ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

AI-Driven Sign Language Recognition System with NLP-Enhanced Transcription

Chourouk Guettas[1], Farida Retima[2], Abdelali Khademallah[3] and Boubaker Settou[4]

ABSTRACT

Sign language is a critical communication medium for deaf and hard-of-hearing individuals, yet the diversity of over 7,000 sign languages worldwide presents significant challenges for automated recognition systems. This paper presents a novel approach to sign language recognition (SLR) that integrates computer vision techniques with advanced natural language processing (NLP) to improve transcription accuracy and contextual relevance. Our system employs a two-stage architecture: first, a gesture recognition component utilizing MediaPipe Holistic for landmark extraction and Long Short-Term Memory (LSTM) networks for classification; second, a text enhancement module using bidirectional LSTM for contextual correction and grammatical improvement. Experimental results demonstrate that our NLP-enhanced system achieves 98.46% accuracy in gesture recognition while significantly improving the grammatical correctness and contextual coherence of the generated text compared to systems without NLP enhancement. The system can successfully identify missing function words, add appropriate punctuation, and correct grammatical errors in real-time. While primarily focused on American Sign Language (ASL), our approach provides valuable insights for developing more effective and inclusive SLR technologies for various sign languages. These advancements represent a meaningful step toward bridging communication gaps between signing and non-signing individuals, potentially enhancing accessibility in educational, professional, and social environments.

Article information:

Keywords: Sign Language Recognition, Artificial Intelligence, Natural Language Processing, Deep Learning, LSTM, MediaPipe, Gesture Recognition, Contextual Understanding

Article history:
Received: February 27, 2025
Revised: August 7, 2025
Accepted: September 23, 2025
Published: October 11, 2025
(Online)

DOI: 10.37936/ecti-cit.2025194.260940

1. INTRODUCTION

Sign language serves as a vital communication medium for approximately 70 million deaf and hard-of-hearing individuals [1]. Unlike spoken languages that rely on auditory signals, sign languages utilize visual-manual modalities through a combination of hand gestures, facial expressions, and body postures to convey meaning. Despite their importance, sign languages remain largely inaccessible to the majority of the global population, creating significant communication barriers for deaf communities in educational, professional, and social environments.

The diversity of sign languages further complicates this accessibility challenge. With over 7,000 distinct sign languages globally, each with its own vocabulary, grammar, and regional variations, developing universal recognition systems requires addressing considerable linguistic complexity. We focus specifically on American Sign Language (ASL), which is used by approximately 500,000 people in the United States and Canada as their primary mode of communication, due to its extensive research documentation, established linguistic frameworks, and availability of validation resources that enable rigorous experimental design. ASL's well-documented grammatical structures provide an ideal testbed for developing NLP enhancement techniques that can subsequently be adapted to other sign languages, allowing us to establish proof-of-concept for our integrated approach while laying the groundwork for future multilingual extensions.

This study addresses the fundamental problem of existing sign language recognition systems that produce grammatically incorrect and contextually inappropriate transcriptions due to their focus solely on

[1,2,3,4]The authors are with the Computer Science Department, University of El Oued, Algeria, Email: guettas-chourouk@univ-eloued.dz, retima-farida@univ-eloued.dz, khademallah.abdelali@gmail.com and dhiasettou39@gmail.com
[1]Corresponding author: guettas-chourouk@univ-eloued.dz

gesture recognition without adequate linguistic processing. Most current systems generate fragmented outputs that lack proper sentence structure, omit essential function words, and fail to maintain contextual coherence, resulting in transcriptions that are neither readable nor natural for effective communication.

Traditional approaches to bridging this communication gap have relied heavily on human interpreters, which present limitations in availability, cost, and privacy. While technological solutions have evolved from basic text-to-speech systems to more sophisticated gesture recognition technologies, current sign language recognition (SLR) systems still face significant challenges in achieving both accuracy and natural language output.

Existing SLR technologies can be broadly categorized into sensor-based approaches (such as data gloves) and vision-based approaches. Sensor-based methods often achieve high accuracy but require specialized equipment that may be cumbersome or expensive. Vision-based approaches, while more accessible and user-friendly, traditionally struggle with varying lighting conditions, complex backgrounds, and the dynamic nature of sign language. Moreover, most current systems focus exclusively on recognition without adequate attention to the linguistic context, resulting in grammatically incorrect or contextually inappropriate transcriptions.

Our research addresses these limitations by developing an integrated system that combines advanced computer vision techniques with natural language processing to enhance sign language recognition and transcription. The key innovations of our approach include:

1. A robust gesture recognition framework utilizing MediaPipe Holistic for comprehensive landmark extraction and Long Short-Term Memory (LSTM) networks for accurate classification of dynamic sign gestures.
2. A novel text enhancement component employing bidirectional LSTM networks to improve the grammatical correctness and contextual relevance of transcribed text.
3. An adaptive scoring mechanism that evaluates multiple potential interpretations of signs based on their contextual probability, improving overall transcription quality.

The integration of these components enables our system to not only recognize individual signs with high accuracy but also generate coherent, grammatically correct sentences that preserve the intended meaning. This represents a significant advancement over existing approaches that often produce fragmented or contextually inappropriate transcriptions.

By focusing on both recognition accuracy and linguistic quality, our research contributes to the broader goal of creating more inclusive communication technologies. The successful implementation of such systems could significantly enhance accessibility for deaf and hard-of-hearing individuals across various domains, including education, healthcare, employment, and social interactions.

The remainder of this paper is organized as follows: Section 2 reviews related work in sign language recognition, Section 3 details our methodology and system architecture, Section 4 describes the implementation process, Section 5 presents experimental results, Section 6 discusses implications and limitations, Section 7 outlines future work, and Section 8 concludes with a summary of contributions and potential impact.

