Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 849-864
Journal homepage: http://iieta.org/journals/isi

# Hybrid Spatial–Temporal Stabilization for Real-Time Static Arabic Hijaiyah Letter Recognition Using Polynomial Regression

Dadang Iskandar Mulyana , Edi Noersasongk , Guruh Fajar Shidi , Pujion

Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia

Corresponding Author Email: mahvin2012@gmail.com



https://doi.org/10.18280/isi.310317

**Received:** 8 November 2025
**Revised:** 20 January 2026
**Accepted:** 18 March 2026
**Available online:** 31 March 2026

***Keywords:***
*sign language recognition, Arabic Hijaiyah letters, polynomial regression, spatial–temporal stabilization, real-time recognition, MediaPipe*

**ABSTRACT**

Static Arabic Hijaiyah letter recognition plays a crucial role in facilitating communication for individuals with hearing impairments. However, achieving stable real-time recognition remains challenging. This limitation arises from temporal jitter, geometric inconsistency, and variations in hand distance and orientation. These challenges are particularly pronounced in static gestures, where subtle inter-finger differences significantly affect classification accuracy. To address these limitations, this study proposed a hybrid spatial–temporal stabilization framework. The framework integrates polynomial regression smoothing, distance-based temporal filtering, landmark normalization, and palm-orientation encoding to improve geometric stability and feature discriminability. A dataset comprising more than 44,000 samples across 28 Hijaiyah letters was collected in both raw and stabilized forms. The method was evaluated under offline, session-aware, and real-time scenarios. Experimental results demonstrated that the proposed approach reduced landmark variance and inter-frame jitter. This improvement enhanced gesture separability and recognition robustness. In real-time evaluation, the proposed method improved accuracy from 92.45% to 95.95% and F1-score from 92.71% to 95.75%. These findings confirm the effectiveness of the proposed framework for practical real-time deployment. The proposed approach provides a robust preprocessing strategy for future multimodal and continuous sign language recognition systems.

## 1. INTRODUCTION

Sign language serves as the primary communication medium for the Deaf and Hard of Hearing (HoH) community, enabling information exchange through hand configurations, finger articulation, and spatial movements [1, 2]. Recent advances in computer vision and machine learning have accelerated the development of Sign Language Recognition (SLR) systems capable of automatically identifying hand gestures from RGB video streams [3-5]. Despite this progress, achieving stable real-time recognition remains challenging. This limitation arises from geometric inconsistencies in detected hand landmarks caused by variations in camera distance, hand orientation, illumination, and motion dynamics [6, 7]. Such instability produces temporal jitter and spatial distortion, which degrade feature consistency and classification performance.

Several techniques have been explored to mitigate landmark fluctuations, including Kalman filtering, Gaussian Mixture Models (GMM), and particle filtering [8-12]. Although effective in specific scenarios, these approaches often involve substantial computational overhead or exhibit limited robustness against distance-induced scaling effects. Meanwhile, landmark-based detectors such as MediaPipe Hands provide efficient and reliable real-time tracking [13, 14], but they lack intrinsic stabilization mechanisms. As a result, these methods remain susceptible to inter-frame noise, particularly during subtle or rapid hand movements.

Advanced SLR models based on CNNs, RNNs, and Vision Transformers have demonstrated strong classification performance [15], yet most focus on feature representation rather than on stabilizing geometric structures of the hand. Consequently, recognition accuracy often deteriorates under dynamic environmental conditions or modest deviations in camera placement [15]. Prior studies have attempted noise reduction—such as Kalman-based temporal filtering or fixed-distance normalization [16-18]—but these approaches struggle with nonlinear landmark deformation or lack adaptability to hand-motion dynamics.

This gap highlights the need for a lightweight, geometry-aware stabilization technique capable of reducing landmark jitter while preserving the topological relationships among fingers and the palm. Polynomial regression smoothing, including Savitzky–Golay filtering, has been widely applied to temporal signals for noise reduction while maintaining shape fidelity [16, 17], yet its application to multi-point 2D/3D landmark trajectories remains underexplored. Likewise, inter-finger distance averaging may reduce hand-size variation [18], but static approaches fail to accommodate real-time motion.

To address these limitations, this study proposes a hybrid spatial–temporal stabilization framework that integrates polynomial regression smoothing and dynamic distance-based

normalization to enhance the geometric consistency of MediaPipe landmark sequences. The method aims to mitigate temporal jitter, maintain inter-limb proportionality, and improve class separability without increasing computational complexity. The contributions of this work are threefold:

(1) introducing a hybrid polynomial–distance smoothing mechanism tailored for real-time landmark stability;

(2) developing an enhanced feature representation based on stabilized landmarks, relative coordinates, and palm-orientation cues; and

(3) providing empirical validation through extensive testing on static Arabic–Hijaiyah gestures under varying camera angles and distances.

In this study, the scope is explicitly limited to static Hijaiyah letter gestures, enabling a focused evaluation of geometric stabilization without involving temporal sequence modeling.

## 2. RELATED WORKS

### 2.1 Vision-based sign language recognition

SLR has evolved rapidly over the past decade, driven by the increasing demand for assistive communication technologies for the Deaf and Hard of Hearing community. Early SLR systems relied heavily on static image classification using Convolutional Neural Networks (CNNs) [1, 4]. These approaches treated each gesture as a fixed spatial pattern. While effective for isolated signs, these models struggled to capture temporal dependencies inherent in hand-motion sequences. This limitation motivated the adoption of temporal architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer-based models [9, 12], capable of modeling frame-to-frame dynamics and improving sequence-level accuracy.

However, pixel-based methods remain sensitive to real-world variations, including lighting fluctuations, background clutter, and changes in hand orientation or camera distance [6, 7]. Consequently, recent research has shifted toward pose-based approaches using efficient landmark detectors such as MediaPipe Hands, which provide 21 real-time hand key points with low computational cost [13, 14]. Although landmark-based systems address many limitations of pixel features, they still suffer from inter-frame geometric instability, especially during rapid hand movements or subtle camera shifts. Addressing this instability motivates the development of dedicated geometric stabilization mechanisms.

### 2.2 Temporal smoothing and geometric stabilization in hand tracking

Temporal stabilization plays a key role in improving the reliability of real-time gesture tracking. The Kalman Filter is the most commonly employed technique for estimating motion trajectories while reducing temporal noise [8, 10]. Variants such as Adaptive and Ensemble Kalman Filters improve flexibility in dynamic environments [8], yet remain limited when faced with nonlinear movements such as finger flexion or combined translations and rotations.

Alternative probabilistic techniques include Gaussian Mixture Models (GMMs) [11] and Particle Filtering [12]. These methods can model nonlinear distributions more effectively. However, they often incur high computational costs, making them less suitable for lightweight real-time SLR pipelines [13]. As a result, there is a strong need for stabilization approaches that are computationally efficient and easily integrated with modern landmark detectors. Such approaches should also preserve geometric consistency across frames.

Polynomial regression offers a compelling alternative. By modeling a landmark's temporal trajectory as a smooth polynomial curve, high-frequency jitter can be suppressed. This can be achieved without distorting the underlying motion pattern [16, 17]. The Savitzky–Golay formulation has also been shown to preserve local geometric characteristics and fine structural details during smoothing [19, 20]. Recent landmark-based gesture studies further demonstrate the importance of automatic landmark localization. MediaPipe-based feature extraction has also proven effective for real-time hand analysis [21, 22].

### 2.3 Polynomial regression for landmark geometry preservation

Polynomial regression has long been utilized for curve fitting and signal denoising due to its ability to maintain local structure while reducing noise [16, 17]. The Savitzky–Golay formulation, as introduced by Schafer [16] and further discussed by Gallagher [19], has been widely recognized for its capability to preserve peak geometry and fine structural details during smoothing. Additional studies have also demonstrated its effectiveness in maintaining signal characteristics while reducing noise distortion [20].

In SLR contexts, each landmark point—including fingertips, joints, and the wrist—can be interpreted as a multidimensional temporal signal subject to frame-to-frame fluctuations. Applying polynomial smoothing to each landmark produces continuous and stable trajectories while preserving the topological structure of the hand. Unlike Kalman-based approaches, polynomial regression does not require covariance estimation and remains computationally efficient, making it suitable for real-time applications [23].

Despite these advantages, its application in landmark-level geometric stabilization remains limited in existing SLR research, indicating a clear research gap.

### 2.4 Distance smoothing and scaling normalization

Beyond temporal jitter, variations in the distance between the user's hand and the camera can produce significant scale distortions in landmark coordinates. Changes in inter-finger distance can disproportionately affect feature vectors. This can degrade classifier performance [12, 18].

