Computer Modeling in
Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.068726

ARTICLE

Fusing Geometric and Temporal Deep Features for High-Precision Arabic Sign Language Recognition

Yazeed Alkhrijah[1,2], Shehzad Khalid[3], Syed Muhammad Usman[4,*], Amina Jameel[3] and Danish Hamid[5]

[1]King Salman Center for Disability Research (KSCDR), Riyadh, 11614, Saudi Arabia
[2]Department of Electrical Engineering, Imam Mohammad ibn Saud Islamic University (IMSIU), Riyadh, 11623, Saudi Arabia
[3]Department of Computer Engineering, Bahria University, Islamabad, 44000, Pakistan
[4]Department of Computer Science, Bahria University, Islamabad, 44000, Pakistan
[5]Department of Creative Technologies, Air University, Islamabad, 44000, Pakistan

*Corresponding Author: Syed Muhammad Usman. Email: smusman.h11@bahria.edu.pk

Received: 05 June 2025; Accepted: 04 July 2025; Published: 31 July 2025

ABSTRACT: Arabic Sign Language (ArSL) recognition plays a vital role in enhancing the communication for the Deaf and Hard of Hearing (DHH) community. Researchers have proposed multiple methods for automated recognition of ArSL; however, these methods face multiple challenges that include high gesture variability, occlusions, limited signer diversity, and the scarcity of large annotated datasets. Existing methods, often relying solely on either skeletal data or video-based features, struggle with generalization and robustness, especially in dynamic and real-world conditions. This paper proposes a novel multimodal ensemble classification framework that integrates geometric features derived from 3D skeletal joint distances and angles with temporal features extracted from RGB videos using the Inflated 3D ConvNet (I3D). By fusing these complementary modalities at the feature level and applying a majority-voting ensemble of XGBoost, Random Forest, and Support Vector Machine classifiers, the framework robustly captures both spatial configurations and motion dynamics of sign gestures. Feature selection using the Pearson Correlation Coefficient further enhances efficiency by reducing redundancy. Extensive experiments on the ArabSign dataset, which includes RGB videos and corresponding skeletal data, demonstrate that the proposed approach significantly outperforms state-of-the-art methods, achieving an average F1-score of 97% using a majority-voting ensemble of XGBoost, Random Forest, and SVM classifiers, and improving recognition accuracy by more than 7% over previous best methods. This work not only advances the technical state-of-the-art in ArSL recognition but also provides a scalable, real-time solution for practical deployment in educational, social, and assistive communication technologies. Even though this study is about Arabic Sign Language, the framework proposed here can be extended to different sign languages, creating possibilities for potentially worldwide applicability in sign language recognition tasks.

KEYWORDS: Arabic sign language recognition; multimodal feature fusion; ensemble classification; skeletal data; inflated 3D ConvNet (I3D)

1 Introduction

Effectively communicating through sign language interpretation is extremely vital for the hearing-impaired. It adds an addictive layer to interactions, lowers social walls, helps inclusion, and paramount accessibility. Despite promising advancements in assistive technologies, Arabic Sign Language (ArSL) brings challenges [1]. Variations in signing style, quick hand motions, and different orientations of hands also make recognition unreliable. The presence of occlusions like overlapping hands also makes detection more difficult [2]. Moreover, accuracy can be affected by changing lighting conditions [3]. There are two challenges pertaining to the recognition of Arabic Sign Language that can be seen within the plots in Fig. 1. Intra-class variability, wherein the same sign may vary across performers due to differences in hand shape, articulation speed, or posture, is the first challenge. The second challenge is inter-class similarity, where two different signs do something similar with respect to hand orientation or motion path. Both majorly affect the performance of the model. All this signifies the importance of integrating structural (skeletal) and motion (temporal) features to strengthen the recognition robustness. The key challenge is the small number of large, annotated datasets, which limits the training of generalizable models. Our method addresses the problem of the scarcity of large-scale annotated datasets, using a hybrid approach that promotes feature extraction and generalization. Our method leverages skeletal pose dynamics and deep-learned temporal features to model motion patterns, requiring fewer annotations.

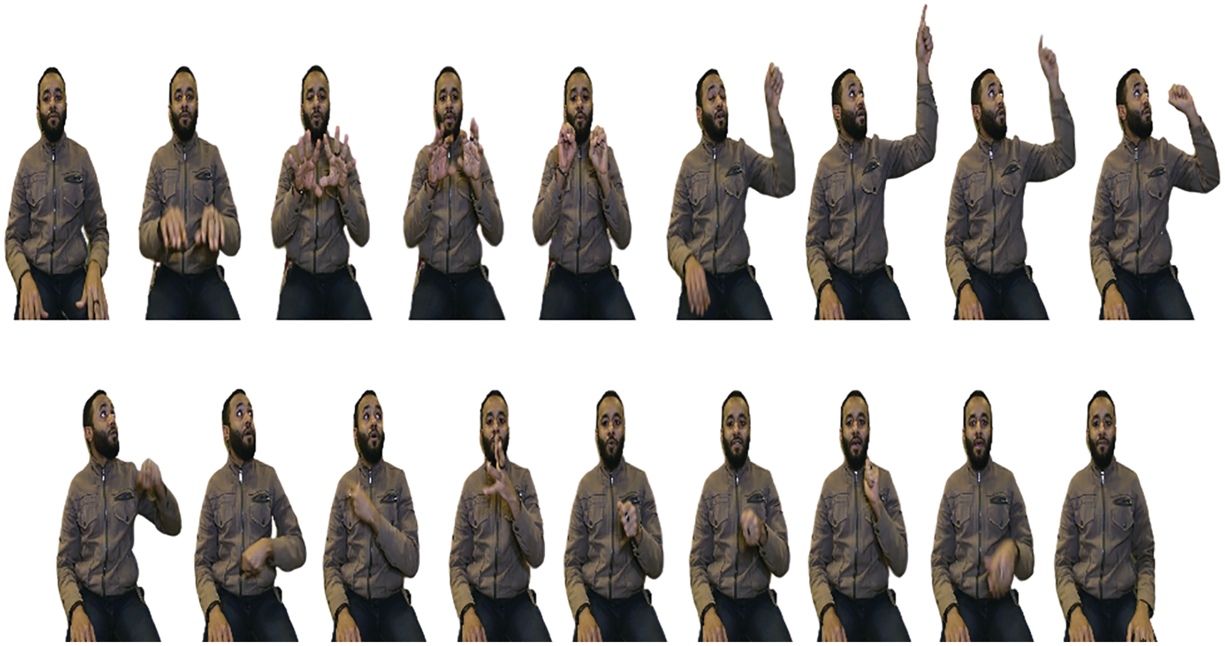


Figure 1: Gesture variability examples in Arabic Sign Language. The figure illustrates within class variability when the same sign may look different due to inter signer styles, articulation, or hand orientations, and inter class similarity, where dissimilar signs may look the same in appearance. Such types of variations are very challenging for reliable and robust sign language recognition systems [4]

One popular approach to gesture recognition is to use skeleton-based analysis, which involves tracking the movements of 3D joints within the human body. Although this approach successfully captures the spatial configuration of human motion, real-world scenarios are still plagued by issues like variability in gestures, occlusions, and environmental noise [5,6]. Numerous approaches have been developed to tackle these challenges [7,8]. The early approaches consisted of using traditional machine learning algorithms such as k-nearest neighbors (KNN) [9], support vector machines (SVM) [10], and decision trees to classify skeleton data [11]. These algorithms are interpretable and computationally efficient; however, they are limited by the variability of real-world gestures and tend to overfit in dynamic environments. The introduction of deep learning techniques led to a paradigm shift in skeleton-based action recognition, where deep models such as convolutional neural networks (CNNs) [12] and recurrent neural networks (RNNs) [13], have demonstrated the ability to effectively capture the intricate spatial and temporal dependencies embedded within skeleton data. Although they provide high accuracy, deep learning models need larger labeled datasets and have substantial computational resource requirements. This may restrict their use in cases where recognitions must be performed in real time or on smaller datasets [14,15].

