Journal of Computer Science

Research Article

Recognizing Sign Language Gestures Using a Hybrid Spatio-Temporal Deep Learning Model

Meryem Cherrate, My Abdelouahed Sabri, Ali Yahyaouy and Abdellah Aarab

Department of Computer Science, Faculty of Sciences Dhar-Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco

Article history
Received: 08-02-2025
Revised: 29-06-2025
Accepted: 06-07-2025

Corresponding Author:
Meryem Cherrate
Department of Computer Science, Faculty of Sciences Dhar-Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco
Email: meryem.cherrate@usmba.ac.ma

Abstract: Recognizing gestures in American Sign Language (ASL) from video data presents significant challenges due to the intricate combination of hand gestures, facial cues, and body motion. In this work, we introduce a hybrid deep learning framework that integrates Convolutional Neural Networks (CNNs) for extracting spatial characteristics with Long Short-Term Memory (LSTM) networks for capturing temporal sequences. The model was trained and evaluated on a subset of 25 classes from the WLASL dataset, a comprehensive video collection comprising over 2,000 labeled ASL signs. Achieving an accuracy of 96%, the proposed system demonstrates superior performance compared to traditional methods. These findings underscore the strength of spatio-temporal modeling in sign language recognition. With a design geared toward scalability and real-time deployment, the approach shows strong potential to support communication and accessibility for individuals with hearing impairments. Future developments will aim to mitigate class imbalance, broaden applicability to other sign languages, and assess the benefits of Transformer-based models for enhanced recognition.

Keywords: American Sign Language, Gesture Recognition, WLASL Dataset, Deep Learning, Communication, Assistive Technology

Introduction

American Sign Language (ASL) is a fully developed visual language widely used by deaf and hard-of-hearing communities in the United States and beyond. It relies on a complex combination of hand shapes, movements, facial expressions, and body posture to express meaning and grammatical structure naturally and expressively. Despite its complexity and expressive power, the absence of mutual intelligibility between ASL users and non-signers continues to hinder communication, contributing to social, educational, and professional disparities. The development of automated, real-time ASL recognition and translation systems presents a promising avenue for enhancing accessibility and fostering inclusive interaction between individuals who are hearing and those who are deaf or hard of hearing (Pun *et al*., 2011). Progress in Artificial Intelligence (AI) and computer vision has opened up new opportunities for automating ASL recognition. By leveraging video-based datasets, machine learning models can now analyze and classify complex gesture sequences (Abdelouahed *et al*., 2022; Adeyanju *et al*., 2021). Despite these advancements, the field of ASL recognition faces significant challenges, including the variability of gestures across individuals, nuanced communication styles, and the need to capture spatiotemporal dynamics from unstructured video data. Traditional approaches, such as hardware-based solutions using connected gloves and sensors, have demonstrated limited scalability and usability (O'Connor *et al*., 2017; Lee and Lee, 2018). Recent advances have emphasized software-driven approaches leveraging deep learning to tackle the challenges inherent in dynamic gesture recognition. Specifically, models that integrate Convolutional Neural Networks (CNNs) to capture spatial details along with Long Short-Term Memory (LSTM) networks for managing temporal relationships have shown remarkable effectiveness in understanding intricate, time-dependent gestures (Cherrate *et al*., 2025a; Hafeez *et al*., 2024). However, challenges such as imbalanced datasets, inter-individual variability, and annotation inconsistencies remain unresolved, necessitating further research (Xiong *et al*., 2025; Ur Rehman *et al*., 2022). This study proposes a hybrid deep learning framework that combines CNNs and LSTMs to process video sequences comprehensively. The

integration of these architectures enables the system to capture both spatial details (e.g., hand shapes and positions) and temporal dependencies (e.g., motion trajectories and gesture sequences). To validate the effectiveness of our approach, we utilize the Word-Level American Sign Language (WLASL) dataset, a large-scale video corpus containing over 100,000 annotated videos spanning more than 2,000 ASL signs (Baskoro, 2025).

This study presents a hybrid deep learning framework for American Sign Language (ASL) recognition, designed to address the combined spatial and temporal complexity of sign gestures. Our primary contribution is the development of a robust CNN-LSTM architecture that captures both the spatial features and sequential dynamics inherent in ASL. The model is trained and evaluated using the WLASL dataset, enabling high recognition accuracy across a diverse subset of signs. Benchmarking against existing methods demonstrates a significant performance improvement, with our approach achieving 96% recognition accuracy on 25 sign classes, highlighting its superior accuracy and robustness. Beyond technical metrics, we discuss the practical implications of this work, including its potential application in next-generation accessibility tools, interactive educational resources, and assistive technologies for the deaf and hard-of-hearing community.

This paper is structured as follows: We begin with a review of recent developments in sign language recognition, focusing on both hardware and software strategies with an emphasis on deep learning architectures. We then present the proposed method, describing dataset preparation, the hybrid CNN-LSTM model, and the training and evaluation protocols. The subsequent section reports the experimental results, followed by a comparative analysis with existing approaches. The paper concludes with a summary and recommendations for future research directions.