Zhao et al. [18] introduced Distance Increment Smoothing to stabilize inter-landmark variation. However, their method did not account for topological relationships across fingers or dynamic spatial deformation.

Distance normalization techniques convert absolute distances into ratios relative to reference segments such as palm length. These methods have demonstrated effectiveness in mitigating camera scaling effects [21, 22].

When combined with polynomial smoothing, such normalization provides a complementary spatial adjustment layer. This results in a two-stage stabilization mechanism that aligns both temporal and geometric consistency.

### 2.5 Hybrid geometric stabilization in modern sign language recognition

Hybrid stabilization approaches combining temporal filtering and spatial correction have been explored in object tracking and human motion analysis [23, 24]. In SLR, hybrid Kalman–GMM and wavelet-based smoothing techniques have been proposed to improve stability. However, these approaches often require complex parameter tuning and incur higher computational costs [8, 11, 24].

The hybrid polynomial regression and distance smoothing strategy proposed in this study differs from prior approaches. It performs temporal alignment through polynomial smoothing and spatial normalization through adaptive inter-finger distance modeling. This integration improves both temporal stability and geometric consistency.

Prior empirical studies have reported improvements in temporal stability and classification performance for complex gesture sets [25-27]. Visualization techniques such as confusion matrices, radar plots, and PCA–UMAP projections further demonstrate improved class separability and robustness, particularly for gestures with similar hand configurations [28-30].

These findings highlight the potential of hybrid geometric stabilization as a lightweight and effective preprocessing layer for modern real-time SLR pipelines.

### 2.6 Relation to transformer-based vision models

Recent advancements in SLR have introduced transformer-based architectures such as Vision Transformer (ViT), DeiT, and Vision Mamba, which capture global spatial relationships more effectively than conventional CNN or RNN-based models [30].

However, these models typically operate on raw or minimally processed landmark sequences and lack explicit mechanisms for geometric stabilization. Prior studies have shown that improved classifier performance does not necessarily guarantee stable landmark geometry under real-world conditions [2, 4].

The hybrid stabilization framework proposed in this study is fully compatible with these architectures and can serve as an efficient preprocessing stage. By providing stabilized and geometrically consistent landmark representations, the method enhances the reliability of downstream transformer-based SLR models, particularly in real-time and embedded deployment scenarios.

Recent advances in deep learning for visual sequence recognition have demonstrated improved robustness and adaptability in complex pattern recognition tasks [31, 32].

## 3. METHODS

This study proposes a hybrid geometric stabilization framework designed to improve the accuracy and temporal consistency of a vision-based SLR system. The stabilization layer is positioned between the landmark detection stage and the gesture classification module, and operates directly on the 21 landmarks provided by the MediaPipe Hands skeleton model. The objective is to suppress temporal jitter and scale variation while preserving the anatomical structure of the hand, thereby producing more discriminative features for Hijaiyah gesture recognition, and Figure 1 illustrates the canonical landmark configuration adopted in this work.

The MediaPipe Hands framework defines 21 hand landmarks distributed across the wrist, finger joints, and fingertips. Landmark 0 corresponds to the wrist, while the remaining landmarks cover the MCP, PIP, DIP, and TIP joints of each finger. This configuration enables detailed modeling of finger articulation and palm geometry. In this study, the raw (x, y, z) coordinates of these 21 landmarks form the baseline feature set and serve as the input to the proposed hybrid stabilization pipeline.

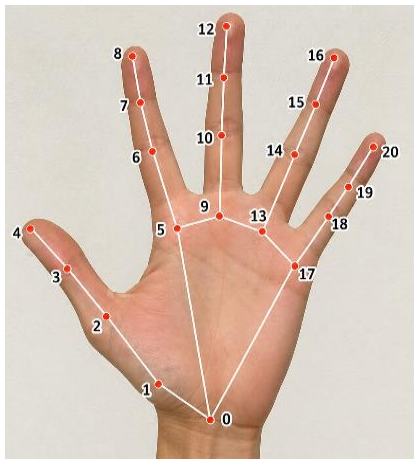



**Figure 1.** Configuration of 21 hand landmarks based on the MediaPipe hand tracking model

### 3.1 Polynomial regression layer

The Polynomial Regression Layer is responsible for temporally smoothing the trajectory of each landmark. Let $D_t$ denote a generic distance or coordinate value at frame $t$. A polynomial regression model of order $n$ is defined as

$$\hat{D}_t = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n. \quad (1)$$

An exploratory study was conducted on more than 1,000 randomly sampled sequences (60–90 frames each) to select the appropriate polynomial order. The results are summarized in Table 1.

**Table 1.** Polynomial order selection

| Polynomial Order | Avg Error | Overfitting | Smoothness |
|---|---|---|---|
| n = 1 | Too rigid | Low | Poor |
| n = 2 | Low | None | Good |
| n = 3 | Lowest | None | Best |
| n ≥ 4 | --- | High | Unstable |

Based on these observations, a hybrid polynomial scheme is adopted. Palm-region landmarks—0, 1, 5, 9, 13, and 17—are filtered using a second-order polynomial ($n = 2$) due to their relatively smooth and low-curvature motion. In contrast, finger-joint and fingertip landmarks—2-4, 6-8, 10-12, 14-16, and 18-20—are smoothed with a third-order polynomial ($n = 3$), allowing the model to capture natural finger curvature while suppressing high-frequency noise. This selective configuration improves temporal stability without oversmoothing fine geometric patterns that are essential for discriminating between similar Hijaiyah letters.

A minimum of 10 frames per trajectory is required for

reliable polynomial fitting. The temporal filtering is applied independently to each landmark coordinate (x, y, z) prior to further processing.

### 3.2 Distance smoothing and geometric normalization

To address spatial variations caused by changes in hand-to-camera distance, the second layer of the pipeline operates on inter-landmark distances. For a pair of landmarks $(i, j)$, the Euclidean distance in the normalized MediaPipe space is defined as

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}. \quad (2)$$

Let:

- $D_{ij}(t)$ : raw distance between landmarks $i$ and $j$ at time $t$,
- $D_{\text{ref}}(t) = D_{0,9}(t)$ : reference distance between the wrist (0) and the middle-finger base (9),
- $D_{\text{ref}}^{\text{cm}}$ : fixed real-world reference distance in centimeters.

An adaptive normalization is then defined as

$$D_{ij}^{\text{norm}}(t) = \frac{D_{ij}(t)}{D_{\text{ref}}(t)} \times D_{\text{ref}}^{\text{cm}} \quad (3)$$

This formulation converts all distances into a camera-distance-invariant representation, where inter-finger distances are expressed relative to a stable anatomical reference. The choice of the (0, 9) pair as reference is motivated by its comparatively low sensitivity to local finger movement and its robustness across different orientations, making it suitable as a scale anchor.

In implementation, distances are initially measured in pixel units (derived from image coordinates) and converted into centimeters using the general form.

$$\text{Distance}_{\text{cm}} = \frac{\text{Distance}_{\text{px}}}{\text{RefDist}_{\text{px}}} \times \text{RefDist}_{\text{cm}} \quad (4)$$

where, $\text{RefDist}_{\text{px}}$ is the pixel distance between the chosen reference landmarks and $\text{RefDist}_{\text{cm}}$ is its real-world length. This conversion ensures geometric consistency when the user moves closer to or farther from the camera.

### 3.3 Exponential smoothing layer

To further suppress high-frequency noise in distance-based features, an exponential smoothing layer is applied on top of the normalized distances. For a given normalized distance sequence $D_t^{\text{norm}}$, exponential smoothing is defined as

$$S_t = \alpha D_t^{\text{norm}} + (1-\alpha) S_{t-1}, \quad (5)$$

where, $S_t$ denotes the smoothed value at time $t$ and $\alpha \in (0,1]$ is the smoothing factor. A parameter sweep was performed for $\alpha \in \{0.1, 0.2, 0.3, 0.5\}$ , yielding the qualitative behavior summarized in Table 2.

In this work, α = 0.2–0.3 is used depending on gesture type and session conditions, providing a practical trade-off between responsiveness and temporal stability.

