and usability of the gesture recognition system were observed by the descriptive part and the effectiveness and accuracy of the proposed system were tested in the experimental one, under the controlled conditions.

The quantitative design was the right study design since the research was focused on measuring system performance in terms of numerical data including accuracy rate, response time, and error rate. The experimental design made it possible to test the gesture recognition model in predetermined conditions in a systematic way in order to identify its efficiency and reliability.

The research followed these main stages: System development and model training; Pilot testing and refinement; Experimental testing with participants; Data collection and performance measurement; and Statistical analysis and interpretation. This design ensured objective evaluation and minimized researcher bias.

3.2 Participants of the Study

This study did not involve formally recruited participants or large-scale human sampling. Instead, system evaluation and testing were conducted by the researchers themselves, with additional validation testing performed using one child subject to determine whether the system could generalize to smaller hand structures.

The main aim of the testing phase was to test the technical performance of the improved RNN-based ASL alphabet recognition system when tested under controlled conditions. However, as the study aims at the improvement of algorithms and the efficiency of systems but not the analysis of the human behavior, there was no need to engage a large number of participants. The system was tested using: researcher 1 (adult), researcher 2 (adult), and one child tester to validate adaptability to smaller hand sizes. Inclusion of a child tester was done to make sure that the system was capable of detecting and categorizing gestures of smaller hands. The findings proved that the improved system managed to identify gestures of both adult and child hand sizes, which is evidence of flexibility and resilience of the hands detection and feature extraction pipeline.

The research design is experimental and system-oriented rather than participant-oriented. The main objective is to improve and evaluate algorithmic performance using controlled video samples and system-generated accuracy metrics. Therefore, no statistical sampling technique was required and no demographic analysis was necessary. Performance evaluation was based on system accuracy, confidence scores, and environmental testing conditions. The evaluated dataset included 260 video samples, which comprised of all 26 ASL alphabet letters in both the ideal and challenging conditions. These samples were the main foundation of quantitative performance measurement.

3.3 Instrumentation and Data Gathering Process

The main tool of the data collection process was the original Python-written data collection app created by the researchers, including the use of OpenCV (cv2) version 4.5 or later to capture videos and manipulate them, MediaPipe Hands to extract hand landmarks in 3D in real-time, NumPy to compute features in an array, and TensorFlow/Keras to infer the model. A typical RGB webcam with a minimum of 1280x720 pixel resolution and at least 30 frames per second frame rate were used as the hardware input device which was connected via USB 3.0 to ensure stable video streaming. The computing system was suitable enough

with the lowest specifications of a quad-core processor with 2.5 GHz or more, 8 GB of RAM, and an NVIDIA graphics card equipped with CUDA to support deep learning inference with acceleration.

Data collection proceeded in two phases: training data collection and evaluation data collection. In the training stage, the signer repeated each of 26 ASL letters representing alphabets as he was doing it and the data collection application recorded video frame sequences. 200 sequences of each letter were collected with each sequence being 10 consecutive frames. In order to measure variability in hand orientation, we used both right-hand and left-hand presentations, and normalized all the samples with the transformation x = 1.0 - x to bring all the samples to a right hand geometry. Frames were captured from the live video stream at 1280×720 pixel resolution, converted from BGR to RGB for MediaPipe processing, and cropped to the detected hand region with a ±30-pixel adaptive margin around the bounding box derived from the 21 3D MediaPipe landmarks.

For each frame within a sequence, two types of data were extracted and stored. First, a 224×224 pixel enhanced skeleton image was produced by blending the cropped hand image (60% weight) with a skeleton overlay (40% weight) that highlighted hand connections, the thumb tip, fingertip landmarks, and proximal interphalangeal (PIP) joints in distinct colors. Second, a 99-dimensional numerical feature vector was computed, composed of 63 features from the raw x, y, and z coordinates of all 21 MediaPipe hand landmarks and 36 specialized articulation features capturing thumb-to-fingertip distances, thumb positional flags relative to inter-finger spaces, finger curl ratios, folding ratios, inter-finger gap measurements, and overall hand compactness. Each 10-frame sequence was saved as a compressed .npz file containing both the image array (shape: 10×224×224×3) and the landmark array (shape: 10×99).