Literature Review

Sign language recognition has become an essential research field, as it plays a key role in enhancing communication and social inclusion for deaf and hard-of-hearing people. The World Federation of the Deaf reports that more than 70 million individuals worldwide use sign language as their primary means of communication, and nearly 80% of them live in developing countries. By 2050, it is estimated that 2.5 billion people will be affected by hearing impairments to varying degrees, with at least 700 million requiring rehabilitation services (WHO, 2025). Researchers have looked into both software- and hardware-based solutions to these problems.

Solutions Based on Hardware

Early attempts in sign language recognition focused on using wearable devices such as connected gloves and sensors to track hand and finger movements (Ji *et al*., 2023). While these approaches provided some success in recognizing gestures, they lacked scalability and practicality for real-world applications. Moreover, they could not capture full-body gestures or facial expressions, both of which are essential components of sign language communication.

Software-Based Solutions

Recent progress in machine learning and computer vision has enabled software-based methods that utilize image and video datasets for recognizing gestures. Static image datasets, such as the Kaggle ASL dataset (Cherrate *et al*., 2025b), have been used for classifying hand signs representing the alphabet, and ASL Finger spelling (Geislinge, 2021; Pugeault and Bowden, 2011). However, these datasets are limited to static gestures and fail to capture the temporal dynamics required for recognizing words and phrases. Dynamic datasets, such as the ASL Lexicon Video Dataset (Athitsos *et al*., 2018; Kataoka and Yoon, 2024), offer video sequences annotated with specific words, enabling more realistic modeling of ASL gestures. However, challenges such as gesture variability, inter-individual differences, and annotation inconsistencies persist.

The WLASL dataset (Li *et al*., 2020; Shen *et al*., 2024) represents a significant advancement in this field. With over 100,000 annotated videos spanning more than 2,000 signs, it provides a comprehensive resource for training deep learning models that can recognize spatio-temporal patterns. The diversity and scale of WLASL make it a valuable tool for developing robust and scalable ASL recognition systems.

Deep Learning in Gesture Recognition

Recent progress in deep learning has further improved the accuracy of sign language recognition systems. CNNs are widely used for extracting spatial features from images, while LSTMs and other recurrent networks model temporal dependencies (Toro-Ossaba *et al*., 2022; Huang and Chouvatut, 2024; Sivaraman *et al*., 2024; Omarkhan *et al*., 2021; Shi, 2015; Tran *et al*., 2015). Transformer-based models have also been explored, treating video sequences as temporal data, and have achieved promising results (Li *et al*., 2020; Hafeez *et al*., 2024). However, challenges such as data imbalance, computational complexity, and generalization to real-world scenarios remain active areas of research.

Comparative Overview

Table 1 presents a comparison of different sign language recognition approaches, highlighting their datasets, methods, and associated challenges.

Table 1: Comparison of Sign Language Recognition Approaches: Datasets, Techniques, and Challenges

Approach	Datasets Used	Techniques	Challenges
Hardware-based	Not dataset-dependent	Connected gloves and sensors	Limited scalability; no full-body or facial gesture recognition
Kaggle ASL Dataset	87,028 images of ASL alphabet	CNN-based classification	Static gestures only; no temporal modeling
ASL Finger Spelling	Over 500 images per sign	CNN-based recognition	Complex backgrounds; limited word recognition
ASL Lexicon Video	Dynamic videos of ASL words and phrases	CNN for visual features	Annotation inconsistencies; gesture variability
WLASL Dataset	2,000 signs; 100,000+ videos	CNN + RNN/Transformer	Temporal dependencies; large-scale training complexity
Proposed Approach	Subset of WLASL (25 classes; 175 videos)	Hybrid CNN + LSTM architecture	Efficient spatio-temporal modeling with 96% recognition accuracy

Despite the progress achieved, limitations such as dataset imbalance, inter-class variability, and computational overhead persist. Our study addresses these challenges by leveraging a hybrid CNN-LSTM model trained on a subset of the WLASL dataset, achieving significant improvements in accuracy and scalability.

Methodology Research

The proposed method for recognizing American Sign Language uses deep learning techniques to effectively process and understand video sequences. This part furnishes an in-depth explanation of the choices made for the dataset, the preprocessing steps taken, the architecture of the model, the training approach, and the metrics used for evaluating the system's performance.

Dataset

This study utilizes the Word-Level American Sign Language (WLASL) dataset, a large-scale video corpus containing over 100,000 annotated videos covering more than 2,000 distinct ASL signs. For this work, we selected a subset of 25 classes, totaling 175 videos, to enable controlled experimentation and effective evaluation of our proposed CNN-LSTM architecture.

