

A Computer Vision-Based System for LESHO: Implementation of a Software System for the Translation of Static Gestures

María Rivera¹ ; Jorge Paz² ; Reyna Valle³ 

^{1,2,3}Universidad Tecnológica Centroamericana, Honduras, maria.rivera@unitec.edu, franciscosauceda041016@unitec.edu, elizabeth.valle@unitec.edu

Abstract—Language enables the transmission of information, ideas, and emotions; however, significant communication barriers persist for the deaf community in many everyday contexts. This paper presents a real-time software system for the translation of Honduran Sign Language (LESHO) based on computer vision and machine learning techniques. The proposed system relies on structured spatial representations extracted from anatomical landmarks of the hands and contextual body cues, rather than raw image data, enabling efficient and robust gesture recognition under real-time constraints. Gesture acquisition, feature extraction, and classification are integrated within a modular architecture implemented in Python using OpenCV, MediaPipe, and Scikit-learn. The system supports fixed-length landmark vectors for static gestures and temporal landmark sequences for dynamic gestures. Experimental evaluation on a multi-class dataset demonstrates high recognition accuracy and interactive response times on consumer-grade hardware. Overall, the results validate the effectiveness of landmark-based modeling for sign language recognition and highlight the system’s scalability for expanding vocabulary coverage and supporting inclusive communication in both face-to-face and technology-mediated environments.

Keywords—Assistive Communication Technologies, Computer Vision, Gesture Recognition, Honduran Sign Language, Machine Learning

I. INTRODUCTION

Communication permits us to transmit our thoughts, feelings, needs, and ideas, allowing for relationships and participation in society. For people with communication disabilities, transmitting their ideas is a challenge, especially when their environment does not master sign language it is more prone for social isolation to occur, affecting their mental health and personal development.

It was not until 2013 that Honduran Sign Language (LESHO) was officially recognized in Honduras with decree 321-2013, which means that many deaf adults did not learn any structured sign system [1].

According to the National Commissioner for Human Rights, in Honduras, it is estimated that more than 100,000 people have hearing impairments, and only 1% have an identification card. The above, together with the deficiency in public infrastructure, hinders access to work and public transportation, leading to discrimination [2].

The Association of Sign Language Interpreters of Honduras has only 19 members, they do not have legal recognition by

the Ministry of the Interior, which prevents them from joining the World Federation of Interpreters [3]. Despite initiatives such as educational programs at the National Autonomous University of Honduras (UNAH), access to LESHO learning remains limited at a national level [4].

Despite regulatory and educational advances, people with communication disabilities in Honduras continue to face structural and social barriers that limit their autonomy and participation. The lack of certified interpreters, low access to Honduran Sign Language (LESHO) learning programs around the country, and late recognition of this language have left thousands of citizens at a communicative disadvantage.

Faced with this scenario, it is imperative to develop inclusive technologies that not only make up for the shortage of human resources, but also empower deaf and/or deaf-mute people with tools that allow them to interact autonomously, with dignity and in real time with their environment [5].

In this context, solutions such as application of machine translation represent a concrete and accessible response to a historically ignored problem. Hand gesture recognition has emerged as a viable technology for human-computer interaction (HCI).

Through hand gestures captured via webcam feed, users can perform various tasks, leveraging computer vision algorithms and ML techniques, hand gesture recognition systems can detect and track hand movements in real-time, enabling intuitive control over digital interfaces [6]. Building upon these advances, several research efforts have explored the application of computer vision and machine learning techniques for sign language recognition and translation [7].

Vision-based approaches have gained significant attention due to their non-intrusive nature and reduced dependency on specialized hardware [8] [9]. These systems typically rely on cameras to capture hand movements, followed by feature extraction and classification using machine learning models [10] [11].

Recent studies have demonstrated promising results using deep learning architectures, such as convolutional neural networks (CNNs), for recognizing sign language gestures from images and video sequences [12] [13]. While these approaches often achieve high accuracy, they generally require large annotated datasets, substantial computational resources, and

controlled acquisition environments, which may limit their scalability and deployment in resource-constrained settings [14] [15].

In contrast, landmark-based approaches combined with classical supervised learning offer a compelling alternative for resource-constrained deployment contexts. Rather than processing raw image pixels, these methods extract structured geometric representations, specifically low-dimensional feature vectors, that reduce computational complexity while maintaining competitive recognition performance [16] [17] [18].

Critically, unlike CNN-based systems that require GPU-level hardware and thousands of training samples per class [14] [15], landmark-based pipelines can be trained and deployed on standard consumer-grade laptops, making them particularly suited to contexts such as Honduras where specialized infrastructure is unavailable. However, most existing implementations of this approach remain limited to internationally dominant sign languages such as American Sign Language (ASL), are evaluated under strictly controlled laboratory conditions, or are not validated on locally representative datasets [16] [17].