Arabic Sign Language (ArSL) acts as the main visual means of communication for the Deaf community in Arabic-speaking countries. Unlike the sign languages of America, ArSL shows heavy regional variation, is not standardized, and is mainly influenced by the dialects of spoken Arabic. Manual and non-manual signs constitute the ArSL, and the signs are more active, especially in situations of conversation when gestures might often run uninterrupted for a while without being segmented. The dynamicity of the structure, but limited availability of sufficiently large annotated datasets, has made the recognition of ArSL even harder than that of some of its counterparts. The ArabSign dataset used in this work contains 50 signs commonly used at the sentence level, which reflect practical scenarios of usage in real-world communication.

In this work, an ensemble method is proposed that takes advantage of both the benefits of deep learning feature representations and geometric features. These complementary representations are combined in order to make the system stronger against real world challenges such as gesture variability, occlusions, and harsh environmental conditions. In Section 5, tests of this robustness are exemplified by the high F1-scores across distinct and often overlapping gesture classes in the ArabSign dataset, despite the above hardships, and relevant analysis. This framework unifies skeleton-based geometric features with deep learning-based temporal features, which were extracted using the Inflated 3D ConvNet (I3D) model. In our technique, after extracting the body joints in each frame, we compute pairwise distances and angles to all other body joints from frame to frame, thus capturing the spatial relationships across frames. Meanwhile, the temporal features of the video are extracted using I3D from another viewpoint. The temporal features are then fused with the skeleton-based geometric features at the feature level [16]. This fusion preserves the pertinent spatial-temporal characteristics in the final feature representation [17,18]. The fused feature vector builds a comprehensive descriptor of gestures that surmounts the hindrance posed by gesture variability and noise [19]. Combining the geometric simplicity of the skeleton features with I3D feature temporal richness provides a powerful framework for gesture recognition. The fine motion details and high-level spatiotemporal patterns concerning our gesture constant will thus be captured for better recognition. This combination strengthens our method against variations in gestures, occlusions, and environmental noises. The approach is therefore intended to perform well in different real-life situations, allowing it to soar from research-based approaches into practical applications. The key contributions of this research are highlighted in the following section.

- A new hybrid framework combining skeletal-based geometrical features, such as pairwise distances and joint angles, with deep spatiotemporal features using I3D is presented at the feature level.
- We present an efficient fusion pipeline that implements feature selection using the Pearson Correlation Coefficient. This helps by reducing dimensionality by keeping the maximal discriminative power.
- Lightweight classifiers (XGBoost, RF, SVM) are chosen for use as ensemble classifiers with majority voting to balance interpretability and robustness.
- We carry out a comprehensive ablation and comparative analysis of the ArabSign dataset, in which we sufficiently outperform other methods (F1-score improvement > 7%).
- The model is applicable for inference deployment into edge assistive devices since it works in real time and is also simple and modular.

In this paper, we propose a practical yet suitable hybrid recognition model for ArSL, which incorporates classical geometric modelling and deep temporal features into a unified, interpretable, and portable system. It is not our intention to develop a new neural network; instead, we plan to propose a low-latency, feature-centric alternative for use in real-world accessibility applications.

2 Related Work

SLR has been researched extensively in different languages, such as ASL, CSL, and DGS, where visual and skeleton data have been used to model gestures [20–22]. The skeleton-based approaches in SLR have described the competency of modelling spatiotemporal dynamics of joint movements using graph convolutions such as ST-GCN and Pose-GCN. On the visual end, I3D, 3D CNNs, and CNN-RNN hybrids have been extensively utilized to extract rich motion features from the RGB or optical flow streams [23,24]. Multimodal fusion approaches have also gained prominence in more recent investigations. For example, references [25] and [26] studied fusing visual and linguistic cues for continuous sign translation. In the ASL domain, CNN-LSTM and transformer-based architectures have been used to model sign language on a sentence level. The earlier works testify that combining multiple modalities would give better performance for recognition. Earlier work on ArSL included finger-spelling recognition with feature vector extraction using ANFIS and reaching 93.55% accuracy [27], and polynomial classifiers being more accurate than the ANFIS on the same dataset, but suffered from limited accuracy owing to the fact that the dataset was not uniformly divided [28].

An enhancement of an Arabic Sign Language (ArSL) recognizer with the addition of a voice translator is proposed to connect users and non-users of sign language by Hemayed and Hassanien [29]. With the help of the Prewitt edge detector and PCA algorithm, a recognition accuracy of 97% was achieved. Misrecognition happened in extreme lighting conditions. In their research, the authors also tracked hand gesture movements from the detected faces and did this tracking with hidden Markov models (HMM) with an accuracy greater than 95%. Also, static Arabic sign alphabet recognition was performed using one-vs.-all Support Vector Machines (SVM) with a histogram of oriented gradients (HOG) descriptors. Nevertheless, static recognition discovered a gap in non-sent handling, targeting practical uses of sign alphabets while communicating. In natural dialogues, all signs are dynamic by default [30]. To overcome this limitation Elons et al. [31] investigated using the Leap Motion sensor to enable dynamic recognition of Arabic sign language, given that the sensor provides high-fidelity trajectories of hand movements. Their method provided enhanced awareness using depth-based tracking which greatly improved accuracy compared to classical image techniques, however, their system was limited by the sensor field of view and sensitivity to environmental factors. However, static recognition of the Arabic sign alphabet revealed a gap in the research, as it did not focus on the practical application of the sign alphabet in day to day life. In real-world conversations, most signs are inherently dynamic.

In general, models for multi-model hand signal recognition are still constrained. Such constraints include differences in training data, sensitivity to illumination, and poor discrimination in kinetic signs, as summarized in Table 1. Computer vision has seen significant advancements in recent years, with applications spanning from sign language recognition to sports analysis, such as recognizing basketball referee signals in online videos. Žemgulys et al. [32] propose methods to recognize and interpret the gestures of hand signals of basketball referees during live games, addressing challenges such as varying lighting conditions and diverse hand orientations to mitigate overfitting and reduce the risk of gesture misrecognition. This was achieved based on a combination of two image segmentation methods along with attributes of the local binary pattern (LBP) and HOG. Using a combination of LBP features and SVM data for identification. This method recorded 95.6% as its accuracy rate. Vaitkevičius et al. [33] report that participants used a Leap Motion device to track their hand and finger movements during gesture execution. Gesture recognition was then performed using the hidden Markov classification (HMC) algorithm. The coupling of data acquired from the algorithm and the Leap Motion device depicted the efficiency of the system in gesture recognition. Within the document, a gesture identification sub-system containing the modules of motion detection, recognition of gestures, and data harvesting was presented. The paper reported the performance in terms of words per minute (WPM) as a metric and measured the error rates using minimum string distance (MSD).

Table 1: summary of key Arabic Sign Language (ArSL) recognition methods including datasets, performance metrics (generally accuracy unless specified otherwise), and associated limitations. Limitations are based on either explicit statements given in original studies or critical observations made during our comparative analysis. Most previous literature compares the studies based on accuracy; however, this study primarily reports the performance through F1-score, as discussed in Section 5

Method	Dataset	Performance	Limitations
Prewitt Edge Detector, PCA, HMM	Custom	97% Accuracy (PCA), >95% (HMM)	Sensitive to lighting; limited real-world applicability
Leap Motion Sensor, Depth Tracking	Leap Motion-based dataset	Qualitative improvement	Limited field of view; environment-sensitive
Pose-Based Transformer	KArSL-100	99.74% Accuracy (SD), 68.2% (SI)	Sharp drop in signer-independent mode
Deep CNN + Transfer Learning	Mixed datasets (40, 23, 10 classes)	98.12%, 100%, 76.67% (SD); 84.38%, 34.9%, 70% (SI)	Poor generalization; data scarcity
2DCRNN, 3DCNN	224 videos, 5 individuals	92% (2DCRNN), 99% (3DCNN)	Small dataset size; low sign diversity
LBP, HOG + Segmentation + SVM	Referee gesture dataset	95.6% Accuracy	Not tailored for sign language; lacks dynamic gesture support
Leap Motion + HMM	Leap Motion-based gestures	WPM, MSD (not percentage)	Sensitive to sensor placement; noise affects accuracy
Manual and Non-manual feature Integration	Small video dataset	73% (20 classes), 80.25% (2 classes) Accuracy	Low signer diversity; poor generalization
CNN + BiLSTM with Attention	ArSL videos and stills	85.60% Accuracy (SI)	Moderate accuracy; lacks robustness across signers
Video Encoder-Decoder	ArabSign (9335 videos, 6 signers)	0.50 WER	Word-level only; limited signer count
CNN-LSTM + Optical flow	mArSL (50 classes, 4 signers)	76% overall, 58.9% (Signer 1, SI)	High preprocessing; Kinect dependency
2D Body and Hand Skeletons	80 videos, 40 signers	88.09% Accuracy (SI)	Limited dataset; low number of signs per signer
Polynomial Classifier vs. ANFIS	ArSL alphabet	Outperformed ANFIS	Trained on non-uniform data
ANFIS for ArSL Alphabet	Static alphabet images	93.55% Accuracy	Focused only on static finger-spelling