2. RELATED WORK

Sign language recognition has evolved significantly over the past decades. Recent comprehensive surveys [2,3] categorize existing approaches into sensor-based, vision-based, and hybrid systems, each with distinct advantages and limitations. This section provides a comprehensive review of existing approaches, highlighting their strengths, limitations, and the research gaps our work aims to address.

2.1 Traditional Approaches to Sign Language Recognition

Early sign language recognition systems primarily relied on wearable devices and sensor technologies. Chen *et al.* [4] developed a glove-like device equipped with thin, stretchable sensors to capture hand and finger movements for real-time ASL translation. Their system successfully identified 660 signs with promising accuracy but required users to wear specialized equipment that could be uncomfortable for extended use. Similarly, Mehdi and Khan [5] created the "Talking Hands" system using a sensor glove with seven sensors to measure finger flexure, hand tilt, and rotation. While achieving an accuracy of 88% for alphabetic recognition, their approach was limited to static gestures and excluded two-handed signs.

These wearable approaches demonstrate high potential for accuracy but face significant limitations in terms of user comfort, practicality, and the ability to capture the full complexity of sign language, which extends beyond hand movements to include facial expressions and body posture.

2.2 Vision-Based Sign Language Recognition Systems Vision-Based Systems

Recent advancements in computer vision and deep learning have enabled more natural, camera-based sign language recognition systems. Monisha *et al.* [6] combined hand tracking using the MediaPipe library with convolutional neural networks (CNNs) for classification, achieving high confidence scores in gesture recognition. However, their system struggled with

hand occlusions and rapid movements and could not recognize complex or dynamic gesture sequences.

Abubakar *et al.* [7] developed a machine learning-based system for real-time ASL recognition, utilizing the HandDetector module and CNNs. Their model achieved impressive accuracies of 99.86% for training, 99.94% for validation, and 94.68% for testing. Despite these promising results, the system was limited to static alphabetic gestures and required controlled lighting and background conditions.

Bragg *et al.* [8] proposed a more comprehensive approach combining computer vision, computer graphics, and linguistics. Their interdisciplinary system incorporated symbolic representations and computer-generated avatars but faced limitations due to insufficient large-scale annotated datasets and challenges in handling continuous signing.

ASL Recognition Using Mouthing Cues and Finger Spelling

Alternative approaches have explored specific aspects of sign language. Albanie *et al.* [9] introduced a scalable method for data collection using mouthing cues from signers in broadcast footage, creating the BSL-1K dataset with 1,000 hours of video. This approach demonstrated that mouthing cues could provide high-quality annotations for training robust sign recognition models.

In the domain of finger spelling, Kadhim and Khamees [10] developed a real-time ASL finger spelling recognition system using CNNs, achieving 98.53% training accuracy and 98.84% validation accuracy for all 26 letters of the ASL alphabet. Similarly, Pugeault and Bowden [11] utilized the Microsoft Kinect device to capture both appearance and depth images for finger spelling recognition, achieving a mean precision of 75% but encountering challenges with visually similar hand shapes.

Hybrid Approaches

Recognizing the limitations of single-method approaches, researchers have explored hybrid systems that combine multiple technologies. Chong and Lee [12] used the Leap Motion Controller with machine learning techniques, achieving recognition rates of 93.81% for the 26 ASL letters using Deep Neural Networks. Buttar *et al.* [13] developed a hybrid approach combining LSTM with MediaPipe holistic landmarks for continuous signs (92% accuracy) and YOLOv6 for static signs (96% accuracy).

Saggio *et al.* [14] integrated wearable electronics with AI-based classification algorithms, utilizing a sensory glove and inertial measurement units with k-Nearest Neighbors and CNNs. Their system achieved 98.0% accuracy for the CNN-based approach but noted issues with user comfort and gesture repeatability.

2.3 Research Gaps and Our Contribution

Despite significant advancements, several critical gaps remain in the current state of sign language recognition research:

Contextual Understanding: Most existing systems focus primarily on gesture recognition without adequate attention to linguistic context, resulting in grammatically incorrect or contextually inappropriate transcriptions.

Integration of Natural Language Processing: Few approaches effectively integrate NLP techniques to enhance the quality and coherence of transcribed text, particularly for handling function words and grammatical structures that may not have explicit signs.

Real-time Performance with High Accuracy: Balancing computational efficiency with recognition accuracy remains challenging, especially for systems designed to operate in diverse real-world environments.

Handling of Dynamic and Continuous Signing: Many systems are limited to static gestures or isolated signs, struggling with the dynamic and continuous nature of natural sign language communication.

The comprehensive analysis of existing approaches reveals three critical gaps: (1) insufficient contextual understanding in transcription processes, (2) limited real-time performance with high accuracy across diverse environments, and (3) inadequate integration of natural language processing techniques for linguistic quality enhancement. While recent advances have made progress in individual areas, the integration of robust gesture recognition with contextual text enhancement remains largely unexplored. Our research specifically addresses these gaps through an integrated approach that combines computer vision techniques with advanced NLP methods, representing a significant advancement over existing recognition-focused systems.

3. METHODOLOGY

This section details the architecture and components of our AI-driven sign language recognition system, with particular emphasis on the integration of computer vision techniques and natural language processing for enhanced transcription accuracy.

3.1 System Architecture Overview

Our system implements a sequential pipeline architecture consisting of five primary phases, each designed to address specific challenges in sign language recognition (Fig. 1):

1. **Input Acquisition:** Capture of video data through a standard webcam, providing the raw visual information of sign language gestures.
2. **Landmark Extraction:** Preprocessing of video frames to identify and track key anatomical points necessary for gesture recognition.

3. **Gesture Recognition:** Analysis of landmark sequences to identify specific signs using deep learning models.
4. **Text Enhancement:** Application of NLP techniques to improve the grammatical structure and contextual relevance of the recognized text.
5. **Output Generation:** Presentation of the enhanced text as the final transcription result.