The subset was carefully chosen based on two primary criteria: Semantic diversity and data quality. Specifically, the selected signs span five meaningful thematic categories: “Emotions, family, colors, politeness, and greetings”, which include commonly used signs and represent a broad range of hand configurations, motion dynamics, and semantic contexts. Furthermore, the selected classes were those for which visually clear and sufficiently numerous samples were available, ensuring a balanced and reliable setup for training and testing.

This focused selection allowed us to fine-tune the model and assess its ability to learn spatio-temporal features effectively. The current study serves as a foundational phase for future research involving larger subsets or the full WLASL dataset, where we aim to address additional challenges such as class imbalance, signer variability, and scalability of the proposed architecture.

Dataset Preparation

Preprocessing: Each video was split into frames, resized to 64×64 pixels, and normalized for uniform input.

To improve model generalization, data augmentation methods such as horizontal flipping and random cropping were used.

Data Split: To ensure a fair assessment of model performance across classes, the dataset was split into 80% training and 20% testing.

The characteristics of the selected ASL dataset subset are summarized in Table 2.

Table 2: Summary of the Selected ASL Dataset Subset

Category	Class	Class ID	Number of videos
Emotions	Angry	0	12
	Confused	6	6
	Happy	12	9
	Sad	20	9
Family	Family	7	7
	Father	8	7
	Son	22	7
	GrandFather	9	9
	GrandMother	10	9
	Aunt	1	5
	Uncle	23	7
	Brother	4	6
	Sister	21	7
Colors	Black	2	5
	Blue	3	7
	Brown	5	6
	Green	11	7
	Orange	15	5
	Pink	16	6
	Purple	18	6
	Red	19	7
	Yellow	24	5
Politeness	Please	17	7
	Help	14	9
Greetings	Hello	13	5

Model Architecture

The proposed model combines Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks to capture significant spatio-temporal characteristics from video sequences. Its architecture is composed of the following key components.

Spatiotemporal Feature Extraction

A ConvLSTM2D layer with 128 filters is employed to simultaneously extract spatial features from individual frames and temporal relationships across frames.

Batch Normalization is applied to stabilize and accelerate training.

A Dropout Layer mitigates overfitting by deactivating random neurons during training.

Temporal Modeling: The outputs of the ConvLSTM2D layers are passed to a Gated Recurrent Unit (GRU) with 64 units, capturing sequential dependencies within the video.

Classification Layer: Fully connected layers process the GRU output to produce high-level features.

A Softmax Activation Function generates probability distributions for the 25 classes.

Model Workflow

The proposed architecture is illustrated in Fig. 1.

Training Process

The model underwent training for 50 epochs utilizing an NVIDIA GPU with a subset of the WLASL dataset. Early stopping was implemented with a patience parameter set to 5 epochs, determined by validation loss. The hyperparameters used in training are as follows:

- Learning Rate: 0.001
- Batch Size: 32
- Optimizer: Adam
- Loss Function: Categorical Cross-Entropy

Evaluation Metrics

To evaluate the performance of the model, several standard classification metrics were employed.

Precision measures the proportion of correct predictions among all positive predictions, with its mathematical definition provided in Equation 1:

$$Precision = \frac{TP+FP}{TP} \quad (1)$$

Recall reflects the model's ability to correctly identify all relevant positive instances, as detailed in Equation 2:

$$Recall = \frac{\sum_{i=1}^{n} recall\ i}{n} \quad (2)$$

The F1-score, representing the harmonic mean of precision and recall, is particularly useful for datasets with imbalanced classes (Eq. 3):