Isolated signs, whether static or dynamic, can directly represent spoken words. In continuous sign language databases, these signs overlap due to variations in their execution. The head and hands move vertically both in manual and non-manual recognition of signs [34]. Previous research on video recognition of manual and non-manual signs achieved a precision of 73% for 20 sign classes and 80.25% for two sign classes. Integrating facial expressions with gesture analysis significantly improves sign language comprehension [35]. Fractal-based deformable landmark models have been proposed that provide a geometric approach to 3D gesture recognition in sign language. The use of CNN in conjunction with LSTM or bidirectional LSTM (BiLSTM) layers is a widely known technique in the recognition of sign language [36–38]. An attention-based Arabic Sign Language recognition system on CNN-BiLSTM architecture was able to identify dynamic sign videos and still images of signs in signer-independent mode with 85.60% accuracy.

Al-Hammadi et al. [39] proposed an efficient deep CNN approach for hand gesture recognition. They employed transfer learning to beat the scarcity of a large labeled hand gesture dataset. The evaluation was done using three gesture datasets from color videos: 40, 23, and 10 classes were used from these datasets. The approach achieved recognition rates of 98.12%, 100%, and 76.67%, respectively, in signer-dependent mode. In signer-independent mode, the recognition rates were 84.38%, 34.9%, and 70%, respectively, in the data sets.

When analyzing sign language databases, factors such as signer diversity, the number of signers per sign, and the context of sign depiction within the database must be considered. Another study examined 80 sign videos and sign dynamic videos performed by 40 signers, employing 2D body and hand skeletons. This study achieved an 88.09% prediction accuracy in signer-independent mode [40]. Boukdir et al. [41] used the 2D convolutional recurrent neural network (2DCRNN) and 3D convolutional neural network (3DCNN) and it was reported that these models achieved an accuracy of 92% for 2DCRNN and 99% for 3DCNN. This approach was trained over a collection of 224 videos of five individuals performing 56 different signs. Another advancement in the recognition of ArSL is the 3D GS-NET, a model which is able to recognize signs via RGB videos [42]. The ArabSign dataset was introduced in [4], including six subjects and more than 9335 video samples. They taught a video sign language encoder-decoder model to understand words and achieved an average WER of 0.50.

Alyami et al. [43] proposed a posture-based transformer model for the KArSL-100 that contains videos of 100 different categories for communication through gesture recognition purposes. The pose-based transformer achieved the highest accuracy of 99.74% and 68.2% in signer-dependent and independent modes, respectively. In a similar work, the mArSL dataset was further expanded, encompassing 6748 videos of 50 classes performed by 4 signers. This multi-modal dataset includes RGB images, depth images, joints of the skeleton positions, and facial images of signs and non-signs. This study applied focal-loss based CNN-LSTM fusion–based Head for LSTM animation units and optical flow, with an overall accuracy of 76% achieved with relatively poor 58.9% accuracy for signer 1 SGI signer independent mode [44]. These techniques usually require large preprocessing stages, complicated networks, and Kinect sensors. While several models perform reasonably well on small datasets, excessive reliance on heavy networks and advanced sensors is not suitable for practical implementations.

Inspired by these advancements, our work proposes a hybrid feature-level fusion framework integrating hand-crafted geometric features (joint distances and angles) with deep temporal features from I3D. Distinct from the previous studies focusing either on skeleton-only or vision-only approaches or relying on end-to-end black-box fusion methods, our method permits an interpretable and sufficiently efficient fusion at the feature level, which allows us to maintain a good balance among recognition performance, model transparency, and real-time applicability. Furthermore, while many prior methods require extensive GPU support and long training time, our ensemble classifier works favorably on a CPU with low inference time while attaining a high accuracy.

Significant strides in ArSl recognition and interesting work done in developing methods employing deep learning architectures. However, some challenges remain in the areas of robust model development that are less sensitive to noise and variability in signing style, as well as those used for non-manual features, for a more complete and natural presentation. Therefore, we present a hybrid framework where the models deploy skeleton-based geometric features and deep learning-based temporal features enhancing their performance in the real world with robustness and accuracy. The advancement in depth learning is associated with many works focusing on either skeleton-based recognition or independent consideration of deep temporal features. To the best of our knowledge, prior works have not attempted skeletal geometry (with distances and angles) and I3D-based features for ArSL in a unified feature-level fusion framework. Also, unlike previous methods, our approach utilizes an ensemble of lightweight classifiers that provide better interpretability and generalization. This unique feature allows our model to have both better accuracy and robustness at a very low computational cost, applicable for real-time applications.

3 Dataset

Arabic Sign Language (ArSL) is the main channel of visual communication for individuals who are Deaf and Hard of Hearing (DHH) in Arabic-speaking parts of the world. ArSL, unlike more standardized sign languages like ASL, recognizes a plethora of regional variations, influences from spoken dialects, and the fact that there is no formal grammatical standardization. In facilitating ArSL recognition, the ArabSign dataset has provided a well-curated collection consisting of 9335 video recordings of 50 commonly used sentence-level signs [45]. Each such sentence is a meaningful complete expression: for instance, "Peace be upon you," or "Where are you going?" or "I am hungry." The expressions included are not isolated words but full, short sentences that have frequently been used in real-life contexts. The dataset provides RGB video streams and skeletal joint coordinates corresponding to each signer. Six different native signers were used to perform each sentence multiple times to provide variation in articulation and style. Each video has been pre-segmented to consist of one sentence-level gesture, therefore making the task a fixed-class classification problem. While this study is confined to the 50 classes, it is to be used as a base for vocabulary extension in the near future. Translating unseen sentences is beyond this paper's scope and will be explored in future work on continuous sign language recognition and sentence composition. Fig. 2 shows the two modalities provided for each sentence sample in the dataset.

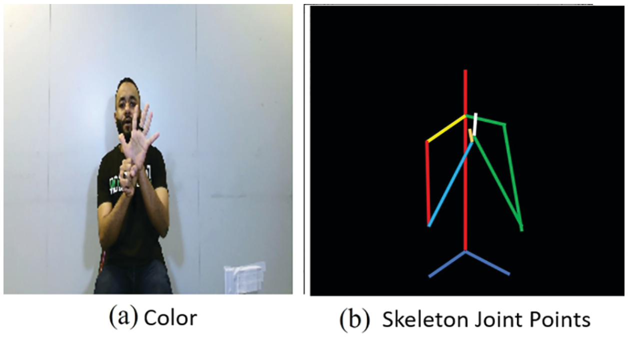


Figure 2: An Illustrative Example from the ArabSign Dataset for the two modalities provided for each sentence sample: (**a**) Color (**b**) Skeleton joint points [45]

4 Proposed Methodology

We propose a multi-modal feature fusion ensemble classification-based approach for Arabic Sign Language (ArSL) recognition through the combination of skeletal and video-based features, followed by ensemble classification. Fig. 3 shows the flow diagram of the proposed method. The first modality considers skeleton data from which we have extracted pairwise geometric features, such as distances and angles between joints, to represent structural motion patterns to be classified. The second modality exploits the video domain, and we customized Inflated 3D ConvNet (I3D) to extract deep temporal features to capture gesture dynamics. In this work, "customized I3D" refers to considering only the RGB stream, which has been adjusted for input resolution and video length with respect to the ArabSign dataset, and extracting intermediate features just prior to final classification. These adjustments thus repurpose I3D into a spatiotemporal feature encoder for fusion with skeleton features without the added side effect of changing anything about the architecture itself. The features pulled out from the two modalities are concatenated, and Feature selection was performed through the Pearson Correlation Coefficient (PCC) on removing those with an absolute correlation value greater than 0.9. The features represented as such have been considered redundant and removed. This threshold was chosen through empirical analysis to balance dimensionality reduction with information retention. The final feature vector retained the most informative uncorrelated features for classification. This feature vector is then fed into three classifiers that include XGBoost, Random Forest, and Support Vector Machine. A majority voting ensemble classifier is then used that takes the output of these three classifiers as input and provides a final decision. The proposed method takes advantage of spatial accuracy from skeletal data and rich motion representation from video-based features, so it fuses both modalities for better recognition. Integrating the two modalities increases the robustness against gesture variability, occlusions, and environmental noise, thus improving the classification performance.