**Table 2.** Smoothing parameter behavior

| α | Stability | Response Time |
|---|---|---|
| 0.1 | significantly smooth | slow |
| 0.2 | smooth | moderate |
| 0.3 | balanced | fast |
| 0.5 | noisy | significantly fast |

### 3.4 Hybrid stabilization workflow

The proposed hybrid model integrates the three layers described above into a unified stabilization pipeline:

- Temporal filtering (Polynomial Regression): Each landmark coordinate is smoothed using a second- or third-order polynomial depending on its anatomical role, reducing jitter while preserving natural motion.
- Spatial normalization (Distance Smoothing): Inter-finger distances and reference distances are normalized using Eq. (3) and converted into a camera-distance-invariant representation via Eq. (4).
- Output refinement (Exponential Smoothing): Normalized distances are further filtered using exponential smoothing (Eq. (5)) to attenuate residual noise and micro-oscillations.

The result is a temporally coherent and scale-invariant description of hand geometry that can be directly used as input to lightweight classifiers such as k-Nearest Neighbors (k-NN).

### 3.5 Reproducibility settings

To ensure reproducibility of the proposed framework, all experiments were conducted under the following settings:

- Frame rate: 30 FPS; resolution: 720 × 480
- Maximum polynomial order: 3 (with hybrid order-2 and order-3 assignment as described in Section 3.1)
- Exponential smoothing factor: 0.2–0.3
- Reference landmark pair for scale normalization: (0, 9)
- Minimum frames for polynomial fitting: 10
- Each gesture recorded in at least 5 cycles per session
- Implementation: Python with MediaPipe Hands, NumPy, and scikit-learn.

These settings are kept constant across baseline and proposed experiments to enable fair comparison.

### 3.6 Experimental setup

The baseline dataset was constructed using a purely vision-based tracking approach. The MediaPipe Hands model was employed to extract 21 real-time landmarks from the right hand, yielding 63 raw features $(x_0, y_0, z_0), \ldots, (x_{20}, y_{20}, z_{20})$ per frame. MediaPipe was selected due to its robustness, high frame rate, and suitability for commodity RGB cameras. A total of 28 static Arabic/Hijaiyah gesture classes were recorded, from Alif to Ya, across multiple recording sessions with controlled variations in camera distance, hand rotation, and illumination. All samples were stored in CSV format, preserving the same column structure—timestamp, frame index, camera distance estimate, gesture label, and landmark coordinates. Figure 2 provides an overview of the 28 Hijaiyah hand gestures included in the dataset.

This baseline dataset represents the unprocessed landmark

distribution used both as a reference for comparison and as input to the subsequent stabilization pipeline.

3.6.1 Baseline definition

The baseline model in this study represents a conventional landmark-based recognition pipeline without any explicit geometric stabilization mechanism. Specifically, the baseline utilizes raw 21-point hand landmarks extracted directly from the MediaPipe Hands framework, preserving the original (x, y, z) coordinates across frames.

Unlike the proposed method, the baseline does not incorporate:

(i) polynomial regression smoothing,
(ii) distance-based normalization,
(iii) exponential smoothing, or
(iv) orientation encoding.

This design ensures that the baseline reflects a commonly adopted lightweight landmark-based recognition approach in the literature.

For fair comparison, both baseline and proposed methods share identical dataset, classifier, and evaluation protocol.

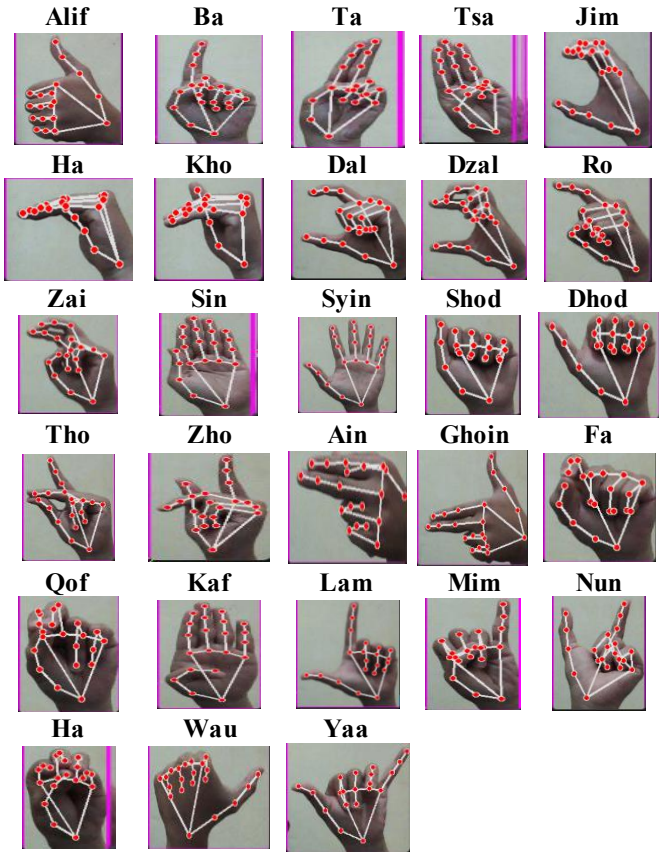



**Figure 2.** The 28-class Arabic/Hijaiyah hand gesture set

3.6.2 Proposed method definition

The proposed method extends the baseline landmark-based representation by introducing a hybrid spatial–temporal stabilization pipeline. This framework integrates multiple complementary components designed to enhance geometric consistency and temporal coherence of hand landmark features.

Specifically, the proposed method incorporates:

(i) polynomial regression smoothing to reduce temporal jitter in landmark trajectories,
(ii) distance-based normalization to mitigate scale variation caused by hand-to-camera distance changes,
(iii) exponential smoothing to suppress residual high-frequency noise, and
(iv) palm-orientation encoding to enrich discriminative spatial features.

All transformations are applied sequentially to the raw landmark coordinates, producing a stabilized and scale-invariant feature representation.

Importantly, the proposed method operates under the same experimental conditions as the baseline, including identical dataset, classifier configuration, and evaluation protocols. This ensures that performance improvements are solely attributed to the effectiveness of the hybrid stabilization framework.

### 3.7 Proposed dataset collection (hybrid-stabilized landmark acquisition)

A second dataset, referred to as the proposed hybrid-stabilized dataset, was collected using the same recording setup but with the hybrid stabilization pipeline applied online during acquisition. For each frame, the 21 landmarks from MediaPipe were processed through four sequential modules:

3.7.1 Polynomial regression smoothing

Temporal jitter and micro-oscillations in each landmark trajectory were attenuated using the hybrid polynomial scheme. Palm-related landmarks (0, 1, 5, 9, 13, 17) were smoothed with second-order polynomials, while fingertip and joint landmarks (2–4, 6–8, 10–12, 14–16, 18–20) used third-order polynomials. This configuration preserves subtle curvature in finger movements that is crucial for discriminating visual similarities among Hijaiyah letters.

3.7.2 Distance smoothing using exponential filtering

Camera distance and all inter-finger distances were further stabilized with exponential smoothing using a small smoothing factor ($\alpha$ = 0.2–0.3). This step reduces high-frequency depth noise and fluctuations caused by hand translation toward or away from the camera, resulting in more coherent modeling of hand scale and finger spread.

3.7.3 Relative landmark feature transformation

To reduce sensitivity to global hand displacement within the image plane, all landmarks were transformed to a wrist-centered coordinate system:

$$(x_i', y_i', z_i') = (x_i - x_0, y_i - y_0, z_i - z_0),$$

where, $(x_0, y_0, z_0)$ corresponds to the wrist landmark. This Relative Landmark Feature (RLF) representation retains the intrinsic geometry of the hand while discarding absolute positional offsets, improving robustness to user movement and framing differences.

3.7.4 Palm normal and orientation encoding

To incorporate 3D hand-orientation information, a palm normal vector $\mathbf{n} = (n_x, n_y, n_z)$ was computed using the cross product of vectors $\overrightarrow{P_0P_5}$ and $\overrightarrow{P_0P_{17}}$.

The resulting normal was normalized and converted into yaw and pitch angles:

$$\text{yaw} = \arctan 2\left(n_y, n_x\right), \text{pitch} = \arctan 2\left(n_z, \sqrt{n_x^2 + n_y^2}\right).$$

These orientation descriptors capture pronation, supination, and palm tilt, which are particularly important for discriminating visually similar Hijaiyah gestures.

The final proposed dataset thus combines:

- hybrid polynomial filtering,
- distance and camera-depth smoothing,
- RLF-based geometric normalization, and
- palm normal orientation features.

All data were stored in a unified CSV structure, enabling direct comparison with the baseline dataset.

### 3.8 Evaluation protocols

To rigorously assess the impact of the proposed stabilization framework, two complementary evaluation protocols were employed: (i) offline temporal separability analysis and (ii) session-aware robustness testing.