Within the Latin American context, recent research efforts have begun addressing national sign languages through computer vision approaches. Mora-Zarate et al. proposed a landmark-based dataset for Colombian Sign Language recognition [9], Rodriguez et al. developed a MediaPipe-based recognition system for Mexican Sign Language achieving 92% accuracy on static signs [19], Garcia-Gil et al. proposed a real-time dynamic Mexican Sign Language recognition framework using compact angular geometric descriptors deployable on low-cost devices [20], and Liu et al. applied spatiotemporal attention mechanisms for Argentine Sign Language recognition on the LSA64 dataset [21].

These efforts demonstrate a growing regional interest in developing sign language technologies for underrepresented national languages; however, they remain focused on larger national contexts with greater institutional support, and no equivalent real-time deployable system currently exists for LESH0.

In Honduras, no publicly available real-time software solution currently supports LESH0 recognition on consumer-grade hardware, and the structural barriers facing the deaf community a shortage of only 19 certified interpreters nationwide and limited access to LESH0 education, make this technological gap particularly consequential [2] [3] [4].

Motivated by these challenges, this work proposes a real-time software system for LESH0 translation that leverages landmark-based feature extraction and classical supervised machine learning to deliver an efficient, scalable, and context-aware solution capable of supporting inclusive communication across a wide range of everyday interaction scenarios, without requiring specialized hardware or large institutional datasets.

II. METHODOLOGY

This section describes the methodological framework adopted for the proposed real-time sign language recognition system. The objective of the methodology is to enable accurate and low-latency recognition by combining structured hand pose representations with temporal modeling, while maintaining computational efficiency suitable for real-time deployment.

A. Research Design

This study follows an applied research design oriented toward the development and empirical evaluation of a real-time software system for Honduran Sign Language recognition. The research proceeds through four sequential phases: (1) requirement analysis and technology selection, in which the functional and non-functional constraints of the target deployment environment were identified; (2) system design and implementation, in which the modular architecture, landmark extraction pipeline, and classification models were specified and developed; (3) data collection and preprocessing, in which gesture samples were recorded and transformed into structured feature representations; and (4) experimental evaluation, in which trained models were assessed using standard classification metrics and cross-validation procedures. All experiments were conducted on a laptop equipped with a 3.00 GHz processor and 31.6 GB RAM, to verify feasibility under resource-constrained conditions representative of typical deployment environments in Honduras.

B. Implemented Technologies

The proposed sign language recognition was implemented using a combination of computer vision, machine learning, and real-time processing technologies selected to balance accuracy, computational efficiency, and deployment flexibility.

The OpenCV library is employed for real-time video capture and basic image processing operations. It provides efficient access to camera streams, frame buffering, and image format handling, which are essential for continuous gesture acquisition [22]. OpenCV acts as the low-level interface between the physical camera hardware and the higher-level hand tracking and processing modules.

MediaPipe is used for hand detection and landmark extraction. It employs a machine learning-based pipeline that combines a palm detection model with a hand landmark regression model to identify and track 21 anatomically defined key points per hand in real time.

Each key point is represented in three-dimensional space by normalized (x,y,z) coordinates [23]. By abstracting raw image data into structured geometric representations, MediaPipe enables the system to operate independently of background complexity and lighting variations.

This landmark-based output forms the foundation of the data processing and feature extraction pipeline. Numpy is utilized for numerical computation and feature manipulation.

Landmark coordinates are stored and processed as multidimensional arrays, allowing efficient vectorized operations for normalization, temporal aggregation, and feature scaling [24].

The use of NumPy ensures computational efficiency while maintaining numerical stability across the data processing pipeline and scikit-learn is employed to implement gesture classification models.

Scikit-learn also provides standardized tools for feature scaling, model evaluation, and cross-validation, ensuring consistency between the training and deployment phase of the system [25]. Classical supervised learning algorithms, such as Random Forest classifiers, are trained using the processed landmark-based feature vectors. These models were selected due to their interpretability, fast training time, and low inference latency compared to deep neural networks.

C. Data Collection

Gesture data was collected using the integrated laptop webcam under controlled indoor conditions across multiple laboratory sessions. Participants were positioned at a distance ranging from 30 cm to 1.80 m from the camera, with optimal landmark detection achieved at approximately 70 to 80 cm, where stable and precise keypoint extraction was consistently observed. A total of three participants contributed to the dataset, comprising a reduced group due to the difficulty of accessing certified LESH0 interpreters.

For static gesture categories (alphabet A to Z and numbers 0 to 9), each gesture was captured as individual frames across multiple repetitions per participant, yielding the labeled sample sets described in Table I. For dynamic gesture categories (basic gestures and days of the week), temporal sequences of consecutive frames were recorded per gesture execution to capture motion trajectories. Preprocessing steps including spatial normalization, wrist-relative coordinate transformation, and frame-level filtering were applied uniformly across all samples prior to model training. Efforts are currently underway to expand the dataset by collaborating with local deaf community organizations, with the goal of incorporating a larger and more diverse signer population in future evaluations.