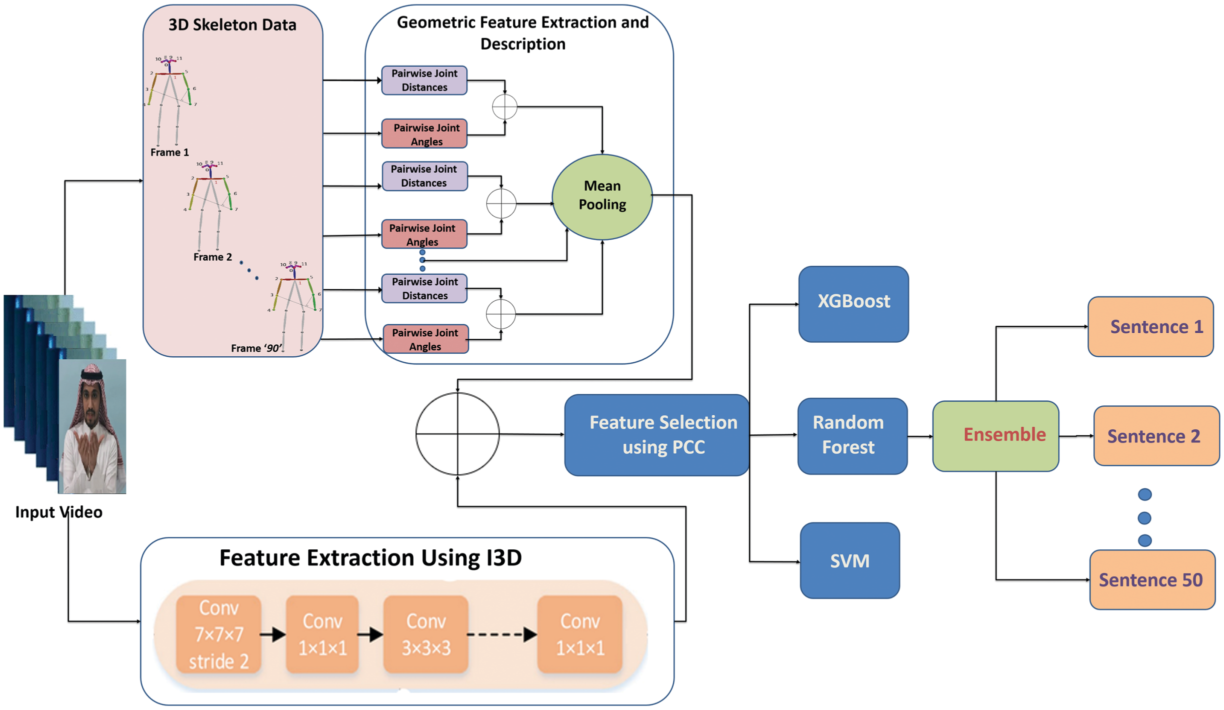


Figure 3: Multi-modal approach for ARSL using skeletel and temporal deep features

4.1 Preprocessing

Gesture recognition in videos is done by providing raw skeleton data in a machine-readable format. The dataset is in the form of `.mat` files, where each file corresponds to a single video. The files contain 3-dimensional coordinates (X, Y, Z) corresponding to 25 joints of the skeleton at every frame. To enable advanced analysis, the `.mat` files are transformed into nicely formatted CSV files, each row of which represents a single frame with the skeleton configuration expressed as a structured vector.

Let a video V be composed of F frames, where each frame contains the 3D coordinates of $J = 25$ joints. We represent a video V as:

$$V = \{F_1, F_2, \ldots, F_F\}, \tag{1}$$

where each frame F_f contains a set of 3D coordinates:

$$F_f = \{(x_j, y_j, z_j) \mid j = 1, 2, \ldots, J\}. \tag{2}$$

Each frame f can be represented as a vector in $\mathbb{R}^{3J}$ space:

$$\mathbf{S}_f = [x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_J, y_J, z_J] \in \mathbb{R}^{3J}. \tag{3}$$

By stacking all frame vectors, we obtain a video matrix representation:

$$\mathbf{M}_V = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_F \end{bmatrix} \in \mathbb{R}^{F\times 3J}. \tag{4}$$

Each row of the matrix corresponds to a frame F_f, and each column represents a joint coordinate (x_j, y_j, z_j). The matrix $\mathbf{M}_V$ is stored as a CSV file, ensuring structured and reproducible processing.

This formalization lays out how raw skeletal data in `.mat` files is transformed into a structured numeric representation $\mathbf{M}_V$, ensuring compatibility with subsequent feature extraction methods, clarifying and standardizing the process, and most importantly, reproducibility and further analysis.

Skeleton Estimation for Unseen Videos: While the ArabSign dataset offers pre-annotated 3D skeletal joint data for each frame, in a real scene or in the case of unseen videos, such skeletal data would require estimation. Our method retains compatibility with most popular pose estimation models like OpenPose, MediaPipe, or Azure Kinect software development kit (SDK) to obtain 2D joint or 3D joint coordinates from an RGB video input. After obtaining these joint coordinates, they could be directly fed into our geometric feature extraction pipeline (pairwise distances and joint angles). Thereby, allowing our method to be executed in real-time as well as new video inputs without altering its underlying structure.

4.2 Feature Extraction

4.2.1 Geometric Feature Extraction and Description Using Framewise Skeleton Data

We obtain geometric features in the form of pairwise joint distances and angles to provide a representation of hand and body movements by obtaining the skeletal data. These properties assist with modeling the spatio-temporal structure of gestures for accurate classification. This ensures efficient recognition while keeping the geometric aspect simple enough for use in real-time applications. After preprocessing, we aim to transform the raw skeleton data into meaningful features that can effectively represent the sign language gestures. The first step in this process involves analyzing the spatial relationships between the joints of the body. The dataset provides 3D coordinates (X, Y, Z) for $J = 25$ distinct joints, representing key points of the human body.

Pairwise Joint Distances: To capture the relative positioning of these joints, we calculate the Euclidean distance between every distinct pair of joints separately for each frame. Since there are $J = 25$ joints, this results in a total of:

$$D = \frac{J(J-1)}{2} = \frac{25 \times 24}{2} = 300 \tag{5}$$

unique joint pair distances per frame.

The Euclidean distance between two joints i and j at any given frame f is computed as:

$$d_{ij}^{f} = \sqrt{(x_i^f - x_j^f)^2 + (y_i^f - y_j^f)^2 + (z_i^f - z_j^f)^2} \tag{6}$$

where x_i^f, y_i^f, z_i^f and x_j^f, y_j^f, z_j^f denote the 3D coordinates of joints i and j at frame f, respectively.

For a video consisting of F frames, the pairwise joint distance feature matrix is represented as:

$$\mathbf{D}_V = \begin{bmatrix} d_{12}^1 & d_{13}^1 & \dots & d_{J-1,J}^1 \\ d_{12}^2 & d_{13}^2 & \dots & d_{J-1,J}^2 \\ \vdots & \vdots & \ddots & \vdots \\ d_{12}^F & d_{13}^F & \dots & d_{J-1,J}^F \end{bmatrix} \in \mathbb{R}^{F \times D}. \tag{7}$$

Joint Angle Representation: Pairwise joint distances are computed separately for each frame, along with the angles between connected joints. These angles describe how different body parts (e.g., arms, hands, shoulders) bend and orient, providing details about the movement dynamics.

The angle θ_{ijk}^f between three adjacent joints i, j, and k at frame f is computed as:

$$\theta_{ijk}^f = \text{atan2}(y_j^f - y_i^f, x_j^f - x_i^f) - \text{atan2}(y_k^f - y_j^f, x_k^f - x_j^f), \tag{8}$$

where atan2 is the four-quadrant inverse tangent function, which calculates the angle between the x-axis and the line connecting two points. This captures the bending of joints, playing an important role in identifying different postures in sign language [46].