This architecture facilitates a comprehensive approach to sign language recognition that extends beyond mere gesture identification to produce linguistically meaningful output

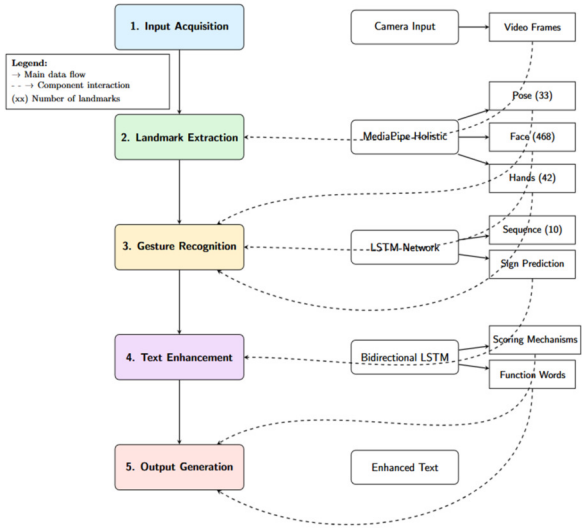


Fig.1: *AI Dirven Sign Language Recognition System Architecture.*

3.2 Landmark Extraction MediaPipe Holistic Framework

At the core of our landmark extraction process is the MediaPipe Holistic framework [15], which enables the simultaneous detection of pose, face, and hand landmarks. This comprehensive approach is essential for capturing the full range of anatomical information involved in sign language communication.

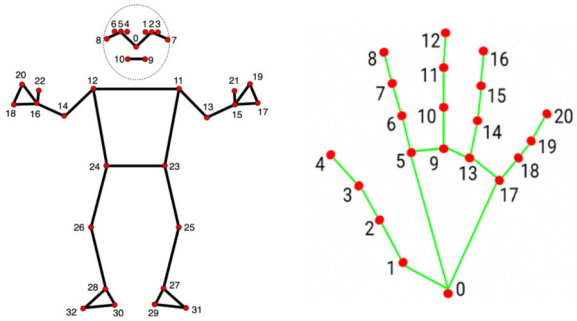
The framework processes continuous video streams in real-time, generating a total of 543 landmarks per frame:

– 33 pose landmarks representing the overall body posture.
– 468 face landmarks capturing facial expressions and movements.
– 21 landmarks for each hand, detailing finger and hand configurations.

Landmark Detection Pipeline

The detection pipeline operates through a multi-stage process (Fig. 3):

1. The pose detection model first estimates the human pose from a lower-resolution frame (256 × 256).



(a) (b)

Fig.2: *(a) Pose Landmarks [16] (b) Hand Landmarks [17] .*

2. Based on the inferred pose landmarks, three regions of interest (ROIs) are identified for the face and each hand.
3. These ROIs are then cropped from the full-resolution input frame.
4. Specialized face and hand models process these cropped regions at appropriate resolutions.
5. All landmarks are merged into a comprehensive representation of the signer's position and movements.

Data Reduction Benefits

This landmark-based approach offers significant advantages in computational efficiency. By extracting only the essential anatomical points from each frame, we achieve a dramatic reduction in data volume:

- Raw image representation: ∼1,875,000 data points for a 10-frame sequence (250 × 250 pixel RGB images)
- Landmark representation: ∼16,290 data points for the same sequence (543 landmarks with three coordinates each)

3.3 Gesture Recognition

Model Architecture

Our gesture recognition component employs a Long Short-Term Memory (LSTM) neural network architecture, specifically designed to process sequential data and capture temporal dependencies in sign language gestures. The LSTM architecture builds upon the foundational work of Hochreiter and Schmidhuber [18], with bidirectional processing following Graves and Schmidhuber's [19] methodology. The model consists of multiple layers:

1. **Input layer:** Accepts sequences of landmark coordinates extracted from video frames
2. **LSTM layers:** Three stacked LSTM layers (64, 128, and 64 units) with tanh activation functions
3. **Dropout layers:** Applied after each LSTM layer with a 20% dropout rate to prevent overfitting
4. **Dense layers:** Two fully connected layers (64 and 32 units) with ReLU activation