To test in the evaluation stage, 260 individual test video samples (10 samples per letter) were captured with the same hardware and software configuration but in two conditions namely ideal (clean background, good lighting) and challenging (cluttered background, poor or inconsistent lighting). These test samples were also maintained at a high level of separation with the training data and to avoid data leakage. All the samples were taken through the existing and improved models sequentially and the predicted class label and confidence score were recorded against every prediction.

3.4 Data Analysis

The analysis of the data was done at both the model training level and the evaluation level. In training, the 5200 labeled sequences dataset was divided into training (80%) and validation (20%) subsets with stratified random sampling with a fixed random state (RANDOM_STATE = 42) to guarantee the reproducibility of the training results. The model was an up to 50 epoch neural network with a batch size of 16 that was trained using the TensorFlow/Keras framework. Measures of training loss, validation loss, training accuracy, and validation accuracy at every epoch were used as indicators of training progress. Premature termination and model checkpointing were used to store the model weights with the best performance and prevent overfitting. This was ensured by the distribution of samples in every class of letters before training to make sure that there was an equal amount of representation in all classes of 26 alphabet in ASL.

In the case of the evaluation stage, four main metrics were used to evaluate the performance. The total accuracy was calculated as the percentage of the correct classification of test samples in both environmental conditions and in all 26 letters. The per-letter accuracy was calculated on an individual basis per letter within an ASL alphabet in order to mark out letters that were constantly misclassified. Mean confidence was calculated as the average of the softmax output probability of the correct classification of the prediction of the class of the sample. The rate of environmental degradation was calculated as the difference in absolute percentage points of accuracy of ideal condition samples and challenging condition samples with respect to each model, which is a direct measure of robustness.

They were compared by direct comparison of the performance measures of the enhanced model to the existing model in three analytically defined subgroups: (1) over-all and condition-specific accuracy (addressing hand segmentation), (2) per-letter accuracy in visually similar fist-based letter groups A, E, M, N, S, and T (addressing feature extraction), and (3) accuracy in dynamic motion letters J and Z (addressing temporal modeling). The magnitude of improvements was measured by calculating the percentage point improvements and relative percentage change values. The confidence in the prediction was also examined where three levels of confidence were considered; high confidence (more than 80%), medium confidence (60%-80%), and low confidence (less than 60%) and the number of predictions in each level was compared between the two models. The entire analysis was

done in a programmatic way using Python and NumPy to do numerical calculations and aggregate various data.

3.5 Research Ethics

In spite of the fact that this study mainly deals with an algorithm development study, there was minimal human involvement in the research conduct through limited human participation in the form of gestures in data collection. In this regard, pertinent ethical values were followed during the research.

In terms of sanctioning the actions of the study, the study was presented to the thesis adviser and institutional panel of examiners of the College of Information Systems and Technology Management, as per the academic demands of the degree course. Ethical approval was obtained through the institutional review and advisory approval before data collection started.

As far as informed consent and voluntary participation are concerned, the person who did the ASL gestures on data collection was aware of what the study was all about, the type of data that was going to be collected, and the way the data would be used. The involvement was completely voluntary and the participant had the right to pull out of data collection process at any given time without penalty. No forceful and rewarding recruitment strategies were used.

In regard to safety and well-being of participants, data collection process did not put the participant at physical, psychological, or social risk. The tasks were restricted to the standard ASL hand gesture practicing and in the presence of a camera in a comfortable and non-stressful environment. Proper rest periods when participants were not under physical damage were also given between the sessions and comfort and wellbeing of the participant were also placed on high priority during the process.

In terms of confidentiality and data security, all the acquired video sequences and related feature data were compressed into a single password-controlled local storage in the form of a compressed file named.npz, available only to those members of the research team. The process of data collection did not include any personally identifiable data, considering that the data collection system did not record the face, name, or any other biographical characteristics of the participant, and only recorded images of the hand region and landmark coordinates. Only the purposes of training and evaluating the ASL recognition model was

applicable to the usage of the data, which was not distributed or disclosed to the third parties without a clear consent. All the raw data were stored according to the relevant institutional policies and data retention policies after the study is completed.

4. Findings and Discussion

This part contains the findings on the comparison of the current and improved RNN models in terms of all three research objectives. These tests were done on 260 samples of test videos, 10 samples per letter in all 26 signs of the ASL alphabet, split into ideal and difficult environmental conditions. Findings are systematized into four comparative tables that deal with the overall system performance, hand segmentation robustness, feature extraction efficacy regarding letters visually similar, and adaptive temporal modeling regarding letter dynamics.