The joint-angle feature matrix for a video V is defined as:

$$\theta_V = \begin{bmatrix} \theta_1^1 & \theta_2^1 & \dots & \theta_K^1 \\ \theta_1^2 & \theta_2^2 & \dots & \theta_K^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^F & \theta_2^F & \dots & \theta_K^F \end{bmatrix} \in \mathbb{R}^{F \times K}, \tag{9}$$

where K is the total number of joint angles computed per frame.

Temporal Feature Aggregation: To obtain a compact representation of the video, we perform average pooling across all frames for each distance and angle:

$$\bar{d}_{ij} = \frac{1}{F} \sum_{f=1}^{F} d_{ij}^f, \quad \forall (i, j), \tag{10}$$

$$\bar{\theta}_{ijk} = \frac{1}{F} \sum_{f=1}^{F} \theta_{ijk}^f, \quad \forall (i, j, k). \tag{11}$$

This results in the final feature vectors:

$$\bar{\mathbf{D}} = \left[\bar{d}_{12}, \bar{d}_{13}, \ldots, \bar{d}_{J-1,J}\right] \in \mathbb{R}^{D}, \tag{12}$$

$$\bar{\theta} = \left[\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_K\right] \in \mathbb{R}^{K}. \tag{13}$$

Thus, each video is represented by a single feature vector:

$$\mathbf{F}_V = \left[\bar{D}, \bar{\Theta}\right] \in \mathbb{R}^{D+K}. \tag{14}$$

This pooling process aggregates frame-level features into a single representative vector for the entire video, effectively capturing both spatial relationships and temporal dynamics within the skeleton motion.

Although the average pooling of geometric features over frames is static to represent the videos. It follows since each video is a single, isolated ArSL sentence. Therefore, aggregating gives a complete picture of the geometric footprint for the given gesture, minus the ambiguity caused by unrelated motion. Very importantly, I3D-based spatio-temporal features from the RGB stream keep temporal dynamics, allowing our framework to capture both movement and structure.

4.2.2 Spatio-Temporal Feature Extraction Using Video Sequences

Sign language involves complex and nuanced hand movements, facial expressions, and body postures that convey meaning. To effectively encapsulate these visual patterns, we use the Inflated 3D (I3D) architecture, a deep learning model specifically designed for spatiotemporal data. Specifically, this work makes use of RGB videos, where each video presents one instance of a sign sentence in ArSL. The I3D model is an architecture that processes dynamic input clips and is pre-trained on a large-scale action recognition dataset, and can utilize 3D convolutional kernels to encode the temporal relationship between adjacent frames. This is an important capability for the model to understand the meaning of sign language gestures, as the model needs to understand motion across time [18].

Let V be a sign language video consisting of T frames:

$$V = \{F_1, F_2, \ldots, F_T\} \tag{15}$$

where F_t represents the feature extracted from frame t. To extract meaningful spatiotemporal representations, I3D processes V and extracts optical flow features:

$$O_t = \mathrm{I3D}(F_t) \tag{16}$$

where O_t is the optical flow representation for frame t. To obtain a fixed-size feature representation, we apply mean pooling:

$$F_{\text{mean}} = \frac{1}{T}\sum_{t=1}^{T} O_t \tag{17}$$

where F_{mean} is the aggregated feature vector that represents the motion dynamics of the video.

Our approach combines two complementary sources of information: dynamic motion patterns captured through optical flow features extracted using I3D, and spatial body configurations derived from 3D skeleton data, including distances and angles between joints. Optical flow features effectively represent the motion patterns and time-dependent behavior of gestures. Body pose features provide formation that expresses the structural relations and spatial configuration of body joints during a performance of the sign language. Our framework aims to achieve superior recognition in terms of accuracy and robustness compared to methods relying solely on one type of feature.

Once the optical flow and body pose features are extracted, the two modalities are combined at the feature level. The feature-level combination takes advantage of the complementary strengths of the two modalities, including the temporal dynamics of the gestures using optical flow as well as the spatial arrangement of the signer's movement using body pose. The new aggregate representation provides a complete picture of the sign language performance, allowing for a richer interpretation of the signer's action. We used Pearson Correlation Coefficient (PCC) to remove compute the correlation between features and removed the highly correlated features in order to reduce the size of the feature vector. In our framework, the ensemble classifier comprises XGBoost, RF, and SVM. These classifiers were chosen because they represent great diversity and good interpretability. They also give good performance results on reduced feature sets after the Pearson correlation filtering process. There were considerations for other ensemble models, such as LightGBM and neural classifiers like BiLSTM and transformer-based classifiers, during the exploratory experiments. However, they did not make it into the final ensemble on account of their greater computational complexity and minimal improvement in performance when compared to the current ensemble, which would be more suited for implementing scenarios with real-time or low-resource deployment capabilities. In addition, the temporal features are already captured through the I3D module, which reduces the need for sequence modeling at the classification stage.

Before combining, both I3D and skeleton characteristics were normalized using min-max scaling to ensure a common value range. Temporal information was consolidated into fixed-length feature vectors for each modality using average pooling across frames. These normalized vectors were then concatenated and used as a consistent feature representation in the ensemble classification stage.

We performed feature selection based on Pearson Correlation Coefficient (PCC) to reduce redundancy in features and improve the classification efficiency for the concatenated feature vectors. Features with absolute values of PCC greater than 0.9 were assumed to be highly correlated and hence, discarded. The selection of this threshold was based on empirical considerations to obtain a balance between reducing the dimensionality and maintaining recognition performance. The 0.9 cutoff consistently accepted the most discriminative features while rejecting close to 21% of redundant features, and thus, reduced the overall computational complexity without a proportional loss in accuracy. This strategy guarantees that the final feature vector carries maximum information content with a bare minimum of redundancy thereby improving model performance and interpretability.

4.3 Classification

For classification, we used an ensemble approach utilizing XGBoost, Random Forests, and Support Vector Machine as base classifiers. The rationale behind our design philosophy lies with the real-time performance, interpretability, and generalized robustness sought by a hybrid system. While deep neural networks, such as the BiLSTM or transformers, may be very expressive, their requirements in terms of the amount of training data, memory, as well as available computational resources, are usually far greater than others. On the other hand, our proposed hybrid system adopts deep learning through the I3D architecture for extracting temporal features. The choice was then made to select lightweight classifiers that would suit real-time deployment scenarios, more so if the environments in consideration are less resourceful. The ensemble structure brings another layer of reliability through majority voting, where the different models cast votes for the prediction.

The comprehensive feature vector with their respective sign language sentences as labels is passed to three classifiers, including XGBoost, Random Forest (RF) and Support Vector Machine (SVM). Each model provides its own prediction, and the final recognition result is achieved by combining the outputs of the three models, for example, by majority voting or weighted averaging. This ensemble approach improves classification strength and accuracy by overcoming the weaknesses of each classifier individually and leveraging their combined strengths, resulting in better performance on the challenging task of sign language recognition. The output of these classifiers is then combined using a majority voting ensemble classifier to get the final output label. This multimodal data fusion approach significantly enhances recognition performance by effectively modeling the variability in poses and gestures that arise due to differences in signer styles, speeds, and contexts. The proposed approach demonstrates the ability to predict all combinations of different poses and gestures accurately by unifying information from temporal dynamics and spatial relationships.

5 Experiment Results and Analysis

We performed multiple experiments by varying approaches on skeleton data as well as RGB videos and compared the performance of all these experiments. The performance of the proposed framework was analyzed using the following metrics: accuracy, precision, recall, and F1-score. We also measured computational time to evaluate the feasibility of the framework for real-time gesture recognition applications. A K-fold cross-validation method was employed for validation where the value of K = 10. The data was split into ten equal partitions, where each partition was taken in turn to be the test set while the remaining folds were used for training. Stratified sampling held the full class distribution in each fold to maintain equity and consistency. This stratification was important to preserve the balance of the dataset across iterations, preventing distortions from uneven representation of classes. This k-fold cross-validation made sure that the whole dataset was used for training and testing, just in different runs. This strategy ensured a thorough assessment of the framework's effectiveness on different data segments, thereby improving its reliability and applicability to novel data.