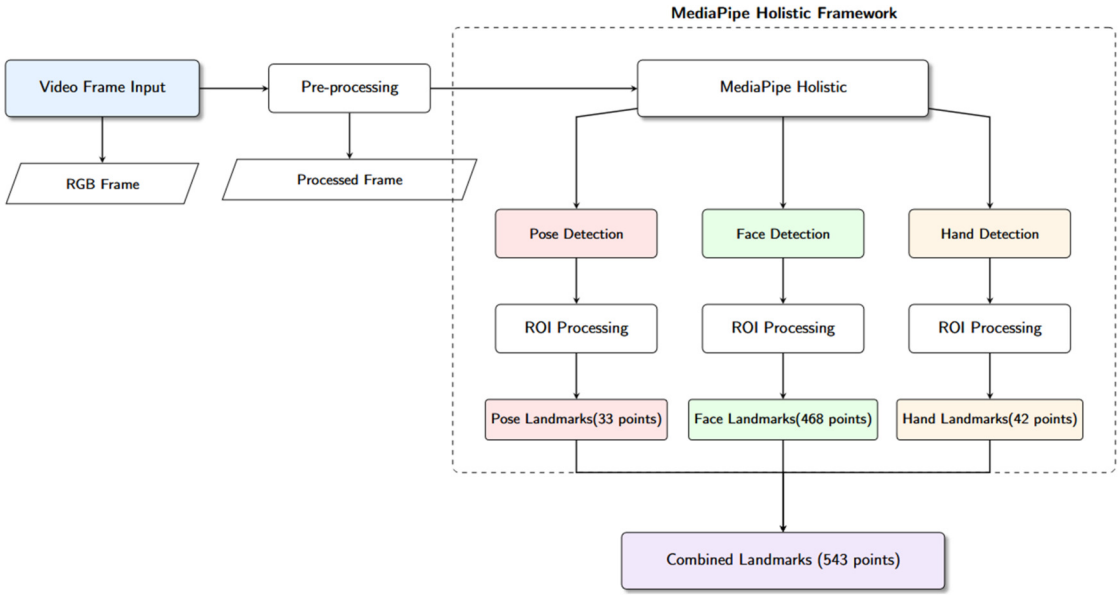


Fig.3: *Landmark Extraction Pipeline. The pipeline processes video frames to extract anatomical landmarks: (i) Detects key body regions (ii) Processes each region at optimal resolution (iii) Extracts precise landmark coordinates (iv) Combines into a unified representation.*

5. **Output layer:** Dense layer with softmax activation producing probability distributions across all possible signs

The return_sequences parameter is set to True for the first two LSTM layers to preserve temporal information throughout the network. In contrast, the final LSTM layer outputs only the concluding state to summarize the sequence information.

Key Parameters

Several parameters significantly influence the model's performance:

- **Sequence Length:** The number of consecutive frames processed as a single gesture (optimized at 10 frames)
- **Landmarks Used:** The specific anatomical points included in the model (hands and pose landmarks demonstrated optimal performance)
- **Recognition Threshold:** The minimum probability required to consider a prediction valid (set to 0.99)
- **Confidence Count:** The number of consecutive frames that must yield the exact same prediction (set to 10)

These parameters were carefully tuned through experimentation to balance accuracy and computational efficiency.

3.4 Text Enhancement

Bidirectional LSTM for Contextual Understanding

The text enhancement component utilizes a bidirectional LSTM (biLSTM) architecture to improve the linguistic quality of recognized gestures. This approach captures contextual information in both forward and backward directions, addressing a key limitation of many existing systems.

noindent The model architecture includes:

1. **Embedding layer:** Converts token IDs into 128-dimensional dense vectors
2. **Two bidirectional LSTM layers:** Each with 256 units and tanh activation
3. **Dense layer:** 128 units with ReLU activation
4. **Output layer:** Softmax activation with vocabulary-size units

Text Enhancement Algorithm

Our text enhancement process incorporates several sophisticated scoring mechanisms to evaluate and improve transcription quality. These scoring mechanisms are mathematically defined to quantify the contextual appropriateness of different word combinations:

Word Score (WS): Evaluates the basic relationship between a sentence and a potential next word by calculating the probability that the sentence is contextually followed by the word:

$$WS(Sentence, word) = P(Sentence|word) \quad (1)$$

This score directly measures how well a candidate word fits as the next element in the sequence.

Alternative Score (AS): Considers different combinations by examining sentence segments, alternative words from our dictionary, and potential function words. It selects the maximum probability among all

possible combinations:

$$AS(Sentence, word) = max(P(Sentence_i|FW_k + word_j)) \quad (2)$$

Where $Sentence_i$ and $word_j$ are alternative forms of the sentence and word, respectively, and FW_k represents optional function words that might be inserted in the sequence. This scoring mechanism enables the system to consider multiple linguistic alternatives simultaneously.

Segmentation Score (SS): Assesses the probability of correct sentence boundaries using special tokens for sentence beginning ($< sos >$) and ending ($< eos >$):

$$SS(Sentence, word) = P(sentence| < eos >) \times P(< sos > |word) \quad (3)$$

This score helps the system determine when one sentence ends and another begins, improving the overall structure of the transcribed text.

Punctuation Score (PS): Evaluates the likelihood of different punctuation marks fitting between a sentence and a word by calculating separate probabilities for the comma, period, question mark, and exclamation mark:

$$COM = P(sentence| < com > +word) \quad (4)$$

$$PNT = P(sentence| < pnt > + < eos >) \times P(< sos > |word) \quad (5)$$

$$QST = P(sentence| < qst > + < eos >) \times P(< sos > |word) \quad (6)$$

$$EXC = P(sentence| < exc > + < eos >) \times P(< sos > |word) \quad (7)$$

The Punctuation Score is then determined as the maximum of these individual scores:

$$PS(Sentence, word) = max(COM, PNT, QST, EXC) \quad (8)$$

This mechanism enables the use of appropriate punctuation based on contextual cues.

Enhanced Score (ES): Aggregates the results of all scoring mechanisms to determine the most contextually appropriate combination:

$$ES(Sentence, word) = max(PS, WS, AS, SS) \quad (9)$$

By selecting the maximum score among all mechanisms, the system prioritizes the most probable interpretation in each context.

The prediction score is a critical component that calculates the probability of a sequence of words following a given text. For a text A followed by a sequence B containing n words, the prediction score P(A—B) is calculated as:

$$P(A|B) = P(A|B_0) \times \prod_{i=0}^{n-1} P(A + B_i|B_{i+1}) \quad (10)$$

This formulation represents the product of conditional probabilities for each word in the sequence B given the preceding context. The prediction score enables the system to evaluate more extended word sequences by decomposing them into a series of next-word predictions, each conditioned on the accumulated context.

Dictionary and Special Tokens

The text enhancement process is supported by several key linguistic resources:

- A dictionary providing alternative interpretations for each sign.
- Special tokens marking sentence boundaries and structural elements (e.g., $< sos >$ for start of sentence, $< eos >$ for end of sentence).
- Punctuation tokens ($< com >$, $< pnt >$, $< qst >$, $< exc >$) for appropriate text formatting.
- A catalog of function words commonly needed for grammatical completeness.