D. Model Definition

The proposed computer vision-based system for LESH0 adopts a modular, landmark-based machine learning architecture composed of multiple specialized recognition models. Each model is designed to address a specific semantic group of sign language gestures, enabling improved classification accuracy, scalability, and efficient real-time performance. This modular approach avoids the limitations of monolithic models and allows independent training, evaluation, and extension of each gesture category.

All models operate on structured spatial features extracted from multi-region body landmarks rather than raw image data. For each video frame, a real-time landmark detection framework identifies a fixed set of anatomical keypoints from the hands, upper body (pose/torso), and face, depending on the gesture category. In particular, each detected hand is represented by 21 anatomical landmarks, while additional landmarks describing upper-body posture and facial configuration are included when required.

Each landmark is encoded by normalized three-dimensional coordinates (x, y, z) relative to the image frame, capturing finger articulation, inter-joint relationships, body posture, and facial orientation [16]. This unified landmark-based representation provides a compact yet expressive description of signing gestures while remaining computationally efficient and robust to background variability.

Depending on the gesture type, the system processes either static or dynamic landmark representations. Static gestures are encoded as fixed-length feature vectors extracted from a single video frame, combining spatial information from the relevant landmark subsets (e.g., hands alone or hands with auxiliary body cues) [6]. Dynamic gestures, in contrast, are modeled as temporal sequences constructed by concatenating landmark vectors across a fixed-length sliding window. When dynamic gestures involve full-body or contextual motion, temporal sequences integrate hand, pose, and facial landmarks to preserve coordinated movement patterns [7] [17]. This temporal modeling strategy enables the system to capture motion dynamics and inter-region dependencies that are critical for accurately distinguishing gestures involving sequential or expressive movement.

Prior to classification, all feature vectors undergo standardization to normalize coordinate distributions and reduce inter-sample variance. Supervised ensemble-based machine learning models are employed for classification due to their robustness to noise, ability to model non-linear decision boundaries, and suitability for structured, medium-dimensional data. Each model is trained independently using labeled datasets corresponding to its specific gesture group.

The division of these models is summarized in Table I. This separation enables task-specific optimization, clearer interpretability, and simplified system maintenance. At the current stage of development, the proposed system incorporates four distinct recognition models, each targeting a well-defined subset of sign language gestures.

TABLE I: SIGN LANGUAGE RECOGNITION MODELS IN THE PROPOSED LESH0 SYSTEM

ID	Category	Samples	Input Type
M1	Alphabet (A–Z)	520,000	Static (126 features)
M2	Numbers (0–9)	10,424	Static (126 features)
M3	Basic Gestures	91	Dynamic (6,480 features)
M4	Days of the Week	117	Dynamic (6,480 features)

This structured model definition enhances system modularity and supports incremental improvements, including user-specific adaptation and the integration of additional gesture categories. Moreover, the proposed design ensures compatibility with real-time interactive applications deployed on consumer-grade hardware.

E. Landmarks Definition

The gesture interpretation software based on computer vision is responsible for analyzing hand movements captured by a camera and converting them into text. The system was implemented in Python using the MediaPipe, OpenCV, and Scikit-

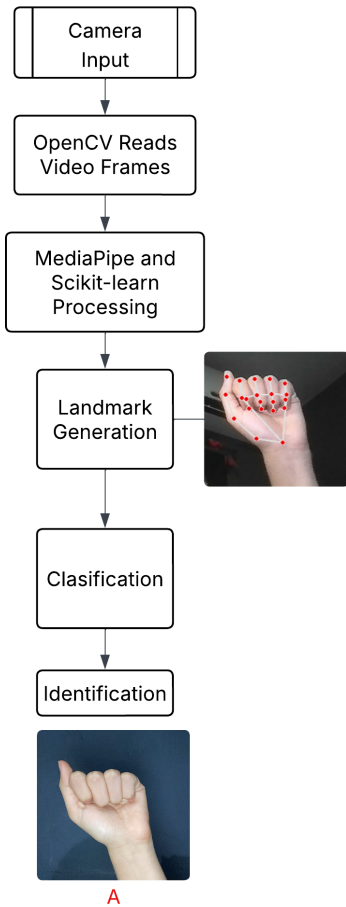


Fig. 1: Workflow Diagram of System Design

learn libraries. MediaPipe is designed for real-time image and video processing through optimized pipelines, enabling robust and efficient hand motion detection without the need for external sensing devices. This combination of libraries ensures a balanced trade-off between accuracy, efficiency, and real-time performance.

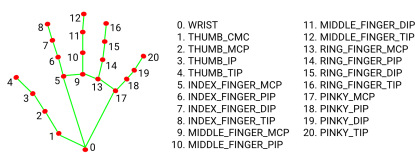


Fig. 2: MediaPipe Hands Model

Fig.1 represents the Workflow hand gesture recognition system is illustrated, which begins with camera input detection followed by landmarks, data processing and classification to identify the signed gesture.

The MediaPipe Hands model was employed, which detects and tracks 21 key hand landmarks within a single video frame while supporting real-time performance for one or both hands. Fig. 2 illustrates the number and position of landmarks across the hand, providing a clear visualization of their spatial distribution.