The framework combines three approaches of gesture recognition, which are Maximum based approach, I3D based approach, and the multi-model framework. Data pre-processing, feature extraction, and classification algorithms all varied depending on the methodology to maximize recognition performance. In the skeleton-based approach, human joint positions were taken from a video sequence. These vectors encapsulated spatial relationships and were used for XGBoost classification. The I3D-Based Approach, on the other hand, utilizes sequences of optical flow that were extracted from video data containing a spatial-temporal flow of features using the I3D model.

For the skeleton data, we performed experiments on three feature types: (i) pairwise joint distances calculated over all frames, (ii) pairwise joint angles extracted over all frames, and (iii) combined features with both distances and angles. For effective capture of temporal patterns, each feature representation was pooled using four different pooling methods: sum, average, minimum, and maximum. Each pooling method's performance was rigorously tested. Similarly, we investigated RGB-based and optical flow-based features derived through the I3D architecture. Individual experiments were carried out with RGB features and optical flow features in isolation, and results were derived using the same four pooling methods (sum, average, min, max) for the sake of consistency and a fair comparison. Averaged results for all the experiments mentioned above are shown in Table 2.

An ablation study was done measuring performance loss caused by different levels of adjustment to the threshold used in PCC-based feature reduction, as shown in Fig. 4. It presents the recognition performances (F1-score and accuracy) against the thresholds (0.8–0.95) and the percentage of features retained. Results indicate that aggressive thresholds (e.g., threshold = 0.8) can hurt accuracy mildly, while the threshold of

0.9 appeared to achieve the best compromise, with about 21% of features being eliminated, but with the highest F1-score (0.97) still being maintained. This is evidence of the effectiveness of our correlation-based pruning strategy.

Table 2: Results achieved from ablation study by varying different experimental settings

Exp#	Method	Min (F1)	Max (F1)	Sum (F1)	Avg (F1)
1	ArSL Recognition using Pairwise Joint Distances	0.70	0.75	0.78	0.81
2	ArSL Recognition using Pairwise Joint Angles	0.68	0.76	0.77	0.82
3	ArSL Recognition using Fused Distances and Angles	0.77	0.82	0.84	0.92
4	ArSL Recognition using Framewise RGB Features	0.71	0.73	0.79	0.80
5	ArSL Recognition using Optical Flow Features	0.74	0.81	0.85	0.87
6	ArSL Recognition using Fused RGB and Flow Features	0.80	0.83	0.88	0.91
7	ArSL Recognition using Hybrid Approach (Geometric and Optical Flow)	0.88	0.90	0.94	0.96
8	ArSL Recognition using Hybrid Approach (Geometric and Optical Flow) With PCC for Feature Selection	0.88	0.91	0.95	0.97

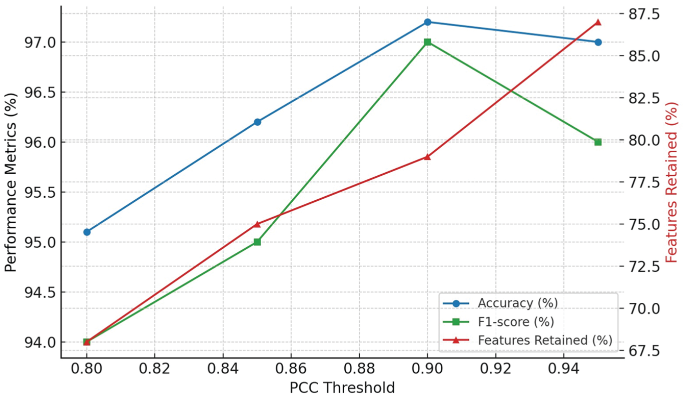
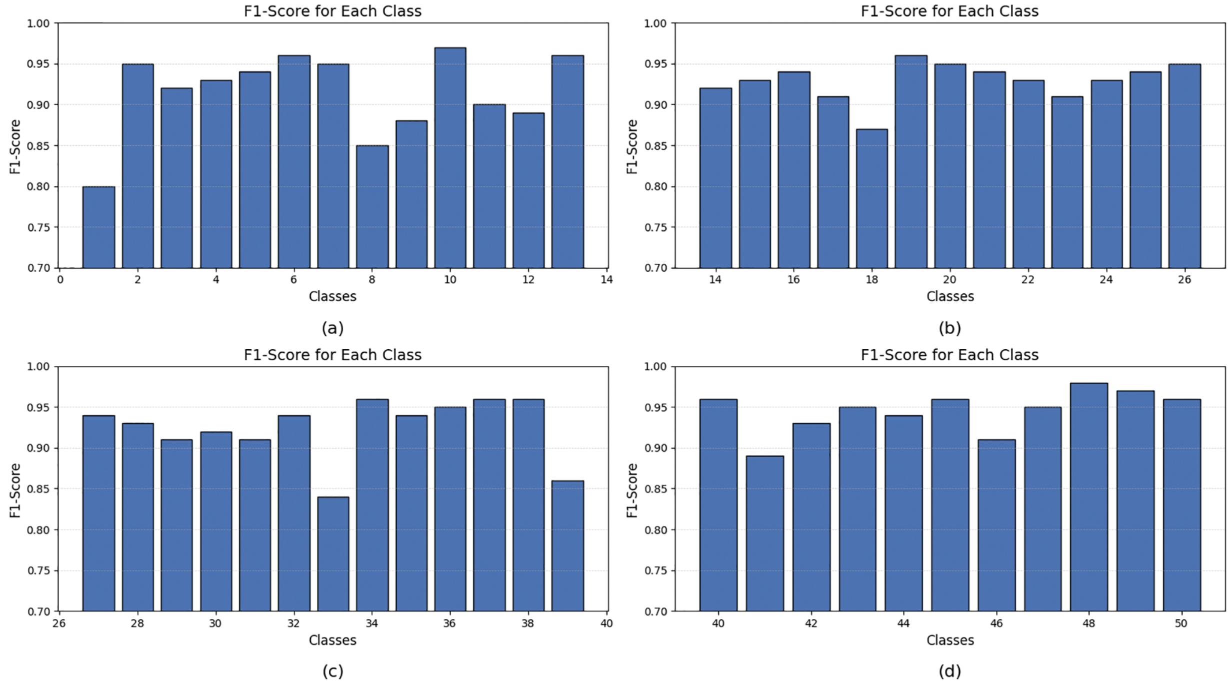


Figure 4: Effect of Pearson Correlation Coefficient (PCC) threshold on feature selection and Arabic Sign Language recognition performance

These features were mean-pooled to retain temporal information and then classified with XGBoost. Next, we combine the body pose features of the Skeleton-Based Approach with the spatiotemporal features from the I3D-Based Approach into a comprehensive feature vector in the Multi-model framework. This integrated feature vector, comprising 102 features, was then classified using the XGBoost algorithm to leverage the strengths of both spatial and temporal information. Experiments were performed using a private desktop computer with Intel(R) Core(TM) i5-8265U CPU 1.60 GHz (up to 1.80 GHz with turbo boost), and 8.00 GB RAM. For efficient processing and reproducibility, the scikit-learn library was used for the implementation and configuration of the classifiers.

The experimental outcome of skeleton data-based ArSL recognition shows the robustness of the system with an average accuracy, precision, and recall of 91%, and an F1-score of 93%. Certain classes, including 13, 19, 36, and 49, experienced outstanding results with accuracy and F1-scores above 97%. Low accuracies (less than 85%) were exhibited in some classes (1, 8, 12, 33, and 39), which was probably because of overlapping motion patterns or due to the representation of the data. Fig. 5 and Table 3 visualize its performance to evaluate the model's fine-grained performance level. The results provide a view of how well the model generalizes over diverse signs while pinpointing its merits and demerits of specifying classes. Such a detailed analysis is important for understanding areas where the model is doing well and those that may need improving.