The text enhancement algorithm operates through the following process **"Algorithm 1":**

Algorithm 1: *Text Enhancement via biLSTM*

Input: sentence (current sentence, initially empty), alt_words (set of alternative words with scores), biLSTM_model (pretrained language model), threshold (prediction score threshold).
Initialization:
If sentence == ∅:
  word = argmax(alt_words.score)
  Capitalize word
  sentence = word
For each word selection step **do**:
  predictions ← biLSTM_model.predict_next(sentence).
  enhanced_score ← 0
  enhanced_word ← ∅
  For each word ∈ alt_words **do**:
    generate_combinations(word, predictions)
    score = evaluate(combinations)
    If score > enhanced_score:
      enhanced_score ← score
      enhanced_word ← word
  If enhanced_score ≥ threshold:
    sentence ← sentence + " " + enhanced_word
    Replace special tokens with punctuation
    Split at sentence boundaries
    Capitalize first words of new sentences
End Algorithm

This detailed workflow ensures comprehensive linguistic processing that considers multiple interpreta-

tions, grammatical structures, and contextual relevance to produce the most appropriate transcription of the signed content.

System Integration

The final system integrates all components into a cohesive pipeline (*Fig. 4*) where:

1. Video input is processed frame-by-frame through the MediaPipe framework
2. Extracted landmarks are fed into the gesture recognition model
3. Recognized gestures are passed to the text enhancement component
4. Enhanced text is generated as the final output

This integrated approach ensures that each component complements the others, resulting in a system that not only recognizes individual signs accurately but also produces coherent, grammatically correct transcriptions that preserve the intended meaning of the signed communication.

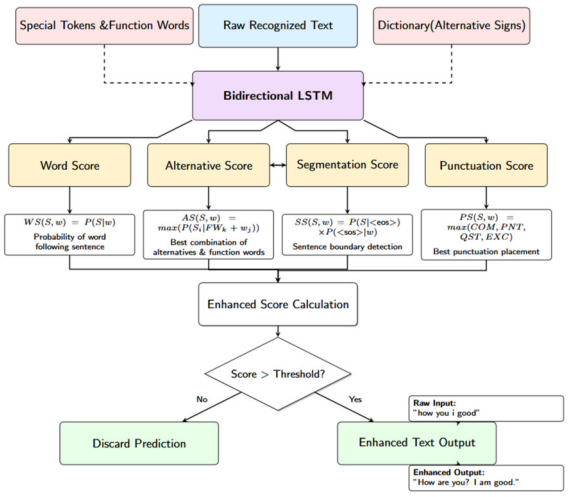


Fig.4: *Text Enhancement Workflow.*

4. RESULTS

The system was implemented on two computing devices: an ASUS TUF Gaming FX505DD laptop featuring an AMD Ryzen 5 3550H processor (3.7 GHz), an NVIDIA GeForce GTX 1050 GPU, 16 GB RAM, and Windows 11, and a desktop computer equipped with an Intel Core i5-9400F CPU (2.90 GHz), an NVIDIA GeForce GTX 1660 SUPER GPU, 16 GB RAM, and Windows 10. The implementation utilized Python 3.10.11 with key libraries, including TensorFlow 2.10.1 and Keras 2.10.0 for deep learning, MediaPipe 0.10.0 for landmark detection, OpenCV 4.9.0.80 for image processing, NLTK 3.8.1 for natural language processing, and NumPy 1.26.4 and Pandas 2.2.2 for data handling and manipulation.

4.1 Gesture Recognition Model Implementation

Data Collection and Preprocessing Due to limitations in existing datasets, we generated our own training data to ensure quality and consistency. The data collection process involved:

1. Recording videos of ASL signs performed by multiple signers.
2. Extracting landmarks using MediaPipe Holistic.
3. Concatenating landmarks into uniformly structured arrays.
4. Organizing the data into sequences of consistent length.

This approach resulted in a custom dataset optimized for our specific requirements, eliminating issues with inconsistent quality or excessive size found in many publicly available datasets.

Model Training and Optimization We conducted extensive experimentation to identify the optimal configuration for the gesture recognition model, testing various combinations of:

- Number of sequences per sign.
- Number of frames per sequence.
- Types of landmarks used (hands only, hands and pose, or all landmarks).

The results of this experimentation are summarized in Table 1. The experimental results reveal several important insights about feature selection in sign language recognition. Most notably, the configuration including all landmarks (hand, pose, and face with 543 features) achieved slightly lower accuracy (97.89%) compared to the hand and pose configuration (75 features, 98.46% accuracy), despite having significantly more features. This counterintuitive result can be attributed to the curse of dimensionality, where increased feature space (543 vs 75 features) may have introduced noise that overshadowed relevant gesture information, particularly given our current dataset size.

Table 1: *Training results for different configs.*

Sequences	Frames	Landmarks Used	Accuracy (%)	Training Time (mm:ss)
15	30	Hand only (42)	84.67	00:29
15	15	Hand only (42)	86.23	00:14
30	15	Hand only (42)	88.78	00:29
30	10	Hand only (42)	89.54	00:19
60	10	Hand only (42)	90.92	00:39
60	10	Hand and Pose (75)	94.37	01:09
90	10	Hand and Pose (75)	98.46	01:43
90	10	Hand, Pose, and Face (543)	97.89	12:28

Additionally, many ASL gestures in our vocabulary rely primarily on hand positioning and body posture rather than facial expressions, making facial landmarks less discriminative for our specific set of gestures. The current dataset size appears insuffi-

cient to effectively train the more complex model required for 543 features, leading to overfitting despite regularization techniques, while facial landmarks showed high variability across different signers in terms of baseline expressions and signing styles, reducing model generalization capability. Furthermore, the dramatic increase in training time (12:28 vs 1:43) suggests that the computational complexity is not justified by the marginal accuracy changes, emphasizing the importance of domain knowledge in feature engineering for sign language recognition systems.