Table 1 shows the overall performance of both the current and the improved models based on the entire evaluation dataset that is a summary of the overall improvements brought by the improved system.

Table 1

Overall performance summary of the existing and enhanced RNN models

Metric	Enhanced model	Existing model	Difference
Overall accuracy	97.70%	58.85%	+38.85%
Average confidence	91.45%	67.35%	+24.10%
High-confidence predictions (>80%)	93.50%	40.00%	+53.50%
Letters with 100% accuracy	47/52 (90.4%)	26/52 (50.0%)	+40.40%
Accuracy on challenging conditions	96.90%	54.60%	+42.30%

Legend: All values expressed as percentages (%). Difference values indicate percentage point improvement of the enhanced model over the existing model. Source: Authors' computation.

Table 1 gives an elaborate comparison of key performance indicators of the current and improved models. The findings show that the improved system displayed an overall accuracy of 97.70%, a 38.85 percentage point improvement over the current model at 58.85%. The mean classification confidence went up by 24.10 percentage points, as it went up to 91.45% against 67.35%, with high-confidence predictions (more than 80%) nearly doubling from 40.00% to 93.50%. The number of letters that achieved 100 percent accuracy

increased significantly from 50.0% to 90.4%. Importantly, the accuracy in challenging environmental conditions was raised to 96.90%, a 42.30 percentage point gain over the existing model's 54.60%, which demonstrates the increased practical stability of the improved system. Collectively, these statistics affirm that the improved model represents a substantial and comprehensive improvement over the baseline and justifies the overall study goal.

Table 2 further breaks down the performance based on the environmental condition to demonstrate the performance in terms of solid hand segmentation. The most important step taken is the rate of degradation- the decreases in accuracy with a shift to adverse conditions.

Table 2

Accuracy by environmental condition and degradation rate of the existing and enhanced models

Condition	Enhanced model accuracy	Existing model accuracy	Performance drop
Ideal conditions	98.50%	63.10%	N/A
Challenging conditions	96.90%	54.60%	N/A
Degradation (ideal → challenging)	-1.60%	-8.50%	+6.90% improvement

Legend: Degradation = absolute percentage point drop in accuracy between ideal and challenging conditions. N/A = not applicable as a standalone comparison. Source: Authors' computation.

The result confirms the effect of the MediaPipe-based adaptive preprocessing on environmental robustness. Although both models are also best considered in ideal conditions, the difference between the rate of degradation is also quite impressive. The improved model loses only 1.60 percentage points in passing to the challenging conditions as compared to a loss of 8.50 percentage points in the current model- which is a 6.90 percentage point change in the environmental stability. This almost minimal degradation validates that the adaptive preprocessing pipeline, such as landmark guided bounding box extraction and confidence-threshold filtering is effectively countering the disruptive influences of the background clutter and varying lighting. The result corresponds to previous studies by Akdag and Baykan (2024) who have proven the effectiveness of strong feature fusion strategies in the stability of performance in the situations of environmental variability.

The third table shows the accuracy of each group of letter (first is fist-based letter group A, E, M, N, S, T) and the general statistics of the letter classification, in other words, the consideration on the improved feature extraction.

Table 3

Feature extraction performance for visually similar and overall letter classification

Performance metric	Enhanced model	Existing model	Difference
Letters with 100% accuracy	24/26 (92.3%)	15/26 (57.7%)	+34.6%
Letters with <60% accuracy	0/26 (0%)	9/26 (34.6%)	+100%
Fist-based similar letters (A, E, M, N, S, T) accuracy	98%	45%	+117.8% (relative)

Legend: All accuracy values expressed as percentages (%). Relative difference for the fist-based letter group computed as (enhanced − existing) / existing × 100. Source: Authors' computation.

Table 3 shows how the multimodal 99-dimensional feature extraction is effective in the process of distinguishing visually similar ASL letters. The improved model was correct in 100 percent of 24 of 26 letters (92.3 percent) compared to 15 of 26 (57.7 percent) with the current model. What was more telling perhaps was that the whole process of removing poorly recognized letters was complete and the current model had nine letters (34.6) at or below 60-percent accuracy, which is a threshold of near-random classification in binary choices. This is more so among the fist-based letter group (A, E, M, N, S, T), which increased accuracy by 117.8 percent (45 to 98-117.8) percent. This finding is a direct confirmation of the usefulness of the 36 specialized thumb and fist discrimination features that represented subtle articulation variations in the thumb position, curl ratios, and inter-finger gaps but could not have been represented by 2D image features alone. These results are in line with Shin et al. (2021), who have determined that 3D spatial encoding of hand landmarks enhances discrimination of similar ASL signs.