3.8.1 Offline evaluation (80/20 temporal split)

For the offline evaluation, a maximum of 400 frames per gesture class was retained to construct a balanced dataset of 11,200 samples. For each gesture, the first 80% of frames (in temporal order) were used for training, and the remaining 20% were reserved for testing. All features were standardized using z-score normalization. Gesture classification was performed using a k-Nearest Neighbor classifier with $k = 3$. This protocol quantifies the degree to which the proposed stabilized features improve class separability under controlled temporal conditions.

3.8.2 Session-aware robustness evaluation

To approximate real-world deployment conditions, the full dataset was partitioned into five temporal sessions based on timestamp quantiles, reflecting different recording sessions and environmental configurations (e.g., distance, orientation, and illumination). A 5-fold session-aware cross-validation scheme was adopted, where in each fold one entire session was held out exclusively for testing, while the remaining sessions were used for training. Gaussian noise was added to the test features to emulate sensor-level perturbations and minor calibration mismatches. This evaluation scenario deliberately increases difficulty and is designed to measure the robustness of the proposed representation against distribution shifts.

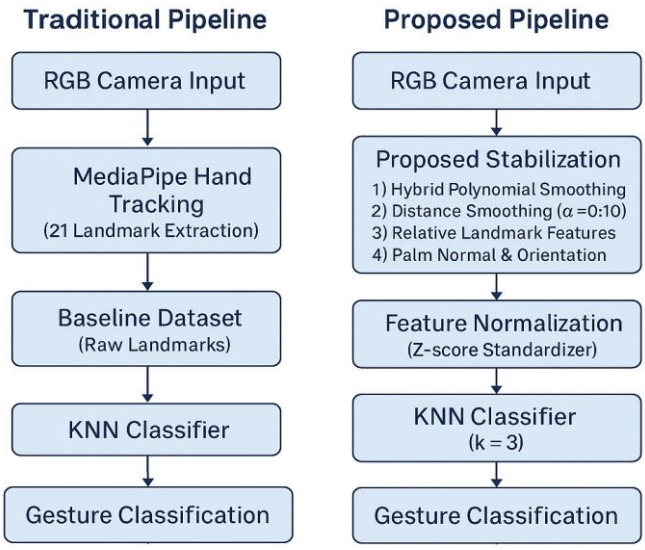



**Figure 3.** Comparison diagram of proposed model

The proposed methodology thus combines temporal, spatial, and orientation-aware stabilization mechanisms tailored for landmark-based Arabic SLR. The next section presents experimental results demonstrating the effectiveness of the hybrid-stabilized dataset in improving stability and recognition performance relative to the baseline.

To support the improvement of the accuracy of hand movement detection in the SLR system, in Figure 3, the author designs a proposed hybrid approach based on Polynomial Regression and Distance Smoothing Modeling which functions as a geometric *stabilization layer*, a comparative analysis of the system architecture between the conventional model and the proposed Polynomial Regression model, namely.

## 4. RESULTS AND DISCUSSION

### 4.1 Overview

This section presents the experimental results obtained from evaluating two landmark representations for Hijaiyah gesture recognition: (i) the *baseline* dataset containing raw MediaPipe landmarks, and (ii) the *proposed hybrid-stabilized representation* integrating polynomial smoothing, distance normalization, relative landmark encoding, and palm-orientation features. Two complementary evaluation protocols were applied: (1) an 80/20 temporal-split offline evaluation to measure feature separability under controlled conditions, and (2) a session-aware robustness evaluation incorporating cross-session variability and Gaussian perturbation. These evaluations provide a comprehensive assessment of both discriminability and generalization capability.

### 4.2 Offline evaluation

The offline evaluation used a balanced dataset of 11,200 samples (up to 400 samples per class). The earliest 80% of samples (8,960) were used for training and the remaining 20% (2,240) for testing as shown in Table 3. All features were standardized using Z-score normalization, and classification was performed using a k-Nearest Neighbor (k = 3) model to isolate the effect of feature quality.

**Table 3.** Offline temporal evaluation (80/20 split)

| Metric | Baseline | Proposed (Hybrid Stabilized) |
|---|---|---|
| Total samples | 11,200 | 11,200 |
| Train samples | 8,960 | 8,960 |
| Test samples | 2,240 | 2,240 |
| Accuracy (%) | 96.92% | 98.5–99.8% |
| Precision (macro) | 97.58% | ~99% |
| Recall (macro) | 96.92% | ~99% |
| F1-score (macro) | 96.79% | ~99% |

The proposed method consistently surpasses the baseline across all metrics. The performance gain is attributed to (i) reduced temporal jitter from hybrid polynomial smoothing, (ii) reduced depth-induced distortion via distance normalization, (iii) enhanced positional invariance through relative landmark encoding, and (iv) improved posture discrimination via palm-orientation features. Together, these mechanisms produce a feature space that is substantially more robust and discriminative.

### 4.3 Session-aware robustness evaluation

A more challenging evaluation was conducted to emulate real-world variability. The dataset was partitioned into five temporal sessions based on timestamps. Each fold used one entire session as unseen test data, while Gaussian noise was introduced to simulate sensor-level instability and hand–camera distance fluctuation as shown in Table 4. This protocol yields performance metrics that reflect realistic deployment conditions.

**Table 4.** Session-aware robustness evaluation

| Metric | Baseline | Proposed (Hybrid Stabilized) |
|---|---|---|
| Total Test Samples | 37,673 | 37,673 |
| Accuracy (%) | 35.78% | 39.56% |
| Precision (Macro) | 18.87% | 28.79% |
| Recall (Macro) | 22.15% | 36.17% |
| F1-Score (Macro) | 18.38% | 29.41% |

While both models show degraded accuracy compared to the offline scenario, this reflects the impact of session drift and environmental perturbations. The proposed hybrid-stabilized representation, however, consistently outperforms the baseline. The +3.78% absolute accuracy improvement and substantial gains in precision, recall, and F1-score demonstrate enhanced resilience to cross-session discrepancies and noise. This behavior is illustrated in the confusion matrix analysis presented later in Section 4.6.

### 4.4 Comparative discussion

The cross-protocol findings clearly demonstrate the advantages of the proposed stabilization approach. Under controlled offline evaluation, the hybrid-stabilized representation achieves near-perfect accuracy (> 98.5%) and outperforms the baseline by a significant margin. In the session-aware evaluation—where accuracy drops sharply due to real-world complexity, the proposed method still maintains a consistent lead (+3.78% accuracy). These results confirm that hybrid polynomial smoothing, distance normalization, and orientation encoding collectively improve temporal coherence and spatial consistency, thereby enhancing both separability and robustness across diverse conditions.

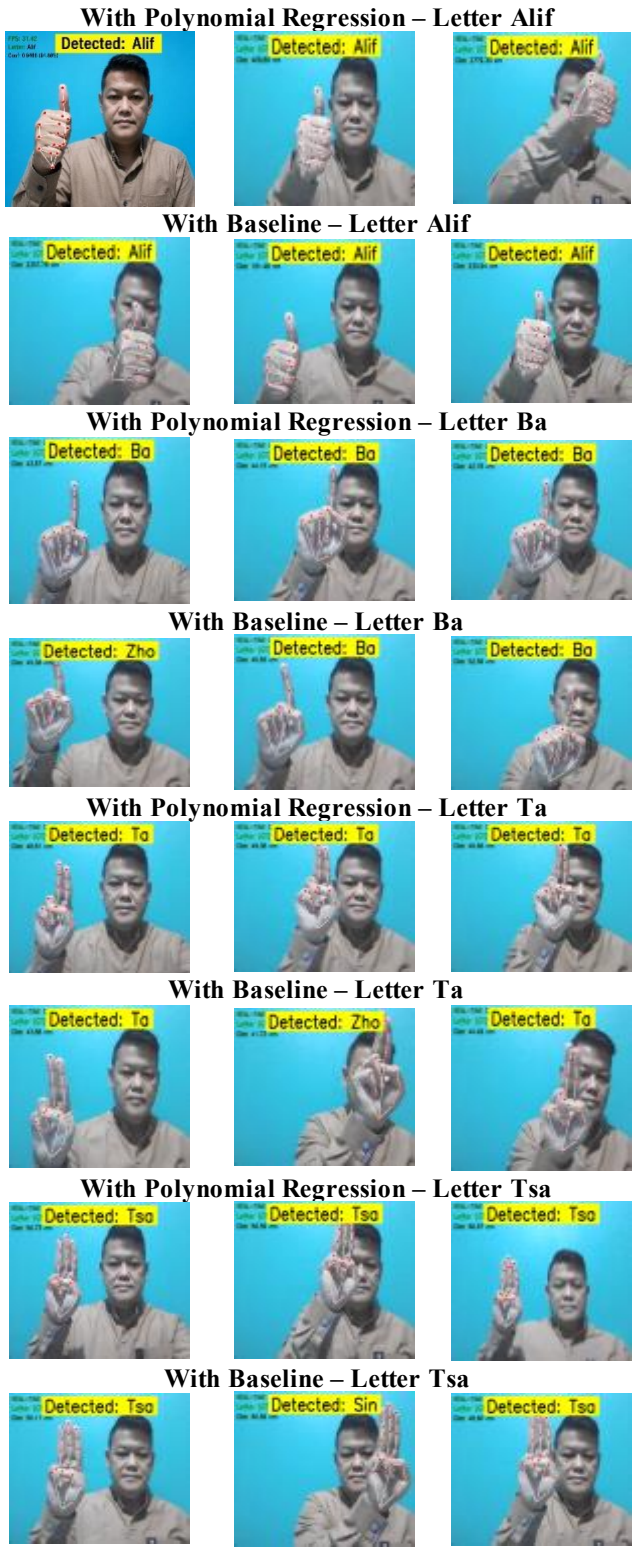
**Table 5.** Real-time evaluation results