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

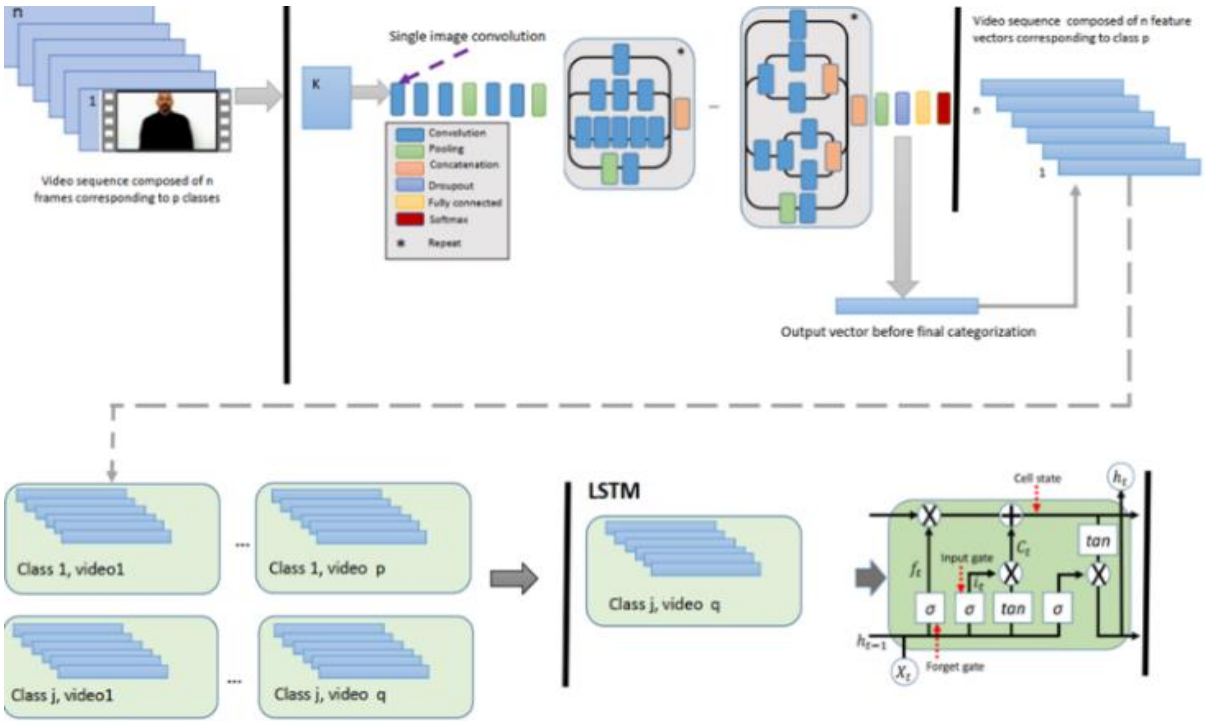

Fig. 1: Process of the Suggested Model for ASL Recognition

Accuracy provides the overall proportion of correctly classified samples relative to the total number of samples, as defined in Equation 4:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Where:

TP (True Positives) : Accurate predictions for the positive categories
TN (True Negatives) : Accurate predictions for the negative categories
FP (False Positives) : Incorrect predictions for the positive categories
FN (False Negatives) : Incorrect predictions for the negative categories

These metrics give an in-depth perspective on the model's effectiveness, assisting in assessing its advantages and drawbacks in identifying ASL gestures.

Results and Discussion

This section presents the results of the proposed CNN-LSTM model when applied to the WLASL dataset. The assessment emphasizes classification effectiveness, training dynamics, and contrasts with current methods. Additionally, we explore the significance of these results and point out potential areas for enhancement.

Dataset Description

For the purpose of recognizing ASL and converting gestures into natural language, we utilized a subset of the WLASL dataset consisting of 25 classes and 175 videos. The data was split into 80% for training and 20% for testing, maintaining a balanced distribution across classes to ensure accurate evaluation. Additional information can be found in the “Dataset” section.

Proposed Approach

The suggested approach integrates Convolutional Neural Networks (CNN) for extracting spatial features and Long Short-Term Memory (LSTM) networks for modeling temporal aspects, allowing it to accurately capture the spatio-temporal relationships present in ASL gestures. The sections titled "Model Architecture" and “Training Process” provide a comprehensive overview of the model design and training methodology.

Results

Table 3 Shows the detailed performance of the proposed model in terms of precision, recall, F1-score, and accuracy for each class, as well as the macro and weighted averages.

Table 3: Measures of Performance of Each Class

	Precision	Recall	F1-Score
0	0.86	1.00	0.92
1	1.00	0.50	0.67
2	1.00	0.86	0.93
3	0.90	1.00	0.95
4	0.88	0.92	0.93
5	1.00	0.60	0.75
6	0.92	0.85	0.92
7	0.94	1.00	0.96
8	0.87	1.00	0.91
9	0.94	0.78	0.85
10	1.00	0.82	0.92
11	0.89	1.00	0.99
12	0.92	0.90	0.88
13	0.94	0.72	0.78
14	1.00	0.86	0.93
15	0.95	1.00	0.92
16	0.89	0.98	0.97
17	1.00	0.58	0.73
18	1.00	0.89	0.92
19	0.90	0.97	0.97
20	0.78	1.00	0.92
21	1.00	0.98	0.92
22	1.00	0.89	0.94
23	0.91	1.00	0.96
24	1.00	0.97	0.94
Accuracy			0.96
Marco avg	0.94	0.84	0.87
Weighted avg	0.93	0.92	0.91

The model attained an overall accuracy of 96%, indicating good performance in recognizing ASL movements. The classification model shows satisfactory overall performance with a high accuracy of 96%, reflecting a good ability to correctly predict the majority of classes. Weighted average metrics show precision of 93%, recall of 92%, and F1-score of 91%, demonstrating balanced performance across majority and minority classes.

However, analysis of performance by class reveals significant variations. Classes such as 7, 11, 19, 23, and 24 show F1-scores above 0.95, indicating that the model recognizes these classes with high precision and near-perfect recall. In contrast, some classes, such as 1, 5, 9, and 17, show lower F1-scores, due to reduced recall (≤ 0.78), suggesting difficulties for the model to effectively detect these signs. Class 1, for example, has a particularly low recall of 0.50 despite perfect precision, indicating a potential problem of data imbalance or visual similarities with other classes. The macro averages (Precision: 94%, Recall: 84%, F1-score: 87%) confirm that the performance of minority or difficult classes slightly drags down the overall metrics.