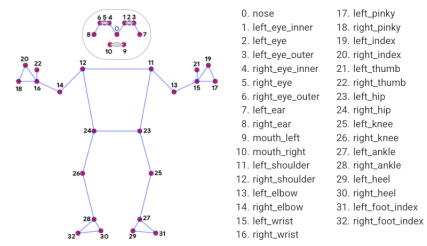


Fig. 3: MediaPipe Holistic Model

Fig. 3 illustrates the landmark configuration of the Medi-aPipe Holistic model, which provides a full-body representation. In this implementation, the framework allows selective disabling of unused landmarks; therefore, only upper-body landmarks were retained, as they are sufficient for the recognition of the target gestures.

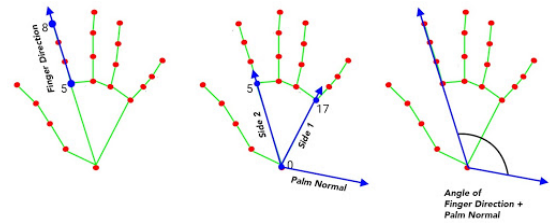


Fig. 4: Spatial Representation of Landmarks

In Fig.4 each landmark is defined by anatomical reference and three-dimensional coordinates (x, y, z), where x and y correspond to spatial coordinates scaled to the frame size, and z represents the depth to the palm plane and provides information to estimate hand orientation and finger positioning.



Fig. 5: Sample Extraction in Software

This spatial representation enables the inference of finger flexion, extension, and palm orientation, supporting the recognition of both static and dynamic gestures.

From an anatomical perspective, the 21 landmarks correspond to key joints and reference points of the human hand, including the wrist, the metacarpophalangeal (MCP) joints, the proximal and distal interphalangeal (PIP/DIP) joints,

and the fingertips. Directly associated with meaningful hand movements and linguistic units in sign language.

In Fig. 5 the landmarks show the process of data extraction to obtain the dataset coordinates that are used for the training phase of the software. Preprocessing steps were applied to normalize variations in scale, rotation, and lighting conditions. By aligning computational landmarks with anatomical structures and linguistic components, the proposed model achieves improved interpretability and robustness, forming a reliable foundation for the real-time translation of LESH0

F. Data Processing

The Data Processing module converts raw RGB video input into structured feature representations suitable for both model training and real-time inference. This module operates as an intermediate stage between the video acquisition subsystem and the gesture classification model.

Each incoming RGB frame is processed by a landmark-based hand tracking algorithm that extracts a fixed set of 21 anatomical keypoints per hand. Each keypoint is represented by three-dimensional coordinates (x, y, z) , forming the spatial description of the hand configuration. The resulting frame-level feature vector for a single hand is defined as:

$$x_t = [x_1, y_1, z_1, \dots, x_{21}, y_{21}, z_{21}] \in \mathbb{R}^{63}. \quad (1)$$

As shown in Eq. (1), each frame is encoded as a fixed-length vector that compactly represents the geometric configuration of the hand.

To reduce variability caused by camera position and hand displacement within the image, spatial normalization is applied to the feature vector defined in Eq. (1). This normalization is performed by expressing all landmark coordinates relative to a reference joint, typically the wrist, improving invariance to global translation and enhancing inter-user consistency.

For dynamic gesture recognition, temporal information is incorporated using a sliding window mechanism that aggregates normalized frame-level feature vectors across consecutive frames. The resulting sequence-level representation is defined as:

$$x_{\text{seq}} = [x_{t-T+1}, \dots, x_t], \quad (2)$$

where each element of the sequence corresponds to a normalized vector as defined in Eq. (1). Equation (2) enables the model to capture both the spatial configuration of the hand and its temporal evolution over time.

Feature scaling is subsequently applied to the sequence representation defined in Eq. (2) using normalization parameters computed from the training dataset. Additionally, frames with incomplete or unreliable landmark detections are discarded prior to sequence construction, ensuring that only valid instances contribute to the final representation.

Through the transformations described in Eqs. (1) and (2), the Data Processing module produces compact, normalized, and temporally coherent feature vectors suitable for low-latency gesture recognition.

G. Model Training

The model training process follows a supervised learning paradigm using the processed feature representations generated by the Data Processing module. Each training sample corresponds either to a single-frame vector defined in Eq. (1) for static gestures, or to a temporal sequence defined in Eq. (2) for dynamic gestures.

For two-handed gestures, the feature vectors corresponding to each hand are concatenated, resulting in a feature dimensionality defined as:

$$d = \begin{cases} 63, & \text{single-hand gestures,} \\ 126, & \text{two-hand gestures.} \end{cases} \quad (3)$$

As indicated in Eq. (3), this formulation allows the model to support both single-hand and two-hand gestures within a unified representation space.

For each gesture class c_i , a labeled dataset is constructed as:

$$D = \{(x_i, y_i)\}_{i=1}^N, \quad (4)$$

where each feature vector $x_i \in \mathbb{R}^d$ follows the dimensionality defined in Eq. (3), and each label $y_i \in \{1, \dots, C\}$ denotes the corresponding gesture class.