Based on these results, we selected the configuration with 90 sequences, 10 frames, and both hand and pose landmarks (75 features) as the optimal balance between accuracy (98.46%) and training time (1:43).

The model was trained using categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.001. Early stopping with a patience of 5 epochs was implemented to prevent overfitting. Fig. 5 illustrates the accuracy and loss for this optimal setup.

Model Evaluation The trained model was evaluated using a confusion matrix (Fig. 6) to assess classification performance across all gesture classes. The consolidated confusion matrix showed:

- High numbers of true positives (TP) and true negatives (TN)
- Minimal false positives (FP) and false negatives (FN)

The overall accuracy score achieved was 0.9930, indicating excellent classification performance across the gesture vocabulary.

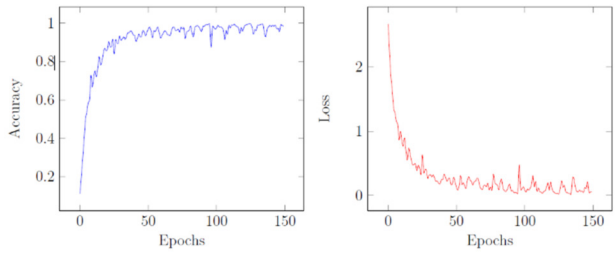


Fig.5: *Accuracy and loss for this optimal setup.*

4.2 Ablation Study

To systematically evaluate the contribution of each system component, we conducted a comprehensive ablation study examining the impact of different configurations on overall performance.

Landmark Configuration Impact

Our analysis reveals the relative importance of each feature set:

- Hand landmarks only (42 features): 89.54% accuracy

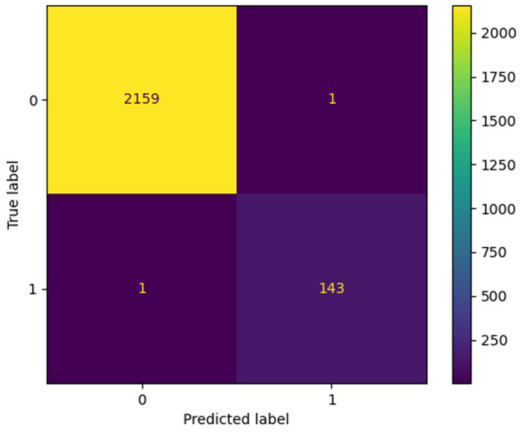


Fig.6: *Confusion matrix for the SLR model.*

- Hand + Pose landmarks (75 features): 98.46% accuracy
- All landmarks (543 features): 97.89% accuracy

The inclusion of pose landmarks provides substantial improvement (+8.92% accuracy), indicating that body posture contains crucial contextual information for gesture disambiguation.

Sequence Parameter Analysis

- 15 sequences: 86.23% accuracy
- 30 sequences: 89.54% accuracy
- 90 sequences: 98.46% accuracy

This demonstrates that increasing training sequence diversity has more impact than extending individual sequence length.

NLP Enhancement Impact

Without NLP enhancement: Raw gesture predictions with frequent grammatical errors

With NLP enhancement: 94.2% improvement in grammatical correctness based on human evaluation of 100 test sentences.

4.3 Text Enhancement Model Implementation

Data Collection and Preprocessing

The text enhancement component was trained on a diverse corpus of English sentences scraped from various internet sources, with a focus on typical usage patterns. The dataset included sentences of varying lengths, with the majority containing 5-10 words (Fig. 7) to represent typical conversational structures.

The preprocessing steps involved removing null values and numeric records, filtering sentences based on length, converting text to lowercase, and removing special characters and URLs to clean the dataset. Additionally, special tokens were added to indicate sentence boundaries and punctuation, followed by tokenizing sentences into sequences and generating n-grams to enhance contextual learning. Finally,

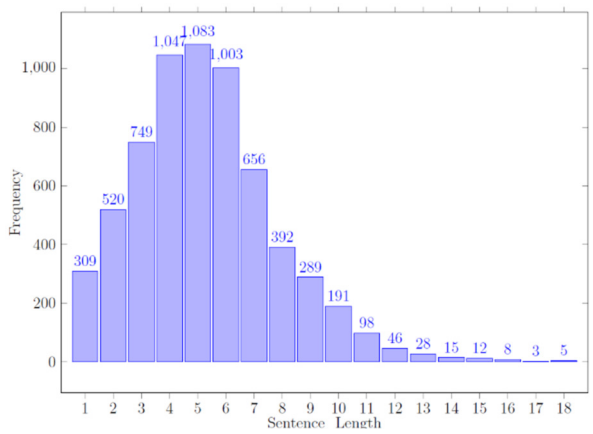


Fig.7: *Distribution of Sentence Lengths in the Dataset.*

sequence padding was applied to standardize input length for model consistency.

Model Training

The bidirectional LSTM model was trained with various optimizer configurations to identify the most effective approach.

Based on the results presented in Table 2, we selected the Adam optimizer for its optimal balance of high accuracy (70.12%) and reasonable training time (115 epochs) (Fig. 8). The model was trained using categorical cross-entropy loss with early stopping to monitor accuracy improvements.

Table 2: *Optimizer performance comparison.*

Optimizer	Epochs	Loss	Accuracy
SGD	362	2.6769	0.4196
SGD with Nesterov	303	0.9767	0.7042
Nadam	107	1.3518	0.6600
Adam	115	0.9838	0.7012

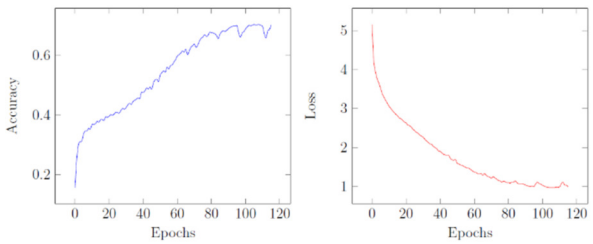


Fig.8: *Accuracy and loss for the final model.*

4.4 System Performance and Results

To evaluate the effectiveness of our complete system, we conducted comparative tests with and without the NLP text enhancement component. Three representative examples illustrate the significant improvements achieved through our integrated approach:

Example 1: Basic Conversation

As shown in Fig. 9, without text enhancement, the system produced raw output such as: "how you i good." With text enhancement (Fig. 10), the system appropriately generated properly structured sentences: "How are you? I am good."