The training and validation results are illustrated through accuracy and loss curves. The first graph depicts the progression of training and validation accuracy across epochs, while the second shows the corresponding training and validation loss trends. Together, these curves offer valuable insights into the model’s learning dynamics, stability, and generalization capability.

The evolution of training and validation accuracy is presented in Fig. 2.

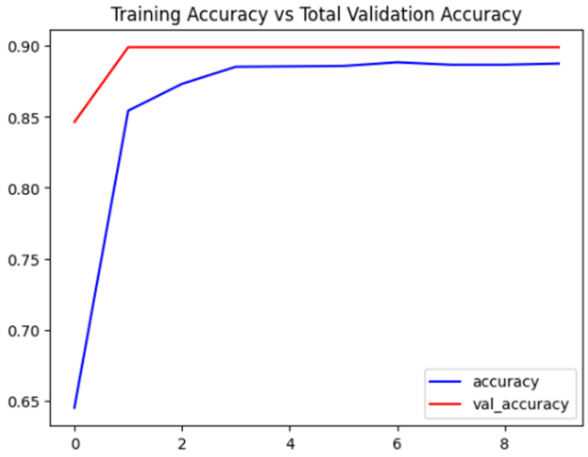

Fig. 2: Training Accuracy vs Total Validation Accuracy

As shown in Fig. 3, both training and validation loss decrease steadily, indicating that the model is learning effectively without overfitting.

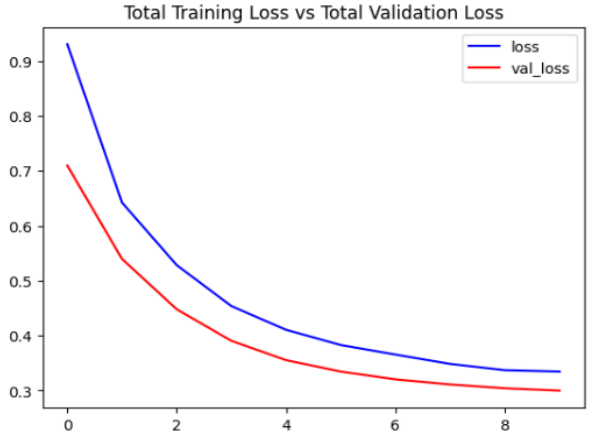

Fig. 3: Evolution of Training vs. Validation Loss over Epochs

The results obtained during the training and validation phases on the 25 classes from the WLASL dataset demonstrate consistent improvement in both accuracy and loss over the epochs. In the first graph, the training accuracy (blue curve) steadily increases, reaching approximately 90%, while the validation accuracy (red curve) quickly converges to a similar level. This indicates strong generalization without significant overfitting. This observation is further supported by the second graph, where the training loss (blue curve) and validation loss (red curve) decrease smoothly and converge to similar values. The minimal gap between the curves, both for accuracy and loss, suggests that the CNN-LSTM model is effectively trained to capture the spatial and temporal patterns present in the videos of the dataset. These results highlight the efficiency of the model architecture and training strategy in recognizing the 25 classes of signs within the context of ASL.

To further analyze the performance of our model, we present the confusion matrix for the 25 ASL classes in Fig. 4. The confusion matrix illustrates the distribution of correct and incorrect predictions across all classes. Most predictions are concentrated along the diagonal, indicating correct classifications. However, some confusion is observed between visually similar signs, particularly in classes such as [e.g., Class 1 vs Class 5], which may share overlapping visual features. This analysis highlights areas where the model could benefit from additional data or attention mechanisms to better distinguish between similar gestures.

The confusion matrix indicates that certain signs are more susceptible to misclassification, often due to visual similarities or a limited number of training samples. For instance, the sign “Aunt” (Class 1) is frequently misclassified as “Sister” (Class 4), while “Help” (Class 5) is often confused with “Please” (Class 14). These errors likely stem from overlapping hand configurations or similar motion trajectories, underscoring the need to enrich the dataset for ambiguous classes and to investigate advanced techniques, such as attention-based mechanisms, to better distinguish subtle gesture differences. To further assess the stability and robustness of the proposed CNN-LSTM architecture, we repeated the training process five times using different random seeds. The model consistently achieved high performance across all runs, with an average accuracy of 96% and standard deviations below 1.5% for precision, recall, and F1-score. These results confirm that the model's performance remains stable regardless of initialization, reinforcing its reliability for real-world deployment.