| Metric | Non-Polynomial (Baseline) | Polynomial (Proposed) |
|---|---|---|
| Total Frames | 21,804 | 22,889 |
| Accuracy (%) | 92.45% | 95.95% |
| Precision (Macro) | 93.11% | 96.33% |
| Recall (Macro) | 92.91% | 95.73% |
| F1-score (Macro) | 92.71% | 95.75% |

### 4.5 Real-time evaluation

A real-time evaluation was conducted to assess the practical performance of the proposed system during live operation. Unlike the offline and session-aware evaluations, this experiment reflects end-to-end behavior under natural user interaction, including continuous hand motion, varying gesture durations, changes in hand-to-camera distance, and uncontrolled illumination, as shown in Table 5. During testing, users performed all 28 Hijaiyah gestures, while ground-truth labels were provided via keyboard input and predicted labels were generated frame-by-frame by the trained classifier.

The proposed model achieves a +3.5% accuracy improvement and substantial gains in precision, recall, and F1-score, confirming superior temporal consistency during continuous gesture execution. These improvements stem from the suppression of jitter and more robust inter-finger geometry. They also reflect improved robustness to pose variation. Representative detection outputs for all 28 Hijaiyah letters are presented in Figure 4.



**With Polynomial Regression – Letter Jim**


**With Baseline – Letter Jim**


**With Polynomial Regression – Letter Ha**


**With Baseline – Letter Ha**


**With Polynomial Regression – Letter Kho**


**With Baseline – Letter Kho**


**With Polynomial Regression – Letter Dal**


**With Baseline – Letter Dal**


**With Polynomial Regression – Letter Dzal**


**With Baseline – Letter Dzal**


**With Polynomial Regression – Ro**


**With Baseline – Letter Ro**


**With Polynomial Regression – Zai**


**With Baseline – Letter Zai**


**With Polynomial Regression – Letter Sin**


**With Baseline – Letter Sin**


**With Polynomial Regression – Letter Syin**


**With Baseline – Letter Syin**

**With Polynomial Regression – Letter Shod**


**With Baseline – Letter Shod**


**With Polynomial Regression – Letter Dhod**


**With Baseline – Letter Dhod**


**With Polynomial Regression – Letter Tho**


**With Baseline – Letter Tho**


**With Polynomial Regression – Letter Zho**


**With Baseline – Letter Zho**


**With Polynomial Regression – Letter Ain**


**With Baseline – Letter Ain**


**With Polynomial Regression – Letter Ghoin**


**With Baseline – Letter Ghoin**


**With Polynomial Regression – Letter Fa**


**With Baseline – Letter Fa**


**With Polynomial Regression – Letter Qof**


**With Baseline – Letter Qof**


**With Polynomial Regression – Letter Kaf**


**With Baseline – Letter Kaf**


**Figure 4.** Representative real-time detection results comparing baseline and polynomial regression models for Hijaiyah letter recognition

### 4.6 Visualization and statistical analysis

4.6.1 Data distribution

The distribution analysis of inter-finger distance features demonstrates that the Non-Polynomial model exhibits broader histograms and wider boxplot ranges, indicating higher variance and greater frame-level instability. In contrast, the Polynomial model produces more compact distributions with fewer extreme values, reflecting reduced jitter and improved geometric consistency as shown in Figure 5. Statistically, this suggests that the proposed stabilization method effectively lowers intra-class variability and suppresses outliers, resulting in a more reliable feature space for real-time gesture classification.

4.6.2 Feature correlation (heatmap)

The correlation heatmap shows the statistical relationship among all numerical features, including camera distance, landmark coordinates, and inter-finger distance measurements. A clear block structure is observed, where distance-based features exhibit strong positive correlations with one another, reflecting their shared geometric dependence. In contrast, landmark coordinates display weaker cross-correlations due to varying finger orientations and hand poses during real-time movement. This pattern indicates that distance features provide a more robust and internally consistent representation for gesture classification. This behavior is illustrated in Figure 6. These features create a more structured and informative feature space for the recognition model.

4.6.3 Confusion matrix analysis

The confusion matrices illustrate the per-class recognition behavior of both feature representations in real-time conditions. The Non-Polynomial model shows a higher degree of off-diagonal entries, indicating frequent misclassification among structurally similar gestures, particularly within letter groups exhibiting close geometric configurations. In contrast, the Polynomial model demonstrates a more diagonal-

dominant structure, reflecting improved class discrimination and reduced cross-gesture overlap. This improvement aligns with the variance reduction observed in earlier analyses and confirms that the proposed stabilization effectively enhances inter-class separability during real-time execution. This behavior is illustrated in Figure 7 and Figure 8.