(a)

(b)

(c)

(d)

Figure 5: A skeleton utilizing geometric features such as pairwise joint distances and angles encoded from 3D pose data obtains class wise F1-scores. This figure shows the recognition performance of the model when only relying on structural features derived from skeleton data, disregarding motion or optical flow information. Due to the high number of classes (50 sign sentences), the results were divided into four subplots for clarity: (**a**) F1-scores for Classes 1–13, (**b**) F1-scores for Classes 14–26, (**c**) F1-scores for Classes 27–39, and (**d**) F1-scores for Classes 40–50. Such a segmented display enhances readability and illustrates fine-grained performance trends across classes under the skeleton-alone model. These results are then used as a benchmark to compare across temporal features (Fig. 6) and the proposed hybrid framework (Fig. 7)

Table 3: Experimental results using skeleton data for recognition of ArSL

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
1	0.73	0.88	0.73	0.80	30	11	4	1780
2	0.94	0.90	0.94	0.95	36	0	4	1785
3	0.85	0.89	0.85	0.92	33	6	0	1786
4	0.94	0.89	0.94	0.92	34	2	4	1785
5	0.95	0.93	0.95	0.95	39	2	2	1782
6	0.94	0.95	0.94	0.96	34	2	1	1788
7	0.93	0.89	0.93	0.94	40	0	5	1780
8	0.80	0.82	0.80	0.81	32	8	7	1778
9	0.94	0.80	0.94	0.87	33	2	8	1782
10	0.94	0.95	0.94	0.96	34	2	1	1788
11	0.90	0.83	0.90	0.90	30	1	6	1788
12	0.81	0.90	0.81	0.88	34	8	1	1782
13	0.97	0.94	0.97	0.97	30	1	1	1793
14	0.94	0.86	0.94	0.93	32	0	5	1788
15	0.92	0.90	0.92	0.93	35	1	4	1785
16	0.94	0.91	0.94	0.94	32	1	3	1789
17	0.89	0.90	0.89	0.92	35	2	4	1784
18	0.86	0.89	0.86	0.87	31	5	4	1785
19	0.97	0.95	0.97	0.96	35	1	2	1787
20	0.91	0.97	0.91	0.96	32	3	0	1790
21	0.89	0.94	0.89	0.94	31	4	0	1790
22	0.97	0.92	0.97	0.95	35	1	3	1786
23	0.86	0.89	0.86	0.89	32	4	4	1785
24	0.92	0.90	0.92	0.93	33	3	2	1787
25	0.94	0.92	0.94	0.93	34	2	3	1786
26	0.94	0.88	0.94	0.94	36	0	5	1784
27	0.95	0.91	0.95	0.94	41	1	4	1779
28	0.88	0.96	0.88	0.93	42	6	0	1777
29	0.86	0.89	0.86	0.91	31	5	1	1788
30	0.89	0.90	0.89	0.91	32	4	2	1787
31	0.89	0.95	0.89	0.94	32	4	0	1789
32	0.92	0.92	0.92	0.94	33	3	1	1788
33	0.81	0.85	0.81	0.83	29	7	5	1784
34	0.97	0.94	0.97	0.96	34	1	2	1788
35	0.92	0.90	0.92	0.94	34	2	2	1787
36	0.97	0.90	0.97	0.95	35	1	3	1786
37	0.92	0.95	0.92	0.96	33	3	0	1789
38	0.94	0.94	0.94	0.94	34	2	2	1787
39	0.81	0.92	0.81	0.87	29	7	2	1787
40	0.94	0.97	0.94	0.96	34	2	1	1788
41	0.89	0.89	0.89	0.89	32	4	4	1785

(Continued)

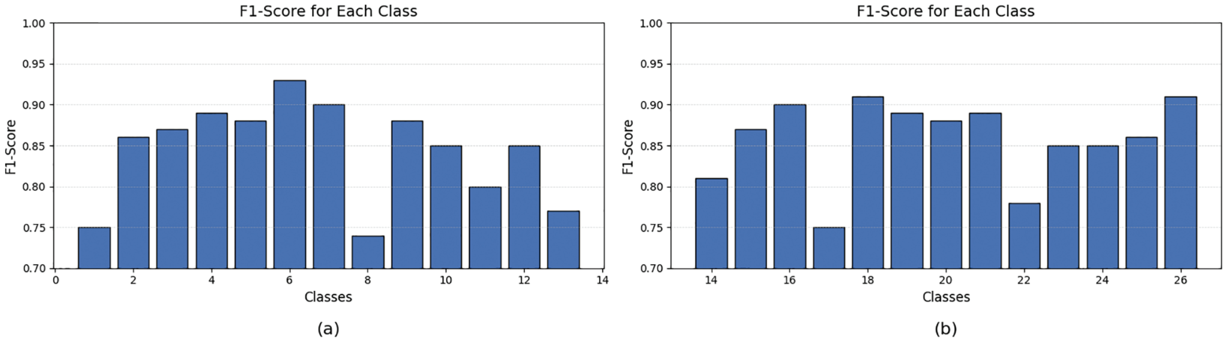
Table 3 (continued)

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
42	0.92	0.94	0.92	0.93	33	3	2	1787
43	0.94	0.95	0.94	0.96	34	2	1	1788
44	0.92	0.90	0.92	0.93	35	1	4	1785
45	0.94	0.97	0.94	0.97	34	2	0	1789
46	0.92	0.95	0.92	0.96	33	3	0	1789
47	0.89	0.94	0.89	0.91	32	4	2	1787
48	0.94	0.95	0.94	0.96	34	2	1	1788
49	0.97	0.97	0.97	0.99	35	1	0	1789
50	0.92	0.92	0.92	0.94	33	3	1	1788
Average	0.91	0.91	0.91	0.92				

Misclassification analysis revealed low false negatives for most classes, indicating that true gestures were rarely missed. However, higher false positives in classes like 8 and 9 suggested confusion with similar gestures. Precision and recall remained consistent across most classes, with some variation highlighting areas for refinement. High F1-scores across the majority of classes underscore the system's potential for applications in education and accessibility for individuals with hearing impairments.

For lower-performing classes, improvements such as advanced feature engineering, hybrid approaches, or integrating additional modalities like RGB or depth data may enhance performance. Testing in real-time and diverse environments would further validate and expand the system's applicability. Overall, the results confirm the system's effectiveness and provide a foundation for future enhancements.

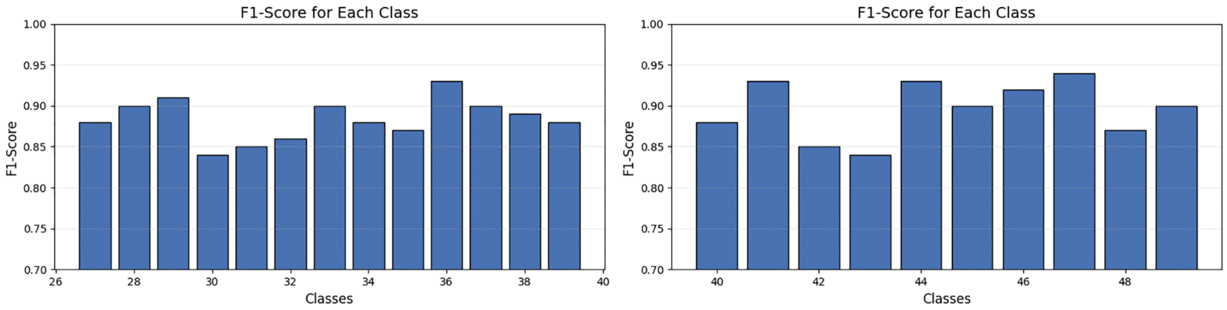
The recognition results based on I3D optical flow features are available in Fig. 6 and Table 4. These insights offer a comparative perspective on the standalone temporal model vs. the hybrid model. With an F1 score of 87%, the system attained an average accuracy, precision, and recall of 85%. Classes 6, 36, and 48 demonstrated strong recognition performance, achieving accuracy and F1-scores exceeding 90%.