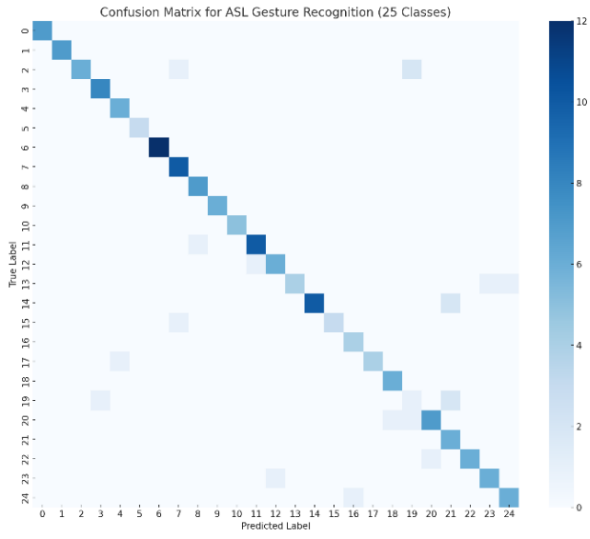

Fig. 4: Confusion matrix showing the classification performance of the proposed CNN-LSTM model on 25 ASL classes

Comparative Analysis

In this comparative study, we evaluate the effectiveness of our proposed approach for sign language recognition using the WLASL dataset against existing state-of-the-art methods. Our model employs ConvLSTM2D for spatio-temporal feature extraction and GRU layers for temporal sequence learning, offering a robust yet computationally efficient solution for dynamic gesture recognition.

Achieving an impressive 96% accuracy on a 25-class subset with 175 videos, our method significantly outperforms prior works in similar settings. While WLASL (Baskoro, 2025) research achieves up to 62.63% top-10 accuracy with pose-based and holistic visual approaches on 2,000 words, our model excels even on a smaller subset, demonstrating superior performance in spatio-temporal modeling. Compared to the ASLLVD study, which emphasizes linguistic annotations and employs a HandShapes Bayesian Network (HSBN) for handshape transitions, our approach focuses on practical and scalable recognition without requiring intricate linguistic details. Our method demonstrates the ability to effectively model spatio-temporal dynamics in small-scale datasets, offering a scalable framework for real-time applications.

Table 4 Presents a comparison between the proposed approach and two referenced works, based on several performance criteria.

From experimental results presented in Table 2, we can conclude that our proposed approach outperforms these methods in accuracy, demonstrating its effectiveness in capturing spatio-temporal patterns with fewer computational resources.

The combination of ConvLSTM2D and GRU enables the proposed model to achieve a balance between accuracy and computational efficiency, making it suitable for real-time and resource-constrained applications.

Ablation Study

To gain deeper insight into the role of each component within the hybrid CNN-LSTM model, an ablation study was performed by evaluating the model's performance under three distinct configurations:

- CNN-only Model: This configuration includes only the convolutional layers followed by fully connected layers for classification. It captures spatial features from individual frames but does not model temporal dependencies
- LSTM-only Model: In this setup, hand-crafted spatial features extracted from individual frames are fed into an LSTM network to model temporal dynamics. No convolutional layers are used
- Hybrid CNN-LSTM Model: This is the full proposed model where CNN extracts spatial features and LSTM/GRU layers capture temporal dependencies

Table 4: Comparative Analysis of Sign Language Recognition Approaches

Criterion	Our Approach (CNN + LSTM)	WLASL Dataset Approach	ASLLVD Dataset Approach
Accuracy	96% (on 25 classes)	Up to 62.63% top-10	Not specified, focuses on linguistic annotations and handshape transitions
Main Method	ConvLSTM2D for spatio-temporal feature extraction and GRU for temporal sequence learning	Visual appearance-based approach and Pose-TGCN (Temporal Graph Convolution Networks)	Handshapes Bayesian Network (HSBN) for handshape transitions in lexical signs
Dataset Used	WLASL (25 classes, 175 videos)	WLASL (2,000 words)	ASLLVD (3,300 signs, 9,800 tokens)
Approach	Single-stream model (simplified and efficient)	Pose-based and appearance-based models	Bayesian network model for handshape transitions
Performance Evaluation	Top-1 accuracy, learning curves, computational efficiency	Top-10 accuracy (up to 62.63%)	Handshape transition analysis, but no direct performance was reported
Advantages of the Approach	High accuracy and efficiency, even on a smaller dataset	Robust approach on a large dataset, but lower accuracy	Linguistic model integrating detailed sign information
Training and Inference Time	Low training and inference time due to a lighter approach	More computationally expensive, especially with large datasets	Can be more complex due to detailed annotations

Table 5: Performance Comparison of CNN, LSTM, and Combined CNN-LSTM Architecture

Model Configuration	Accuracy	Precision	Recall	F1-Score
CNN-only	78%	76%	74%	75%
LSTM-only	65%	62%	61%	61%
CNN + LSTM (proposed approach)	96%	93%	92%	91%

As presented in Table 5, the combined CNN-LSTM architecture outperforms standalone CNN and LSTM. Models, confirming the benefits of integrating spatial and temporal feature extraction. The results clearly show that both components are essential for robust gesture recognition. The CNN-only model is limited by its inability to model temporal changes across video frames, while the LSTM-only model suffers from the lack of spatial abstraction. The hybrid approach significantly outperforms both baselines, demonstrating the advantage of combining spatial and temporal modeling.