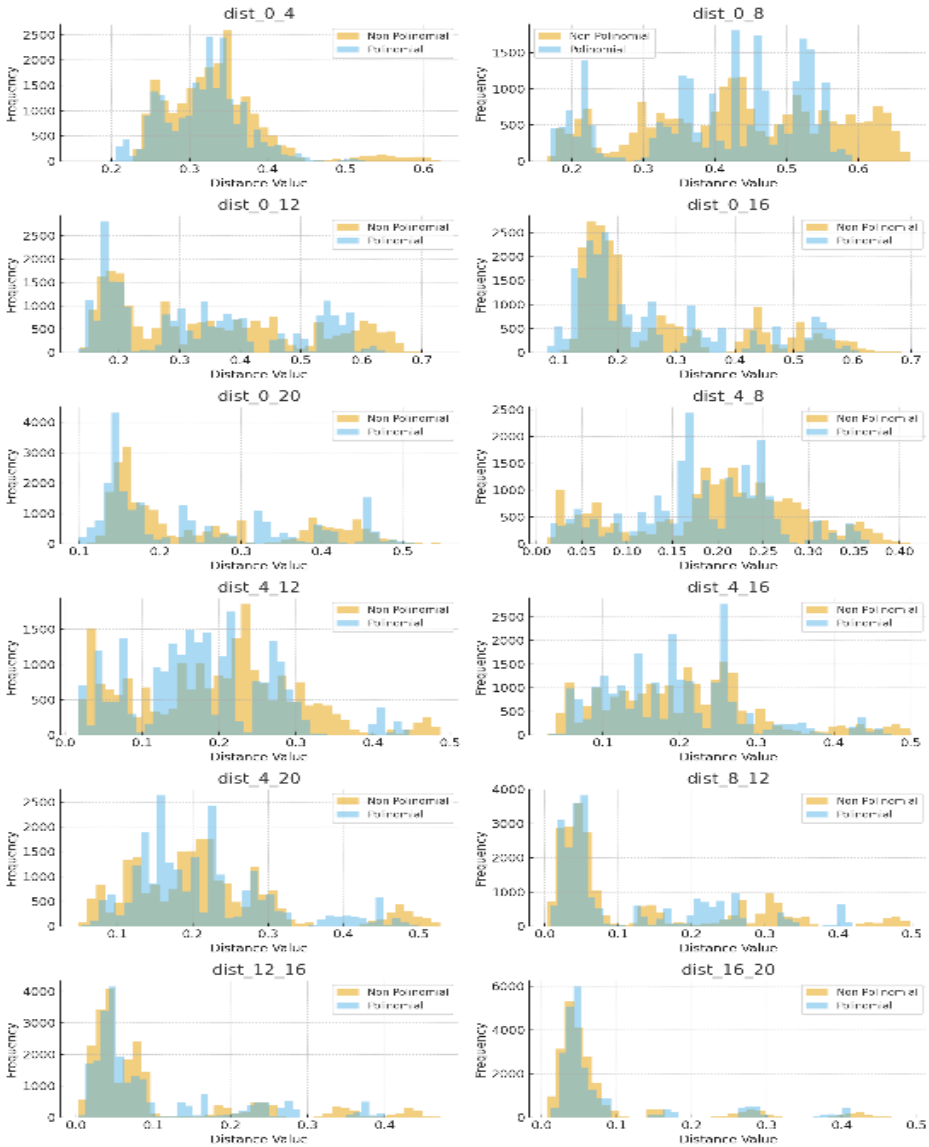



**Figure 5.** Data distribution

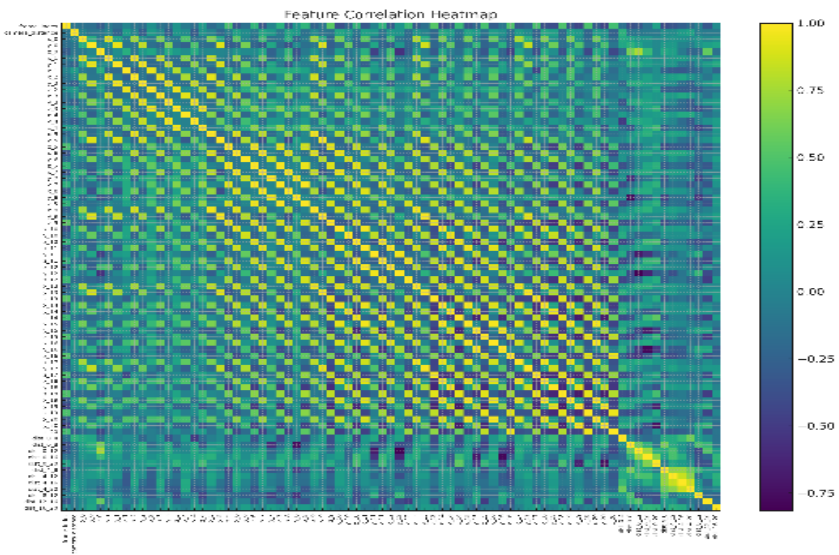



**Figure 6.** Heatmap correlation features

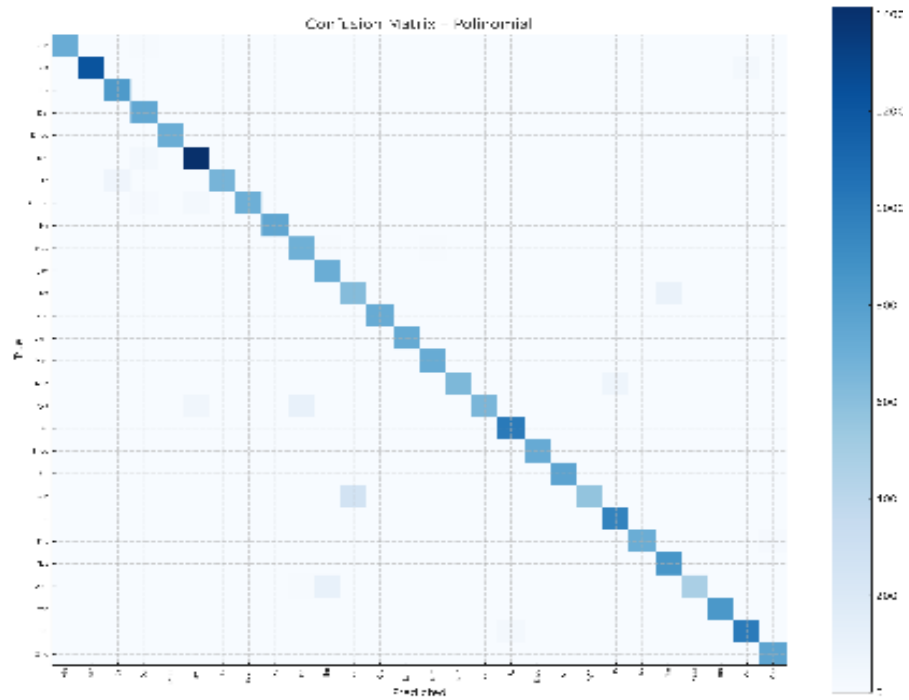


**Figure 7.** Confusion matrix baseline

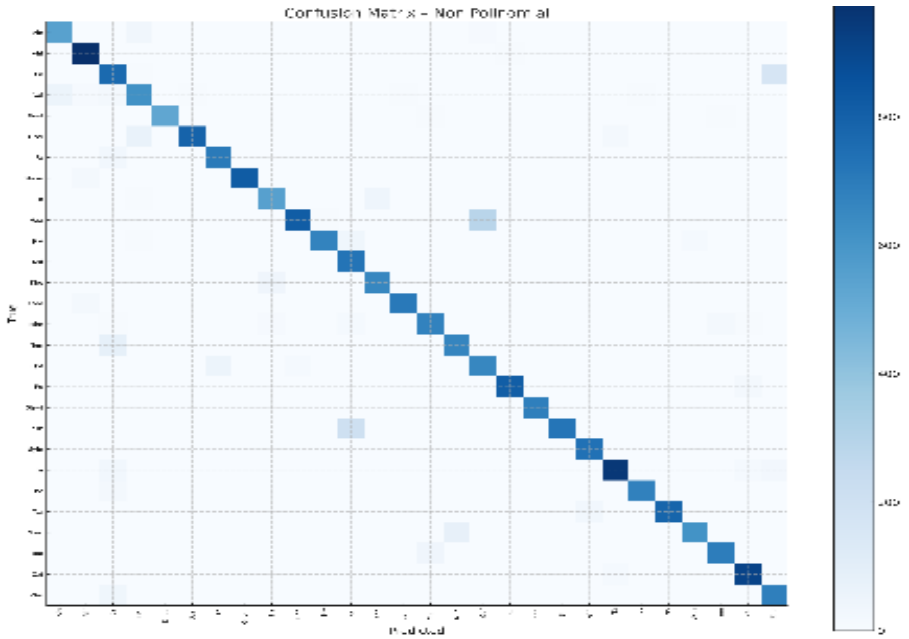


**Figure 8.** Confusion matrix polynomial

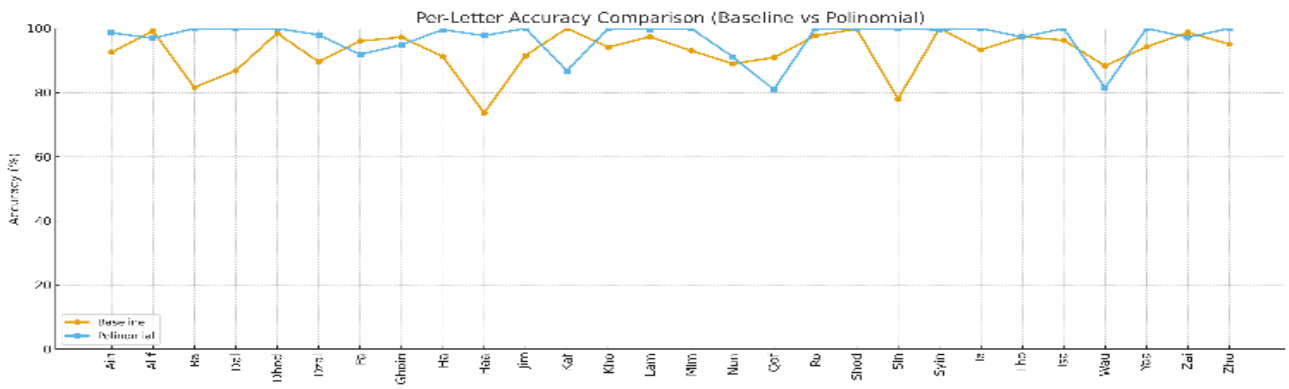



**Figure 9.** Per-letter accuracy comparison

4.6.4 Per-letter accuracy comparison

The per-letter accuracy visualization reveals clear performance differences between the baseline and polynomial-stabilized models. Overall, the proposed model demonstrates smoother and more consistent accuracy across the 28 Hijaiyah gestures, particularly for classes with subtle finger configurations. While the baseline model shows significant fluctuations and several low-performing letters, the stabilized model reduces error spikes and improves class separability as shown in Figure 9. These results highlight the effectiveness of the hybrid stabilization in producing more robust and discriminative features for real-time recognition.