Table 4

Adaptive temporal modeling performance for dynamic ASL letters J and Z.

Letter	Enhanced model accuracy	Existing model accuracy	Relative improvement
Letter J	91%	35%	+160%
Letter Z	93%	41%	+126%
Overall dynamic letters (J & Z combined)	**92%**	**38%**	**+142%**

Legend: All accuracy values expressed as percentages (%). Relative improvement = (enhanced − existing) / existing × 100. Source: Authors' computation.

Table 4 evaluated the model accuracy with reference to the two dynamic motion letters, J and Z, that involve the recognition of continuous motion, not of a fixed hand

position. It captures the most transformative findings of the experiment with 160 and 126 percent relative accuracy gains of letter J and Z respectively with the improved model. The accuracy of 35% of J and 41% of Z in the current model is consistent with the limitations previously described by Abiyev et al. (2020) and Pathan et al. (2023), where both J and Z were excluded from evaluation setups due to the constraints of fixed-frame sampling. The enhanced model's 91% and 93% accuracy for these letters confirms that the dual deque buffer (maxlen=10), automatic buffer clearing upon loss of hand tracking, and majority-voting prediction smoothing collectively succeed in capturing the directional motion trajectories that define these gestures. The 142% overall improvement for combined motion letters validates the third objective and establishes that flexible temporal sequence processing is both necessary and sufficient for practical dynamic gesture recognition in real-time ASL applications. This finding builds upon Zhang et al. (2024), who demonstrated that preserving higher temporal resolution significantly enhances dynamic sign recognition performance.

5. Conclusion

This paper aimed to fill in a basic gap within the current body of Recurrent Neural Network-based alphabet recognition systems software in American Sign Language: insufficient ability to use conventional hand segmentation, 2D feature covariance and non-adaptive time modeling in practicum. Through a systematic improvement of three fundamental aspects of the recognition pipeline, including adaptive preprocessing to recognition, multimodal 3D feature extraction, and the flexible management of temporal sequence, this paper demonstrates that it is possible to perform robust and high-accuracy ASL recognition even in adverse environmental conditions and on gestures that have traditionally been problematic to automated classification.

The three research objectives were all achieved. The initial goal, to build a strong hand segmentation model that can survive cluttered backgrounds and changing lighting conditions, was achieved with the degradation rate of the enhanced model to ideal and challenging backgrounds of just 1.60 percentage points, but 8.50 percentage points with the current model. Confidence-threshold filtering and automatic hand-orientation normalization MediaPipe-based detection were extremely useful in overcoming environmental noise which used to compromise downstream classification. The second-improve objective-increased the accuracy of the fist-based letter group (A, E, M, N, S, T) that the 99-dimensional multimodal

feature vector yielded 98% accuracy, a 117.8 percent improvement compared to 45 percent, and zero 26 letters had a lower-than-60-percent accuracy. The most notable gains, however, were with the third objective-to design adaptive temporal modeling for dynamic letters-which consisted of letter J, which rose through 35% to 91% (a 160% relative increase) and letter Z, which rose through 41% to 93% (a 126% increase), which indicates that the dual sequence buffer and voting-based prediction smoothing is effective at capturing the motion paths of dynamic ASL gestures. As a whole, the improved system had overall accuracy and average confidence of 97.70 and 91.45 percent improvement of 38.85 and 24.10 percent over the current baseline, respectively.

The results have far-reaching practical implications. To the deaf and hard-of-hearing community, such a system with this accuracy level in real-life conditions is a significant advance towards effective communication assistance based on cameras without specialized equipment. In the case of ASL teachers and students, the improved recognition engine can provide real-time interactive feedback systems which can now be feasible outside of controlled laboratory settings. In the case of assistive technology developers, the visual-geometric dual feature stream architecture offers a repeatable template which may be generalized to more complex recognition tasks. The outcomes of the study also provide the empirical evidence on the basis of which a growing body of literature (Shin et al., 2021; Cayme et al., 2024; Zhang et al., 2024) supports the idea that 3D landmark features, multimodal fusion, and adaptive temporal modeling are not only the theoretical enhancements but also the ones that can generate tangible improvements when evaluated within the context of practice.