Discussion

The results indicate that the hybrid CNN-LSTM architecture effectively addresses the challenges of ASL recognition. Key observations include.

Strong Performance across Classes: The model achieves high precision and recall in most classes, with an F1-score above 0.90 for many.

Challenges in Certain Classes: Classes such as 1 (Aunt) and 5 (Help) exhibit lower recall due to potential data imbalances or visual similarities with other classes.

Generalization: The minimal gap between training and validation metrics suggests strong generalization capabilities.

Limitations

The model's performance in imbalanced or low-data classes highlights the need for data augmentation or advanced techniques, such as dynamic attention mechanisms.

Summary of Results

The proposed CNN-LSTM model demonstrates state-of-the-art performance in recognizing ASL gestures from the WLASL dataset subset. While achieving high accuracy, it highlights the importance of balancing datasets and addressing inter-class variability. Future work can focus on incorporating attention mechanisms and Transformer architectures to further improve performance.

Conclusion

This study introduces a hybrid deep learning model designed to recognize ASL gestures from video sequences, utilizing a selected subset of the Word-Level American Sign Language (WLASL) dataset. By integrating Convolutional Neural Networks (CNNs) for extracting spatial features with Long Short-Term Memory (LSTM) networks to capture temporal dependencies, the model achieved a classification accuracy of 96% across 25 gesture categories.

The findings demonstrate the effectiveness of spatio-temporal modeling in ASL gesture recognition, with the proposed method outperforming comparable approaches under similar experimental setups. The model successfully captures subtle hand shapes and motion dynamics inherent in ASL gestures while maintaining computational efficiency. These results suggest promising applications for real-time ASL translation, educational tools, and assistive communication technologies for individuals who are deaf or hard of hearing.

Nonetheless, the study also highlights certain challenges, particularly in recognizing gestures with limited training samples or those visually similar to others. Addressing these issues will be key to improving model robustness and ensuring reliable performance in practical deployment scenarios.

Limitations and Future Work

While the proposed ConvLSTM GRU architecture demonstrates high recognition accuracy and effective spatio-temporal modeling on a selected subset of the WLASL dataset, several limitations remain. Performance disparities across certain gesture classes persist, mainly due to class imbalance and visual similarities between signs, and will be addressed in future work through weighted loss functions, targeted data augmentation, and oversampling techniques. Although the model performs well under controlled experimental conditions, its robustness in real-world environments characterized by variations in lighting, occlusions, background complexity, and signer-specific differences requires further improvement.

To enhance generalization, future research will focus on integrating skeletal keypoint extraction, multimodal fusion with depth information, and more advanced spatio-temporal augmentation strategies. From an architectural perspective, extensions of the current ConvLSTM GRU framework will explore the incorporation of attention mechanisms, 3D convolutional layers, and transformer-based components to better capture long-range temporal dependencies and subtle gesture dynamics. Finally, to support the practical deployment objectives outlined in this study, lightweight and computationally efficient versions of the model will be developed for real-time implementation on resource-constrained devices such as smartphones and wearable systems.

Acknowledgment

The authors would like to express their sincere gratitude to the Faculty of Sciences, University Sidi Mohamed Ben Abdellah, Fez, Morocco, for providing the academic environment and support that contributed to the completion of this research. The authors also thank all

individuals who were directly or indirectly involved in this work.

Funding Information

The authors declare that they have not received any funding or financial support to report this study.

Author's Contributions

Meryem Cherrate: Conceptualization, methodology, data curation, validation, investigation, writing original draft, writing review and editing.

My Abdelouahed Sabri: Conceptualization, validation, investigation, writing review and editing.

Ali Yahyaouy and Abdellah Aarab: Conceptualization, Writing review and editing.

Ethics

This study used publicly available video datasets that were collected and distributed with ethical approval and participant consent by the original investigators. No new data involving human participants were collected by the authors.

Data Availability Statement

The data used in this study are publicly available from https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed.

References

Abdelouahed, S. M., Meryem, C., Ali, Y., & Abdellah, A. (2022). Moroccan sign language recognition based on machine learning. *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco. https://doi.org/10.1109/iscv54655.2022.9806116

Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, *12*, 200056. https://doi.org/10.1016/j.iswa.2021.200056

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Quan Yuan, & Thangali, A. (2008). The American Sign Language Lexicon Video Dataset. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (CVPR Workshops), Anchorage, AK, USA. https://doi.org/10.1109/cvprw.2008.4563181

Baskoro, R. (2025). WLASL-Processed. *Kaggle*.