The learning objective is to approximate a decision function:

$$f_\theta : \mathbb{R}^d \rightarrow \{1, \dots, C\}, \quad (5)$$

which maps the feature vectors defined in Eq. (4) to their respective gesture classes. The parameters θ of the function in Eq. (5) are learned by minimizing a classification loss over the training subset.

Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrices. After training, the learned parameters of the decision function defined in Eq. (5), along with the preprocessing statistics associated with Eqs. (1) and (2), are serialized and deployed for real-time inference.

H. Classifier Selection and Evaluation Protocol

The Random Forest algorithm was selected as the classification model for all four gesture categories based on properties well-suited to the structure of landmark-based feature vectors. Specifically, Random Forest is capable of capturing non-linear relationships in three-dimensional coordinate data without requiring complex feature transformations, exhibits high tolerance to noise arising from real-time variations in lighting, hand distance, and positioning, and delivers efficient inference suitable for real-time deployment. Compared to k-Nearest Neighbors, which requires exhaustive neighbor search at each prediction, and Support Vector Machines, which may incur high computational cost depending on kernel configuration, Random Forest provides a more favorable balance between accuracy and inference latency for this application. Hyperparameters were initialized using scikit-learn defaults, with minor targeted adjustments guided by training behavior analysis.

The dataset for each gesture category was partitioned using a stratified 80/20 train-test split, preserving class proportions across both subsets. To estimate generalization performance beyond the training distribution, stratified 5-fold cross-validation was applied over the training partition. Cross-validation accuracy and its standard deviation are reported alongside test accuracy in Table II to characterize model stability and identify potential overfitting under the current small-sample conditions. All preprocessing statistics, including feature scaling parameters, were computed exclusively on the training partition to prevent data leakage.

III. RESULTS

Fig. 6-9 represent the performance and accuracy by the proposed system for LESH0 gesture classification across different categories through normalized confusion matrix. Fig. 6 illustrates the spatial distribution of the 21 hand landmarks used in Model M1 for static alphabet recognition, where finger articulation and relative joint positioning encode distinctive patterns for individual letters.

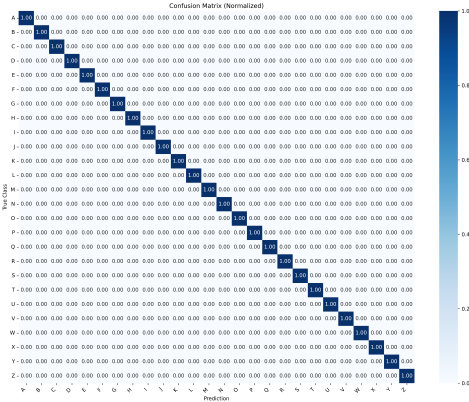


Fig. 6: Normalized confusion matrix representing the accuracy of Model M1.

Fig. 7 presents the corresponding hand landmark representation for numerical gestures, highlighting how differences in finger extension and relative landmark geometry enable reliable discrimination between numerical signs that share similar hand shapes.

Fig. 8 represents the temporal evolution of hand landmarks across consecutive frames in Model M3, capturing motion trajectories and relative landmark displacement that encode dynamic gesture information. This temporal representation allows the system to distinguish gestures with similar static configurations but differing motion patterns, supporting the reported performance for dynamic gestures under controlled evaluation conditions.

Fig. 9 depicts the upper-body landmark representation employed in Model M3 for translating the database of the days of the week, illustrating how the integration of hand motion with contextual body cues such as arm orientation and torso alignment enhances the recognition of complex gestures.

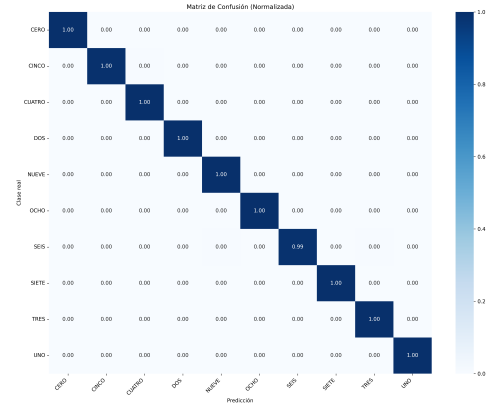


Fig. 7: Normalized confusion matrix representing the accuracy of Model M2.

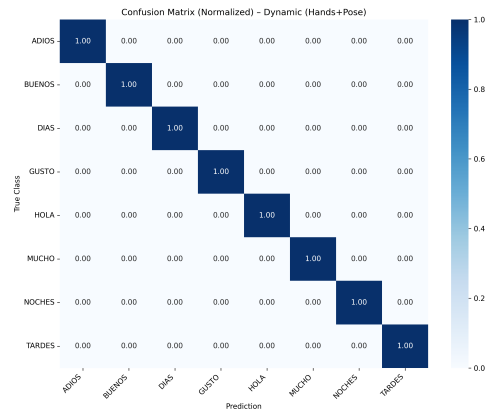


Fig. 8: Normalized confusion matrix representing the accuracy of Model M3.