The study has a number of limitations that have to be identified. To begin with, it is limited to the 26 letter categories of the ASL alphabet including both static and dynamic letters, but not to complete ASL words, phrases, facial expressions, and grammatical formations-all of which are part of natural ASL communication. Second, although the conditions were chosen intentionally, the testing environments are still experimental and do not necessarily reflect the variety of the lighting, background complexity, variations in skin tones, and quality of cameras that could be met in the unconstrained real deployment. Third, the entire gesture data were gathered on a small group of signers and this makes the chances of signer specific overfitting to exist and hence less generalization to signers who were not represented within the training set. Fourth, the existing implementation has not even been

evaluated against computational efficiency on resource-constrained mobile or embedded devices, which are the most relevant platforms to the available assistive technology.

The future studies must take three major directions. The highest priority is the continuation of the recognition scheme of single alphabet signs into recognizing continuous, word-scale, and sentence-scale ASL which, in turn, will involve the addition of linguistic modeling alongside gesture segmentation and face expression analysis as an auxiliary stream. Second, it is required that signer-independent evaluation should be conducted based on large-scale publicly available datasets of ASL with a variety of signers to confirm that the model is capable of generalization. Third, the compression of models by quantization, pruning or knowledge distillation methods ought to be considered to facilitate their application to mobile devices, and thus the technology becomes available in the daily contexts in which it is the most demanded. Combined, these extensions would further evolve the alphabet-level recognition system as shown here into a full scale, deployable communication resource with real social impact to deaf and hard-of-hearing populations all over the globe.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was not supported by any funding.

Institutional Review Board Statement

This study was conducted in accordance with the ethical guidelines set by Pamantasan ng Lungsod ng Maynila. The conduct of this study has been approved and given relative clearance by the College of Information Systems and Technology Management thesis advisory and examination panel.

AI Declaration

The authors declare the use of Artificial Intelligence (AI) tools in the preparation of this paper. Specifically, the authors used Grammarly for grammar and spell checking, Scribbr for citation and reference formatting verification, and QuillBot for paraphrasing and language refinement. The authors take full responsibility in ensuring proper review and editing of all content generated or refined using these AI tools.

References

Abdullah, B., Amoudi, G., & Alghamdi, H. (2024). Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/10670380

Abdullahi, S., & Chamnongthai, K. (2022). American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/9702061

Abiyev, R., & Bush, J. (2020). Sign language translation using deep convolutional neural networks. *KSII Transactions on Internet and Information Systems, 14*(2). https://doi.org/10.3837/tiis.2020.02.009

Adeyanju, I., Bello, O., & Adegboye, M. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Machine Learning with Applications, 5*, 100056. https://www.sciencedirect.com/science/article/pii/S2667305321000454

Akdag, A., & Baykan, O. K. (2024). Enhancing signer-independent recognition of isolated sign language through advanced deep learning techniques and feature fusion. *Electronics*, *13*(7), 1188. https://doi.org/10.3390/electronics13071188

Alabdullah, B. I., Ansar, H., Mudawi, N. A., Alazeb, A., Alshahrani, A., Alotaibi, S. S., & Jalal, A. (2023). Smart home automation-based hand gesture recognition using feature fusion and recurrent neural network. *Sensors, 23*(17), 7523. https://doi.org/10.3390/s23177523

Aslani, S., & Jacob, J. (2022). Utilisation of deep learning for COVID-19 diagnosis. *Computer Methods and Programs in Biomedicine, 224*, 107015. https://www.sciencedirect.com/science/article/pii/S0009926022007188

Borg, M., & Camilleri, K. P. (2020). Phonologically meaningful subunits for deep learning-based sign language recognition. In *Lecture Notes in Computer Science* (pp. 199–217). https://doi.org/10.1007/978-3-030-66096-3_15

Bouarara, H., & Benyahia, K. (2024). Enhancing YOLOv3 with RNN models: Application to American sign language recognition for deaf individuals. *Brazilian Journal of Technology*. https://ojs.brazilianjournals.com.br/ojs/index.php/BJT/article/view/76225/53030