Cherrate, M., Sabri, M. A., Yahyaouy, A., Aghoutane, B., Farhaoui, Y., & Aarab, A. (2025a). *Recognition of American Sign Language Using Hard Voting*. 34–40. https://doi.org/10.1007/978-3-031-88304-0_5

Cherrate, M., Sabri, M. A., Yahyaouy, A., Aghoutane, B., Farhaoui, Y., & Aarab, A. (2025b). *Construction and Prediction of American Sign Language Using Deep Learning*. 33–40. https://doi.org/10.1007/978-3-031-90921-4_5

Hafeez, K. A., Massoud, M., Menegotti, T., Tannous, J., & Wedge, S. (2024). American Sign Language Recognition Using a Multimodal Transformer Network. *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Kingston, ON, Canada. https://doi.org/10.1109/ccece59415.2024.10667296

Geislinge, V. (2021). *ASL Fingerspelling Images (RGB & Depth)*. https://www.kaggle.com/datasets/mrgeislinger/asl-rgb-depth-fingerspelling-spelling-it-out

Huang, J., & Chouvatut, V. (2024). Video-Based Sign Language Recognition via ResNet and LSTM Network. *Journal of Imaging*, *10*(6), 149. https://doi.org/10.3390/jimaging10060149

Ji, A., Wang, Y., Miao, X., Fan, T., Ru, B., Liu, L., Nie, R., & Qiu, S. (2023). Dataglove for Sign Language Recognition of People with Hearing and Speech Impairment via Wearable Inertial Sensors. *Sensors*, *23*(15), 6693. https://doi.org/10.3390/s23156693

Kataoka, J., & Yoon, H. (2024). AVATAR: Adversarial self-superVised domain Adaptation network for TARget domain. *Expert Systems with Applications*, *258*, 125147. https://doi.org/10.1016/j.eswa.2024.125147

Lee, B. G., & Lee, S. M. (2018). Smart Wearable Hand Device for Sign Language Interpretation System With Sensors Fusion. *IEEE Sensors Journal*, *18*(3), 1224–1232. https://doi.org/10.1109/jsen.2017.2779466

Li, D., Opazo, C. R., Yu, X., & Li, H. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison.\ *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA. https://doi.org/10.1109/wacv45572.2020.9093512

O'Connor, T. F., Fach, M. E., Miller, R., Root, S. E., Mercier, P. P., & Lipomi, D. J. (2017). The Language of Glove: Wireless gesture decoder with low-power and stretchable hybrid electronics. *PLOS ONE*,*12*(7), e0179766. https://doi.org/10.1371/journal.pone.0179766

Omarkhan, M., Kissymova, G., & Akhmetov, I. (2021). Handling data imbalance using CNN and LSTM in financial news sentiment analysis. *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, Kaskelen, Kazakhstan. https://doi.org/10.1109/icecco53203.2021.9663802

Pugeault, N., & Bowden, R. (2011). Spelling it out: Real-time ASL fingerspelling recognition. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain. https://doi.org/10.1109/iccvw.2011.6130290

Pun, C.-M., Zhu, H.-M., & Feng, W. (2011). Real-Time Hand Gesture Recognition using Motion Tracking. *International Journal of Computational Intelligence Systems*, *4*(2), 277. https://doi.org/10.2991/ijcis.2011.4.2.15

Shen, X., Zheng, Z., & Yang, Y. (2024). StepNet: Spatial-temporal Part-aware Network for Isolated Sign Language Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *20*(7), 1–19. https://doi.org/10.1145/3656046

Shi, X. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems 28*, 26.

Sivaraman, R., Santiago, S., Chinnathambi, K., Sarkar, S., SN, S., & S, S. (2024). Sign Language Recognition Using Improved Seagull Optimization Algorithm with Deep Learning Model. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India. https://doi.org/10.1109/icoici62503.2024.10696047

Toro-Ossaba, A., Jaramillo-Tigreros, J., Tejada, J. C., Peña, A., López-González, A., & Castanho, R. A. (2022). LSTM Recurrent Neural Network for Hand Gesture Recognition Using EMG Signals. *Applied Sciences*, *12*(19), 9700. https://doi.org/10.3390/app12199700

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile. https://doi.org/10.1109/iccv.2015.510

Ur Rehman, M., Ahmed, F., Attique Khan, M., Tariq, U., Abdulaziz Alfouzan, F., M. Alzahrani, N., & Ahmad, J. (2022). Dynamic Hand Gesture Recognition Using 3D-CNN and LSTM Networks. *Computers, Materials & Continua*, *70*(3), 4675–4690. https://doi.org/10.32604/cmc.2022.019586

WHO. (2025). Deafness and hearing loss. *WHO (World Health Organization)*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Xiong, S., Zou, C., Yun, J., Jiang, D., Huang, L., Liu, Y., & Xie, Y. (2025). Continuous sign language recognition enhanced by dynamic attention and maximum backtracking probability decoding. *Signal, Image and Video Processing*, *19*(2). https://doi.org/10.1007/s11760-024-03718-9