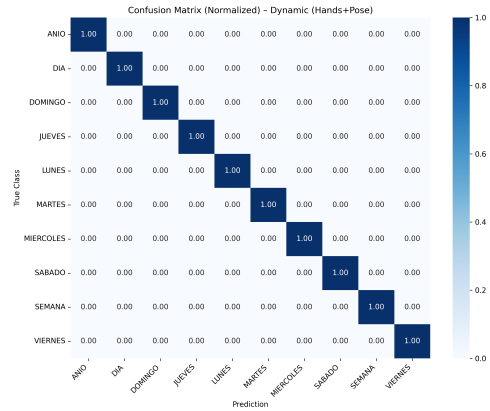


Fig. 9: Normalized confusion matrix representing the accuracy of Model M4.

Together, these figures provide qualitative evidence of how different landmark-based representations encode complementary spatial and temporal information, explaining the performance differences observed across alphabetic, numeric, and dynamic gesture recognition tasks.

The normalized confusion matrices exhibit dominant results values close to 1.0, indicating high class separability for the evaluated gestures, creating separate databases for each category of gestures. This behavior is expected given the controlled acquisition conditions and the use of landmark-based features, which provide discriminative geometric representations for static gestures.

The near-perfect test accuracy observed across all models must be interpreted in the context of the controlled acquisition conditions and the limited signer diversity in the current dataset. To partially mitigate overfitting risk, stratified k-fold cross-validation was applied to each model, yielding cross-validation accuracies that are 4-7 percentage points lower than the corresponding test scores, as reported in Table II. This performance gap is consistent with expected variance under small-sample conditions and highlights the need for expanded inter-subject evaluation.

The alphabet model, which benefits from the largest training set (520,000 samples across 26 classes), exhibits the narrowest gap between test and cross-validation accuracy (99.88% vs. 95.04% $\pm$ 2.98%), suggesting that dataset scale meaningfully supports generalization for static gestures. Future work will prioritize diverse signer recruitment and evaluation under unconstrained environmental conditions to quantify real-world robustness.

TABLE II: PERFORMANCE SUMMARY OF THE PROPOSED LESHORECOGNITION MODELS

Model	Samples	Features	Test Acc.	CV Acc.
Alphabet (A-Z)	520,000	126	99.88%	95.04% $\pm$ 2.98%
Numbers (0-9)	10,424	126	99.86%	92.95% $\pm$ 2.35%
Basic Gestures	91	6,480	100%	94.62% $\pm$ 6.75%
Days of the Week	117	6,480	100%	96.67% $\pm$ 6.67%

A consolidated summary of model performance metrics is presented in Table II, reinforcing the quantitative results discussed throughout this section and highlighting the scalability of the proposed approach.

The experimental results demonstrate the effectiveness and robustness of the modular modeling strategy. The dynamic gesture model, trained on 91 temporal samples with 6,480 features per sequence and covering eight dynamic gestures, achieved a test accuracy of 100% and a cross-validation accuracy of 94.62% $\pm$ 6.75%, with a training time of only 0.42 seconds. These results demonstrate promising performance under controlled evaluation conditions. While results are encouraging, larger and more diverse datasets are required to fully assess generalization for dynamic gestures.

The numerical gesture model exhibited similarly strong performance. Trained on 10,424 static samples across ten numerical classes and using 126 normalized features derived from one or two hands, the model achieved a test accuracy of 99.86% and a cross-validation accuracy of 92.95% $\pm$ 2.35%. Its fast training time (under one second) and scalability make it well suited for large-scale deployment and incremental dataset expansion.

For temporal semantic gestures, the days-of-the-week model was trained using 117 dynamic sequences representing ten temporal gestures. This model achieved a test accuracy of 100% and a cross-validation accuracy of 96.67% $\pm$ 6.67%, while maintaining efficient training times below one second. These results validate the effectiveness of the temporal feature representation for structured, sequential gestures.

The alphabet recognition model constitutes the most extensive component of the system. Trained on 520,000 static samples evenly distributed across 26 alphabet classes, with 126 normalized features per sample, the model achieved a test accuracy of 99.88% and a cross-validation accuracy of 95.04% $\pm$ 2.98%. Despite the large dataset size, the complete training process was completed in approximately 2.25 minutes, highlighting the computational efficiency and scientific robustness of the proposed approach.

IV. DISCUSSION

A. System Performance and User Experience

The proposed system demonstrates strong real-time performance using two landmark configurations: a hand-only representation and a combined hand-face-torso representation. The hand-only configuration provides low-latency inference and stable performance for isolated gestures, making it suitable for resource-constrained environments. The multimodal configuration improves recognition robustness and expressiveness, particularly for gestures involving facial cues or upper-body posture, while maintaining interactive response times on consumer-grade hardware. From a user perspective, both configurations allow natural gesture execution with immediate visual feedback, although recognition quality remains influenced by lighting conditions, camera placement, and landmark visibility.