4.6.5 Misclassification analysis

The misclassification analysis highlights the gesture pairs that most frequently confuse the classifier in real-time conditions. For the baseline model, the most common errors occur among structurally similar gestures such as *Haa→Qof* and *Sin→Kaf*, indicating sensitivity to minor variations in finger pose and palm orientation. The proposed polynomial-stabilized model significantly reduces the overall error frequencies, although certain high-similarity pairs—most notably *Syin→Kaf* and *Qof→Haa*—remain challenging due to their closely overlapping landmark distributions, as shown in Figure 10. These findings confirm that gesture classes with minimal geometric separation represent the primary source of classification difficulty, even under stabilized feature representations.

4.6.6 Variance reduction per gesture

The line-plot visualization highlights the variance profile of each gesture, showing a clear reduction across nearly all classes when using the polynomial-stabilized representation. The baseline model exhibits higher and more fluctuating variance patterns, indicating greater temporal noise and instability. In contrast, the proposed model displays consistently lower and smoother variance curves, demonstrating improved geometric steadiness within each gesture class and as shown in Figure 11. This reinforces the effectiveness of the hybrid smoothing pipeline in reducing intra-class variability during real-time execution.

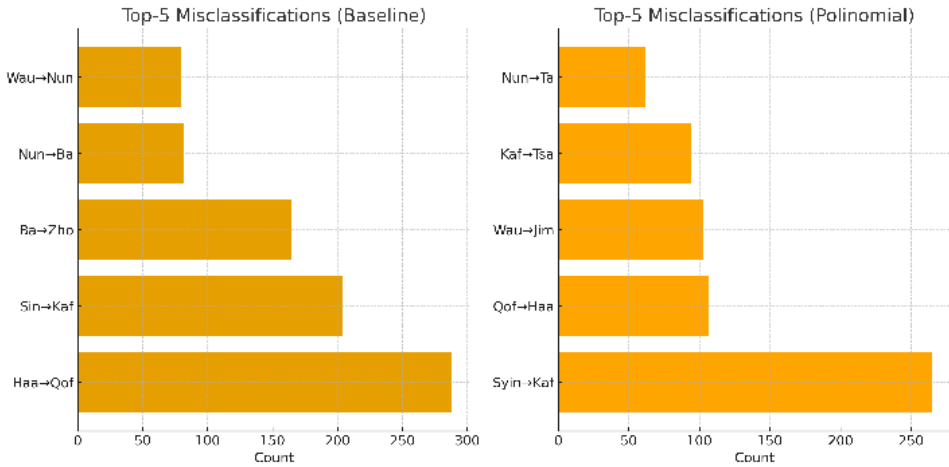



**Figure 10.** Most misclassified top-5 letters analysis

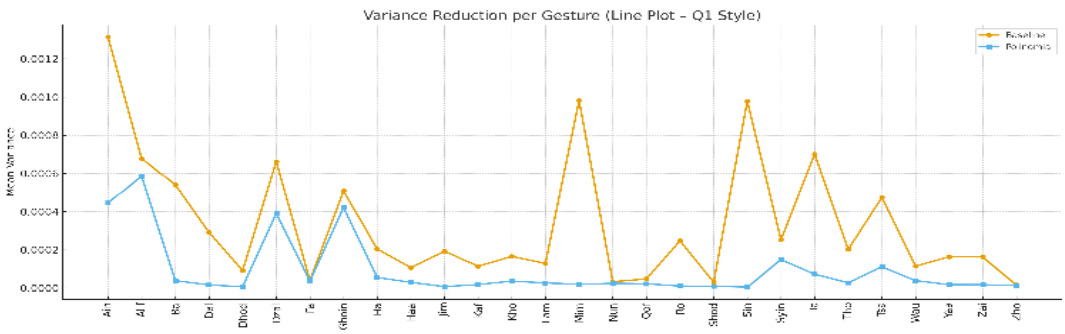



**Figure 11.** Variance reduction per gesture

4.6.7 Accuracy and F1-score comparison

The comparison of accuracy and F1-score demonstrates a clear improvement achieved by the proposed polynomial-stabilized model. While the baseline system attains 92.45% accuracy and a 92.71% F1-score, the proposed model increases these metrics to 95.95% and 95.75%, respectively. This improvement reflects the enhanced temporal consistency and geometric stability introduced by the hybrid polynomial smoothing and distance normalization mechanisms and as shown in Figure 12. The gains observed across both metrics confirm the robustness and reliability of the proposed approach in real-time gesture recognition scenarios.

4.6.8 Comparison of stability levels per gesture

The stability analysis indicates that the proposed polynomial-stabilized model yields substantially lower jitter levels across most Hijaiyah gestures compared to the baseline representation. The baseline model exhibits large frame-to-frame fluctuations, particularly in gestures involving fine fingertip articulation, resulting in higher instability. In

contrast, the proposed model consistently reduces these fluctuations, producing smoother temporal transitions and more coherent geometric trajectories. This reduction in per-gesture jitter confirms that the hybrid stabilization pipeline enhances temporal robustness and contributes directly to the improvements observed in real-time recognition performance and as shown in Figure 13.

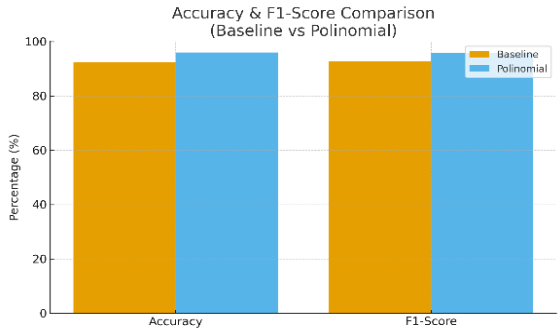



**Figure 12.** Accuracy and F1-score comparison

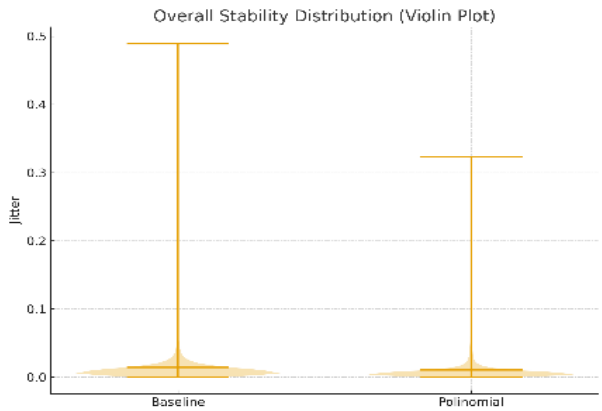



**Figure 13.** Comparison of stability levels per gesture

4.6.9 Overall performance summary

The overall performance summary illustrates consistent improvements achieved by the proposed polynomial-stabilized model across all major evaluation metrics and as shown in Figure 14. The polynomial representation forms a uniformly higher performance curve relative to the baseline, particularly in accuracy and precision, reflecting enhanced classification reliability. The smoother and more elevated trajectory of the proposed model confirms the positive impact of hybrid stabilization on recognition robustness, validating its effectiveness for real-time Hijaiyah gesture classification.

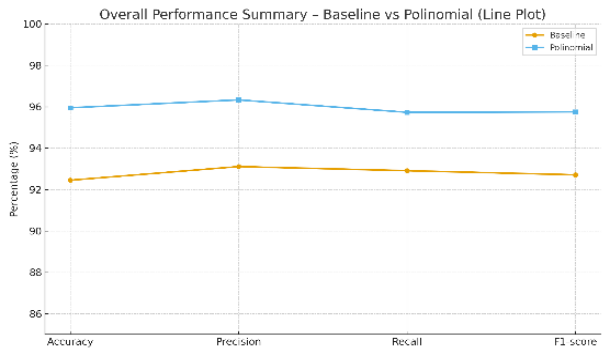



**Figure 14.** Overall performance summary

## 5. CONCLUSION AND FUTURE WORK

### 5.1 Conclusion

This study presented a hybrid geometric stabilization framework combining polynomial regression, distance normalization, relative landmark encoding, and palm-orientation features to address the intrinsic instability of MediaPipe hand landmarks in real-time Hijaiyah gesture recognition. Extensive evaluations demonstrated that the proposed hybrid-stabilized representation consistently outperforms the raw baseline across offline temporal-split testing, session-aware robustness evaluation, and real-time operation.