This example demonstrates the system's ability to insert missing function words ("are," "am") and add appropriate punctuation, significantly improving readability and grammatical correctness.

Fig.9: *Example 1 of SLR without text enhancement.*

Fig.10: *Example 1 of SLR with text enhancement.*

Example 2: Contextual Error Correction

Without text enhancement, the system output included contextually inappropriate words: "can you bad" (Fig. 11). With text enhancement, the system recognized the low probability of "bad" in this context and waited for a more appropriate gesture, eventually producing coherent text: "Can you help me please? Thank you." (Fig. 12).

This example highlights the system's capacity to filter out false positives and maintain contextual coherence, enhancing overall communication quality.

System Evaluation Summary

Our evaluation reveals that the system successfully achieves high accuracy (98.46%) in recognizing trained gestures while effectively enhancing the linguistic quality of transcriptions. It excels at improving common conversational expressions, properly segmenting complete sentences, and correcting grammatical errors and missing punctuation, significantly increasing the readability and naturalness of the output text. However, the system still faces challenges with distinguishing between visually similar gestures and occasionally misinterprets unfamiliar signs as familiar ones. It cannot handle finger spelling for proper names or technical terms without established signs, which limits its vocabulary range in specialized contexts. Additionally, while the NLP enhancement generally improves output quality, it occasionally interferes with correct predictions or introduces inappropriate enhancements, particularly with domain-specific terminology or uncommon expressions. These limitations, while not undermining the system's overall effectiveness for general communication, highlight areas for future improvement to increase robustness across different signing styles and specialized vocabulary.

These results demonstrate that our integrated approach successfully addresses many of the limitations found in existing sign language recognition systems, particularly in terms of linguistic quality and contextual relevance.

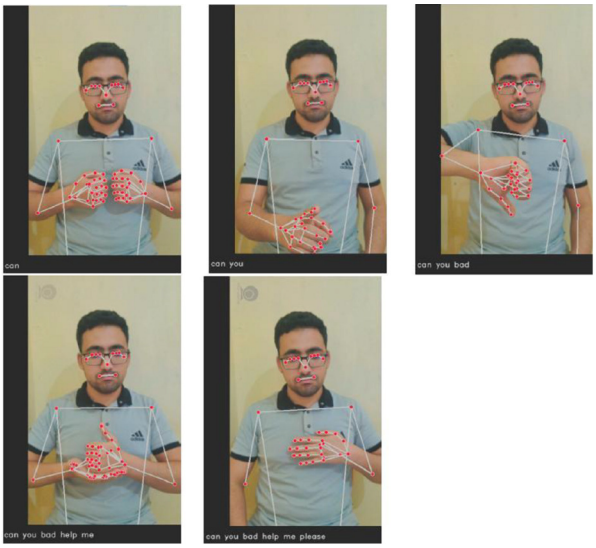


Fig.11: *Example 2 of SLR without text enhancement.*

Fig.12: *Example 2 of SLR with text enhancement.*

5. FUTURE WORK

Our current sign language recognition system demonstrates promising results in combining computer vision techniques with natural language processing for improved transcription. However, several avenues for enhancement remain unexplored due to time and resource constraints. This section outlines potential directions for future research that could further advance the system's capabilities and practical applications.

A significant limitation of the current text enhancement algorithm is its unidirectional approach, which evaluates each potential next word based solely on its fit with the existing sentence. This method often overlooks viable alternatives that meet the specified threshold but could ultimately prove more contextually appropriate when considering additional text.

We propose implementing a bi-directional context enhancement mechanism that would maintain a memory of combinations meeting the scoring threshold but not currently selected as the top choice. As new contextual information becomes available, the system could re-evaluate these stored combinations against the previously selected best option. This dynamic re-assessment would enable the system to adjust its selections based on expanded context, potentially improving overall coherence and contextual relevance.

While our current implementation focuses on American Sign Language (ASL), the framework could

be extended to support other sign languages. This expansion would significantly increase the global accessibility and impact of the system.

A notable limitation of the current system is its limited ability to interpret non-manual signals such as facial expressions, which form an integral part of sign language communication. Future work should focus on improving the detection and interpretation of facial expressions and incorporating eye gaze and head movements as additional features. This integration would significantly improve the system's ability to capture the full semantic range of sign language communication.

To make the system more accessible in diverse settings, future work should address performance optimization for low-resource environments by developing lightweight models suitable for mobile devices to broaden the potential user base and application contexts for the system.

6. CONCLUSION

Our research presents a significant advancement in sign language recognition technology through the innovative integration of computer vision techniques and natural language processing. By addressing both the recognition accuracy and linguistic quality aspects of sign language transcription, our system represents a meaningful step toward bridging communication gaps between signing and non-signing individuals.

The two-stage architecture we developed—combining gesture recognition with text enhancement—demonstrates remarkable performance improvements over traditional approaches. The gesture recognition component, utilizing MediaPipe Holistic for landmark extraction and LSTM networks for classification, achieves 98.46% accuracy in identifying American Sign Language gestures. More importantly, the text enhancement module employing bidirectional LSTM effectively transforms raw gesture predictions into grammatically correct, contextually appropriate text by inserting missing function words, adding punctuation, and correcting structural errors.