B. Limitations

Despite the improved expressiveness of the multimodal configuration, the system remains limited by its focus on isolated gesture recognition, restricting full sentence-level interpretation. Landmark-based approaches are sensitive to occlusions, fast movements, and extreme viewing angles, with these effects becoming more pronounced when multiple landmark sets are tracked simultaneously. Additionally, the dataset was collected from a limited number of signers under controlled conditions, which may affect generalization to diverse users. While classical machine learning models offer computational efficiency, they may not fully capture complex spatiotemporal dependencies present in continuous sign language.

C. Implementation Challenges

Supporting multiple landmark configurations introduced challenges related to feature dimensionality, normalization, and synchronization. Ensuring consistent preprocessing across hand-only and multimodal inputs was essential for fair comparison and stable inference. The increased feature space in the multimodal configuration required careful optimization to

preserve real-time performance, particularly for dynamic gestures. Maintaining reliable landmark detection across varying camera distances and body orientations also required iterative parameter tuning.

D. Future Work and System Enhancements

Future work will focus on extending both configurations toward continuous sign language recognition through improved temporal segmentation and sequence modeling. Expanding the dataset to include a larger and more diverse signer population will be critical for improving robustness and generalization. Additional optimizations for mobile and embedded deployment are planned, along with the integration of higher-level language modeling and speech synthesis modules to enable complete assistive communication systems.

E. Practical Implications

The proposed system has direct implications for inclusive communication in Honduras, where over 100,000 people with hearing impairments face structural barriers due to a shortage of certified LESHO interpreters and limited access to sign language education [2] [3]. By enabling real-time gesture-to-text translation on consumer-grade hardware without internet connectivity or specialized sensors, the system offers a deployable tool for contexts such as healthcare consultations, public service counters, and educational settings where human interpreters are unavailable.

From an institutional perspective, the modular architecture facilitates incremental vocabulary expansion, allowing organizations such as the National Autonomous University of Honduras (UNAH) or government agencies to adapt the system to domain-specific terminology without retraining the full pipeline [4]. The low training time demonstrated across all models, under 2.25 minutes even for the largest 520,000-sample alphabet model, further supports rapid iteration and community-driven dataset expansion.

More broadly, the landmark-based approach demonstrated here is language-agnostic and could be extended to other underrepresented national sign languages across Latin America, contributing to regional efforts in assistive technology development, as evidenced by recent work on Mexican Sign Language [20] and Argentine Sign Language [21].

V. CONCLUSIONS

This work presented the design, implementation, and evaluation of the proposed computer vision-based system for LESHO, a real-time sign language recognition framework based on hand landmark representations and classical machine learning models. By leveraging structured spatial and temporal features instead of raw image data, the proposed approach achieves high recognition accuracy while maintaining low computational cost and real-time feasibility on consumer-grade hardware.

Additionally, the proposed design not only implements single sign language detection, but also supports dynamic sign recognition by analyzing sequences of frames, which allows

the system to accurately interpret gestures involving temporal evolution and movement patterns. This capability extends the applicability of the framework beyond static signs, enabling more natural and expressive communication for users and increasing the overall usability of the system in real-world scenarios.

Cross-validation results confirm that the performance observed under controlled evaluation conditions partially generalizes beyond the training distribution, with cross-validation accuracies consistently above 92% across all four models. The gap between test and cross-validation accuracy underscores the importance of dataset diversity, and future efforts to recruit broader signer populations will be critical for validating real-world robustness.

Overall, the proposed LESHO system demonstrates that landmark-based representations combined with classical supervised learning techniques can deliver high accuracy, robustness, and real-time performance for sign language recognition tasks. The modular architecture enables task-specific optimization, simplifies system maintenance, and facilitates future extensions, including personalized user adaptation, additional gesture categories, and multilingual sign language support.