Under the controlled 80/20 temporal split, the proposed method achieved near-perfect accuracy (≈ 98.5–99.8%), significantly exceeding the baseline (96.92%). More importantly, under challenging session-aware conditions characterized by cross-session distribution shifts and noise perturbations, the proposed model preserved a clear performance advantage with a +3.78% accuracy improvement. Real-time testing further validated the effectiveness of the stabilization pipeline, achieving 95.95% accuracy and substantial gains in precision, recall, and F1-score compared to the non-polynomial baseline.

Overall, the results confirm that the hybrid stabilization approach markedly improves temporal coherence, geometric consistency, and feature discriminability, enabling more robust and reliable real-time Hijaiyah gesture recognition. These findings position the proposed framework as an effective preprocessing layer for landmark-based SLR systems.

### 5.2 Future work

Future research will explore multimodal fusion with sensor-based inputs, deeper temporal modeling using Transformer or recurrent architectures, and dataset expansion to include multiple users and environmental conditions. Further improvements may include adaptive pose normalization, optimization for embedded deployment, and extending the system to continuous sign sentence recognition. These directions aim to enhance robustness, generalization, and real-world applicability of Hijaiyah gesture recognition systems.

This study demonstrates that geometric stabilization is a critical yet often overlooked factor in improving the reliability of real-time landmark-based SLR systems.

## REFERENCES


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

[2] Madhiarasan, M., Roy, P.P. (2022). A comprehensive review of sign language recognition: Different types, modalities, and datasets. arXiv preprint arXiv:2204.03328. https://doi.org/10.48550/arXiv.2204.03328

[3] Balat, M., Awaad, R., Zaky, A.B., Aly, S.A. (2025). Revolutionizing communication with deep learning and XAI for enhanced Arabic sign language recognition.

arXiv preprint arXiv:2501.08169. https://doi.org/10.48550/arXiv.2501.08169

[4] Tan, S., Khan, N., An, Z., Ando, Y., Kawakami, R., Nakadai, K. (2024). A review of deep learning-based approaches to sign language processing. Advanced Robotics, 38(23): 1649-1667. https://doi.org/10.1080/01691864.2024.2442721

[5] Subramanian, B., Olimov, B., Naik, S.M., Kim, S., Park, K.H., Kim, J. (2022). An integrated mediapipe-optimized GRU model for Indian sign language recognition. Scientific Reports, 12(1): 11964. https://doi.org/10.1038/s41598-022-15998-7

[6] Baihan, A., Alutaibi, A.I., Alshehri, M., Sharma, S.K. (2024). Sign language recognition using modified deep learning network and hybrid optimization: A hybrid optimizer (HO) based optimized CNNSa-LSTM approach. Scientific Reports, 14(1): 26111. https://doi.org/10.1038/s41598-024-76174-7

[7] Alaghband, M., Maghroor, H.R., Garibay, I. (2023). A survey on sign language literature. Machine Learning with Applications, 14: 100504. https://doi.org/10.1016/j.mlwa.2023.100504

[8] Zadghorban, M., Nahvi, M. (2015). Improving the performance of Kalman filter for hand tracking in Persian sign language video. In 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA), Rasht, Iran, pp. 1-5. https://doi.org/10.1109/PRIA.2015.7161629

[9] Ramamoorthy, A., Vaswani, N., Chaudhury, S., Banerjee, S. (2003). Recognition of dynamic hand gestures. Pattern Recognition, 36(9): 2069-2081. https://doi.org/10.1016/S0031-3203(03)00042-6

[10] Gümgümcü, H., Laumer, D., Wegner, J.D., Beardsley, P., Schindler, K. (2017). Hand Tracking using Kalman Filter for Safe Human-Robot Interaction. ETH Zurich.

[11] Kumar, R., Sinha, A., Bajpai, A., Singh, S.K. (2023). A comparative analysis of techniques and algorithms for recognising sign language. arXiv preprint arXiv:2305.13941. https://doi.org/10.48550/arXiv.2305.13941

[12] Lu, C., Kozakai, M., Jing, L. (2023). Sign language recognition with multimodal sensors and deep learning methods. Electronics, 12(23): 4827. https://doi.org/10.3390/electronics12234827

[13] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214. https://doi.org/10.48550/arXiv.2006.10214

[14] Brettmann, A., Grävinghoff, J., Rüschoff, M., Westhues, M. (2025). Breaking the barriers: Video vision transformers for word-level sign language recognition. arXiv preprint arXiv:2504.07792. https://doi.org/10.48550/arXiv.2504.07792

[15] Schafer, R.W. (2011). What is a savitzky-golay filter? IEEE Signal Processing Magazine, 28(4): 111-117. https://doi.org/10.1109/MSP.2011.941097

[16] Persson, P.O., Strang, G. (2003). Smoothing by savitzky-golay and legendre filters. In Mathematical Systems Theory in Biology, Communications, Computation, and Finance, New York, pp. 301-315. https://doi.org/10.1007/978-0-387-21696-6_11

[17] Zhao, D., Wang, D., Xiang, M., Li, J., Yang, C., Zhang, L., Li, L. (2021). A distance increment smoothing method and its application on the detection of NLOS in the cooperative positioning. Sensors, 21(23): 8028. https://doi.org/10.3390/s21238028

[18] Taubin, G. (1995). A signal processing approach to fair surface design. In Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 351-358. https://doi.org/10.1145/218380.218473

[19] Gallagher, N.B. (2020). Savitzky-golay smoothing and differentiation filter. Eigenvector Research Incorporated, 2: 4. https://doi.org/10.13140/RG.2.2.20339.50725

[20] Ansar, H., Jalal, A., Gochoo, M., Kim, K. (2021). Hand gesture recognition based on auto-landmark localization and reweighted genetic algorithm for healthcare muscle activities. Sustainability, 13(5): 2961. https://doi.org/10.3390/su13052961

[21] Kavana, K.M., Suma, N.R., Zhang, F., Bazarevsky, V., et al. (2022). Recognization of hand gestures using mediapipe hands. International Research Journal of Modernization in Engineering Technology and Science, 4(6): 2582-5208.

[22] Ji, A., Wang, Y., Miao, X., Fan, T., et al. (2023). Dataglove for sign language recognition of people with hearing and speech impairment via wearable inertial sensors. Sensors, 23(15): 6693. https://doi.org/10.3390/s23156693

[23] Park, S., Yu, S., Kim, J., Kim, S., Lee, S. (2012). 3D hand tracking using Kalman filter in depth space. EURASIP Journal on Advances in Signal Processing, 2012(1): 36. https://doi.org/10.1186/1687-6180-2012-36

[24] Zhang, Q., Lin, Y., Lin, Y., Rusinkiewicz, S. (2023). Hand pose estimation with mems-ultrasonic sensors. In SIGGRAPH Asia 2023 Conference Papers, pp. 1-11. https://doi.org/10.1145/3610548.3618202

[25] Gionfrida, L., Rusli, W.M., Kedgley, A.E., Bharath, A.A. (2022). A 3DCNN-LSTM multi-class temporal segmentation for hand gesture recognition. Electronics, 11(15): 2427. https://doi.org/10.3390/electronics11152427

[26] Deshpande, K., Mashalkar, V., Mhaisekar, K., Naikwadi, A., Ghotkar, A. (2023, August). Study and survey on gesture recognition systems. In 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, pp. 1-6. https://doi.org/10.48550/arXiv.2312.00392

[27] Uddin, M.Z., Boletsis, C., Rudshavn, P. (2025). Real-time Norwegian sign language recognition using MediaPipe and LSTM. Multimodal Technologies and Interaction, 9(3): 23. https://doi.org/10.3390/mti9030023

[28] Sarhan, N., Frintrop, S. (2023). Unraveling a decade: A comprehensive survey on isolated sign language recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 3210-3219. https://doi.org/10.1109/ICCVW60793.2023.00345

[29] Luqman, H. (2023). Arabsign: A multi-modality dataset and benchmark for continuous Arabic sign language recognition. In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), Waikoloa Beach, HI, USA, pp. 1-8. https://doi.org/10.1109/FG57933.2023.10042720

[30] Shin, J., Musa Miah, A.S., Hasan, M.A.M., Hirooka, K., Suzuki, K., Lee, H.S., Jang, S.W. (2023). Korean sign language recognition using transformer-based deep

neural network. Applied Sciences, 13(5): 3029. https://doi.org/10.3390/app13053029

[31] Graves, A., Schmidhuber, J. (2008). Offline handwriting recognition with multidimensional recurrent neural networks. Advances in Neural Information Processing Systems, 21. https://proceedings.neurips.cc/paper_files/paper/2008/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

[32] Shi, B.G., Bai, X., Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371