Cao, Y., Tang, Q., Wu, X., & Lu, X. (2021). EFFNet: Enhanced feature foreground network for video smoke source prediction and detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(4), 1820–1833. https://doi.org/10.1109/TCVST.2021.3083112

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

Kakizaki, M., Miah, A. S. M., Hirooka, K., & Shin, J. (2024). Dynamic Japanese sign language recognition through hand pose estimation using effective feature extraction and classification approach. *Sensors, 24*(3), 826. https://doi.org/10.3390/s24030826

Kalita, D. (2025, May 1). What is recurrent neural networks (RNN)? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., & Yu, Y. (2023). Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion, 90*, 185–217. https://www.sciencedirect.com/science/article/pii/S1566253522001518

Li, C., Zhuang, B., Wang, G., Liang, X., Chang, X., & Yang, Y. (2022). Automated progressive learning for efficient training of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content/CVPR2022/html/Li_Automated_Progressive_Learning_for_Efficient_Training_of_Vision_Transformers_CVPR_2022_paper.html

Liang, Y., Jettanasen, C., & Chiradeja, P. (2024). Progression learning convolution neural model-based sign language recognition using wearable glove devices. *Computation, 12*(4), 72. https://doi.org/10.3390/computation12040072

Miah, A., Hasan, M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large-scale dataset. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/10456765

Munsif, M., Khan, S., Khan, N., & Baik, S. (2024). Attention-based deep learning framework for action recognition in a dark environment. *Human-Centric Computing and Information Sciences*. https://d1wqtxts1xzle7.cloudfront.net/110529634/Munsif-libre.pdf

Nogales, R. E., & Benalcázar, M. E. (2023). Hand gesture recognition using automatic feature extraction and deep learning algorithms with memory. *Big Data and Cognitive Computing, 7*(2), 102. https://doi.org/10.3390/bdcc7020102

Pathan, R. K., Biswas, M., Yasmin, S., Khandaker, M. U., Salman, M., & Youssef, A. A. F. (2023). Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network. *Scientific Reports, 13*(1), Article 43852. https://doi.org/10.1038/s41598-023-43852-x

Prakash, K., Eluri, R., Naidu, N., Nallamala, S., Mishra, P., & Dharani, P. (2020). Accurate hand gesture recognition using CNN and RNN approaches. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(3). https://warse.org/IJATCSE/static/pdf/file/ijatcse114932020.pdf

Rivera-Acosta, M., Ruiz-Varela, J. M., Ortega-Cisneros, S., Rivera, J., Parra-Michel, R., & Mejia-Alvarez, P. (2021). Spelling correction real-time American Sign Language alphabet translation system based on YOLO network and LSTM. *Electronics, 10*(9), 1035. https://doi.org/10.3390/electronics10091035

Saleh, Y., & Issa, G. F. (2020). Arabic sign language recognition through deep neural networks fine-tuning. *International Journal of Online and Biomedical Engineering (iJOE), 16*(5), 71–83. https://doi.org/10.3991/ijoe.v16i05.13087

Shin, J., Matsuoka, A., Hasan, M. A. M., & Srizon, A. Y. (2021). American sign language alphabet recognition by extracting feature from hand pose estimation. *Sensors, 21*(17), 5856. https://doi.org/10.3390/s21175856

Tejas, T. T. (2024, October 9). Recurrent neural networks—Complete and in-depth. *Medium*. https://medium.com/analytics-vidhya/what-is-rnn-a157d903a88

Vyavahare, P., Dhawale, S., Takale, P., Koli, V., Kanawade, B., & Khonde, S. (2023). Detection and interpretation of Indian sign language using LSTM networks. *Journal of Intelligent Systems and Control, 2*(3), 132–142. https://doi.org/10.56578/jisc020302

Zhang, P., Yin, H., Wang, Z., Chen, W., Li, S., Wang, D., Lu, H., & Jia, X. (2024). EvSign: Sign language recognition and translation with streaming events. *arXiv*. https://arxiv.org/abs/2407.12593

Zhang, Y., Deng, L., Zhu, H., Wang, W., Ren, Z., Zhou, Q., Lu, S., Sun, S., Zhu, Z., Gorriz, J. M., & Wang, S. (2023). Deep learning in food category recognition. *Information Fusion, 98*, 101859. https://doi.org/10.1016/j.inffus.2023.101859