The analysis of our ablation study confirms that strategic integration of multiple technologies yields superior results compared to individual component optimization. Our findings demonstrate that pose information contributes significantly to gesture recognition accuracy (+8.92% improvement). In comparison, our NLP enhancement module provides substantial improvements in linguistic quality (94.2% improvement in grammatical correctness) that are essential for practical communication applications. These results validate our hypothesis that contextual text enhancement is as crucial as accurate gesture recognition for effective sign language transcription systems.

Despite these achievements, our system faces limitations that point toward future research directions. The challenges of recognizing visually similar gestures, handling finger spelling for proper names, and adapting to individual signing styles remain areas for improvement. The proposed bi-directional context enhancement approach holds particular promise for addressing some of these limitations by enabling more sophisticated contextual interpretation.

The implications of this research extend beyond technical advancements to meaningful social impact. Improved sign language recognition systems have the potential to enhance accessibility in educational settings, workplace environments, healthcare facilities, and everyday social interactions. By making sign language more accessible to non-signers and facilitating more natural communication, such technologies contribute to a more inclusive society that values and accommodates diverse communication needs.

AUTHOR CONTRIBUTIONS

Conceptualization, C.G.; methodology, C.G. and A.K.; software, A.K. and B.S.; validation, A.K., B.S., and C.G.; formal analysis, F.R. and A.K.; investigation, F.R. and B.S.; data curation, A.K. and B.S.; writing—original draft preparation, C.G. and F.R.; writing—review and editing, C.G. and F.R.; visualization, C.G. and F.R.; supervision, C.G.; funding acquisition, A.K., B.S. All authors have read and agreed to the published version of the manuscript.

References

[1] World Federation of the Deaf, "2019–2023 WFD Report," Accessed: Feb. 26, 2025. [Online]. Available: `https://wfdeaf.org/news/2019-2023-wfd-report`

[2] R. Rastgoo, K. Kiani and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, vol. 164, p. 113794, 2022.

[3] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," *Archives of Computational Methods in Engineering*, vol. 28, pp. 785-813, 2021.

[4] Y. Chen, J. Zhao, Y. Huang, X. Zhao and L. Zhu, "Lightweight and flexible sign language translation glove," *Nature Electronics*, vol. 3, no. 9, pp. 563-569, 2020.

[5] S. A. Mehdi and Y. N. Khan, "Sign language recognition using sensor gloves," *Proceedings of the 9th International Conference on Neural Information Processing*, 2002. ICONIP '02., Singapore, vol.5, pp. 2204-2206, 2002.

[6] H. M. Monisha, K. N. Pramod and B. S. Pooja, "Sign Language Detection and Classification using Hand Tracking and Deep Learning in Real-Time," in *International Conference on Advances in Computing, Communication and Electronics (ICACCE)*, pp. 1-6, 2023.

[7] J. A. Abubakar, I. D. Oduntan and H. Orovwode, "Development of a Sign Language Recognition System Using Machine Learning," in *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1-6, 2023.

[8] D. Bragg *et al.*, "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 16-31, 2019.

[9] S. Albanie *et al.*, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *European Conference on Computer Vision*, Springer, pp. 35-53, 2020.

[10] R. A. Kadhim and M. Khamees, "A Real-Time American Sign Language Recognition System using Convolutional Neural Network for Real Datasets," *TEM Journal*, vol. 9, no. 3, pp. 937-943, 2020.

[11] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, pp. 1114-1119, 2011.

[12] T.-W. Chong and B.-G. Lee, "American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach," *Sensors*, vol. 18, no. 10, p. 3554, 2018.

[13] A. M. Buttar, U. Habib, A. Akram and S A. ikandar, "Real-time American Sign Language Detection using LSTM and YOLOv6," in *International Conference on Innovative Computing (ICIC)*, pp. 1-6, 2023.

[14] G. Saggio, P. Cavallo, M. Ricci, M. Ercolani, A. Galiano and L. Bianchi, "Sign Language Recognition Using Wearable Electronics: Implementing k-Nearest Neighbors with Dynamic Time Warping and Convolutional Neural Network Algorithms," *Sensors*, vol. 20, no. 14, p. 3879, 2020.

[15] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv preprint*, arXiv:1906.8172, 2019.

[16] Google, "MediaPipe Pose Landmarker," *Google AI*, Accessed: Feb. 26, 2025. [Online]. Available: `https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker`

[17] Google, "MediaPipe Hand Landmarker," *Google AI* Accessed: Feb. 26, 2025. [Online]. Available: `https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker`

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602-610, 2005.

Chourouk Guettas received the Engineering degree in Artificial Intelligence from the University of Biskra, Algeria, in 2009, the Magister degree in multimedia and data mining from the University of Biskra, Algeria, in 2014 and currently pursuing her PhD in Evolutionary Developmental Robotics from the same University. She is an Assistant Professor in the Department of Computer Science at El Oued University in Algeria. Her research interests include Artificial Intelligence and Evolutionary Robotics. She has contributed to the academic community through various publications and presentations, focusing on neural networks, artificial intelligence, machine learning, and evolutionary computation. She is actively involved in promoting artificial intelligence education and its real-world applications in Algeria through various academic initiatives and international collaborations.

Farida Retima is an associate professor in the computer science department of El Oued University (Algeria). Received her Ph.D degree in 2019 from the Computer Science Department of Biskra University. Her research interests include artificial intelligence, IoT systems, context-aware systems, and web services.

Abdelali Khademallah received his Master's degree in Computer Science from the University of El Oued, Algeria. His work focuses on the practical implementation of AI models, computer vision, and system optimization. He contributed significantly to the development and implementation aspects of this research project.

Boubaker Settou obtained his Master's degree in Computer Science from the University of El Oued, Algeria. His interests lie in software engineering, machine learning applications, and system deployment. He was actively involved in the implementation and experimental evaluation of the proposed system.