REFERENCES

- [1] L. M. A. García, "DIARIO OFICIAL DE LA REPÚBLICA DE HONDURAS DECANO DE LA PRENSA HONDUREÑA PARA MEJOR SEGURIDAD DE SUS PUBLICACIONES," no. 33, 2014.
- [2] Comisionado Nacional de los Derechos Humanos, "Discriminados en su "silencio" más de 100,000 personas sordas en Honduras – Comisionado Nacional de los Derechos Humanos," Mar. 2024. [Online]. Available: <https://conadeh.hn/?p=1907>
- [3] W. Holly, "A Sociolinguistic survey of the Honduran Deaf community." [Online]. Available: <https://www.sil.org/resources/publications/entry/9181>
- [4] U. A. de Honduras, "Inauguran el curso Lesho: Lengua de Señas Hondureñas." [Online]. Available: <https://blogs.unah.edu.hn/dircom/unah-inaugura-curso-de-lesho-el-lengua-de-senas-de-honduras/>
- [5] S. Chadha, K. Kamenov, and A. Cieza, "The world report on hearing, 2021," *Bulletin of the World Health Organization*, vol. 99, no. 4, pp. 242–242A, Apr. 2021. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8085630/>
- [6] S. Dhamodaran, P. Phukan, M. Singh, and S. Nandakum, "Implementation of Hand Gesture Recognition using OpenCV," *International Journal of Research Publication and Reviews*, vol. 5, pp. 8039–8050, May 2024.
- [7] F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," *Neural Computing and Applications*, vol. 37, no. 2, pp. 841–857, Jan. 2025. [Online]. Available: <https://doi.org/10.1007/s00521-024-10395-9>
- [8] Y. Meng, H. Jiang, N. Duan, H. Wen, Y. Meng, H. Jiang, N. Duan, and H. Wen, "Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System," *Sensors*, vol. 24, no. 19, Sep. 2024, company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher. [Online]. Available: <https://www.mdpi.com/1424-8220/24/19/6262>
- [9] J. E. Mora-Zarate, C. L. Garzón-Castro, and J. A. Castellanos-Rivillas, "LSC-54: A landmark-based dataset for colombian sign language," *Data in Brief*, vol. 63, p. 112145, Dec. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352340925008662>
- [10] A. Anand, A. Pandey, S. Mehndiratta, H. Goyal, P. Kaushik, and R. Rathore, *Smart AI based Volume Control System: Gesture Recognition with OpenCV & Mediapipe Integration*, May 2024, pages: 6.
- [11] V. Adithya and R. Rajesh, "Hand gestures for emergency situations: A video dataset based on words from Indian sign language," *Data in Brief*, vol. 31, p. 106016, Aug. 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352340920309100>

- [12] M. D. Dere, R. O. Dere, A. Adesina, and A. R. Yauri, "An end-to-end framework for translation of American sign language to low-resource languages in Nigeria," *Scientific African*, vol. 21, p. e01809, Sep. 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S246822762300265X>
- [13] B. Márquez, A. Alanis, A. Quezada, and J. Magdaleno-Palencia, "Development of a Mobile Application with Artificial Intelligence for Mexican Sign Language Recognition," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 19, pp. 122–139, May 2025.
- [14] A. Kasapbaşı, A. E. A. Elbushra, O. Al-hardanee, and A. Yilmaz, "DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100048, Jan. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666990021000471>
- [15] "Intelligent Sign Language Recognition System for E-Learning Context," *Computers, Materials and Continua*, vol. 72, no. 3, pp. 5327–5343, Apr. 2022. [Online]. Available: <https://www.sciencedirect.com/org/science/article/pii/S154622182200902X>
- [16] P. Gogoi, B. Karsh, R. K. Karsh, R. H. Laskar, and M. K. Bhuyan, "Vision-Based Real-Time Gesture-to-Speech Translation for Sign Language," *Procedia Computer Science*, vol. 258, pp. 2050–2059, Jan. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050925015583>
- [17] R. Dagde, S. Thakre, S. Chopade, L. Rokde, and V. Kakani, "A hand sign recognition based signal system for mute people using machine learning," *MethodsX*, vol. 15, p. 103670, Dec. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S221501612500514X>
- [18] A. Tyagi, V. S. Prasanth, P. Lunawat, and S. Ashok, "Real-time sign language gesture recognition for Women's safety using dynamic time warping and voting algorithms," *Franklin Open*, vol. 13, p. 100428, Dec. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2773186325002166>
- [19] M. Rodriguez, O. Oubram, A. Bassam, N. Lakouari, and R. Tariq, "Mexican Sign Language Recognition: Dataset Creation and Performance Evaluation Using MediaPipe and Machine Learning Techniques," *Electronics*, vol. 14, no. 7, p. 1423, Jan. 2025. [Online]. Available: <https://www.mdpi.com/2079-9292/14/7/1423>
- [20] G. Garcia-Gil, G. d. C. López-Armas, and Y. E. Ramirez-Pulido, "Multimodal Minimal-Angular-Geometry Representation for Real-Time Dynamic Mexican Sign Language Recognition," *Technologies*, vol. 14, no. 1, p. 48, Jan. 2026. [Online]. Available: <https://www.mdpi.com/2227-7080/14/1/48>
- [21] X. Liu, Z. Zhou, E. Xia, and X. Yin, "SSTA-ResT: Soft Spatiotemporal Attention ResNet Transformer for Argentine Sign Language Recognition," *Sensors*, vol. 25, no. 17, p. 5543, Jan. 2025. [Online]. Available: <https://www.mdpi.com/1424-8220/25/17/5543>
- [22] "OpenCV: Introduction." [Online]. Available: <https://docs.opencv.org/4.x/d1/dfb/intro.html>
- [23] "Hand landmarks detection guide | Google AI Edge | Google AI for Developers." [Online]. Available: https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker
- [24] "NumPy documentation — NumPy v2.4 manual." [Online]. Available: <https://numpy.org/doc/stable/>
- [25] "scikit-learn: machine learning in Python — scikit-learn 1.8.0 documentation." [Online]. Available: <https://scikit-learn.org/stable/index.html>