(a)

(b)

Figure 6: (Continued)



(c)

(d)

Figure 6: Class-wise percentages of F1 scores obtained using the optical flow features extracted through the I3D (Inflated 3D ConvNet) architecture. The figure thus shows the recognition performance with respect to temporal motion dynamics captured solely from RGB videos neglecting skeletal geometry features. Due to the total of 50 sign sentence classes, results were divided into four subplots for readability: (**a**) F1-scores for Classes 1–13, (**b**) F1-scores for Classes 14–26, (**c**) F1-scores for Classes 27–39, and (**d**) F1-scores for Classes 40–50. This accordingly facilitates the visual representation of individual class performance under purely temporal feature settings. The results are then further used for comparison against the skeletal based model (Fig. 5) and the hybrid fusion (Fig. 7)

Table 4: Experimental results using optical flow features (I3D) for recognition of ArSL

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
1	0.68	0.78	0.68	0.75	29	12	7	1784
2	0.83	0.83	0.83	0.86	30	6	4	1792
3	0.85	0.84	0.85	0.88	36	3	7	1786
4	0.83	0.92	0.83	0.88	30	6	2	1794
5	0.90	0.84	0.90	0.87	37	4	7	1784
6	0.94	0.90	0.94	0.93	34	2	3	1793
7	0.90	0.85	0.90	0.90	39	2	7	1784
8	0.85	0.84	0.85	0.86	35	5	6	1786
9	0.74	0.76	0.74	0.75	26	9	8	1789
10	0.83	0.87	0.83	0.87	30	6	3	1793
11	0.83	0.73	0.83	0.80	32	4	12	1784
12	0.76	0.83	0.76	0.85	32	10	1	1789
13	0.71	0.72	0.71	0.77	25	6	9	1792
14	0.75	0.75	0.75	0.81	24	8	3	1797
15	0.89	0.84	0.89	0.86	32	4	6	1790
16	0.91	0.84	0.91	0.90	32	1	6	1793
17	0.73	0.79	0.73	0.76	27	10	7	1788
18	0.89	0.85	0.89	0.91	35	1	6	1790
19	0.83	0.91	0.83	0.89	31	5	3	1793
20	0.88	0.83	0.88	0.87	30	3	6	1793
21	0.91	0.83	0.91	0.88	33	2	7	1790
22	0.72	0.82	0.72	0.78	26	10	5	1791
23	0.86	0.86	0.86	0.86	31	5	5	1791
24	0.86	0.84	0.86	0.85	31	5	6	1790

(Continued)

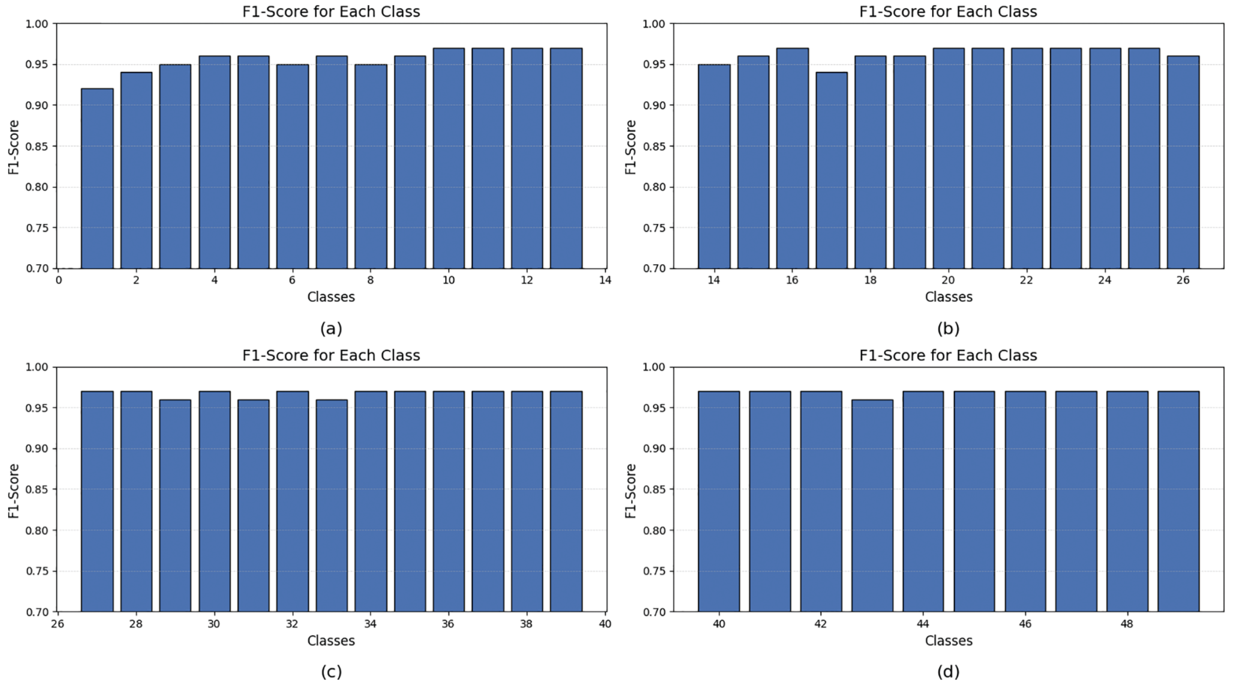
Table 4 (continued)

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
25	0.89	0.79	0.89	0.85	33	3	9	1787
26	0.92	0.87	0.92	0.91	34	2	5	1791
27	0.89	0.85	0.89	0.88	42	4	7	1779
28	0.88	0.87	0.88	0.90	42	6	3	1781
29	0.89	0.92	0.89	0.91	32	4	2	1794
30	0.78	0.85	0.78	0.84	29	8	3	1792
31	0.86	0.82	0.86	0.85	32	4	7	1789
32	0.89	0.91	0.89	0.90	32	4	3	1793
33	0.86	0.84	0.86	0.85	31	5	6	1790
34	0.86	0.87	0.86	0.88	30	5	3	1794
35	0.86	0.89	0.86	0.87	31	5	4	1792
36	0.94	0.92	0.94	0.93	34	2	3	1793
37	0.86	0.92	0.86	0.90	31	5	2	1794
38	0.83	0.92	0.83	0.88	30	6	2	1794
39	0.86	0.83	0.86	0.88	33	3	6	1790
40	0.86	0.89	0.86	0.87	31	5	4	1792
41	0.92	0.90	0.92	0.93	33	3	2	1794
42	0.83	0.79	0.83	0.85	29	6	4	1793
43	0.81	0.88	0.81	0.84	29	7	4	1792
44	0.91	0.89	0.91	0.93	32	3	2	1795
45	0.92	0.89	0.92	0.90	33	3	4	1792
46	0.86	0.89	0.86	0.89	32	4	4	1792
47	0.86	0.94	0.86	0.91	31	5	1	1795
48	0.94	0.92	0.94	0.94	34	2	2	1794
49	0.86	0.79	0.86	0.86	34	2	9	1787
50	0.89	0.89	0.89	0.89	32	4	4	1792
Average	0.85	0.85	0.85	0.87				

In contrast, classes 1, 9, and 22 exhibited lower performance, with accuracies below 75%. These disparities in performance may be attributed to factors such as overlapping gesture patterns or limitations in the training data for these specific classes. The system attained an average of 85% accuracy, precision, and recall, with 87% F1-score. 339 patterns were found common to most classes, based on true positives (TPs), false negatives (FNs), and false positives (FPs) analysis. Classes with high false positives like 9 and 13 experienced high misclassification that reduced the precision. Likewise, high false negatives for some classes, 22 and 17 among them, affected recall. It exhibited well-balanced precision against recall, proving robust performance despite these challenges.

The findings indicate that some classes may benefit from the inclusion of other features, or data modalities to improve recognition. In future work, we could look at hybrid methods, or more complex deep learning methods for improving robustness, and real time applicability, particularly for complex gestures. In general, the results underscore both the system's effectiveness and its capacity for scalable applications in accessibility and education.

The hybrid framework for ArSL recognition demonstrates in Fig. 7 and Table 5 consistently high performance across all 50 classes, achieving an average accuracy, precision, recall, and F1-score of 0.97. These results highlight the effectiveness of the proposed approach. Notably, individual class results reveal that most classes exhibit accuracies exceeding 0.95, with many achieving values of 0.99 or higher, demonstrating the capability of the framework to classify a diverse range of sign-language gestures accurately. Values for all classes of precision and recall are proportionately high, similar to how we typically fall between 0.95 and 1.00, meaning there are very few false positives and false negatives, respectively. The mean scores of 0.97 for the F1 score show a successful balance between precision and recall.



(a)

(b)

(c)

(d)

Figure 7: Class wise F1-scores, derived from the hybrid framework proposed, which integrates geometric features (joint distances and angles) in the form of skeleton data and deep temporal features extracted using the I3D model. This multimodal feature fusion exploits both structural and motion cues, aiming at recognition improvement. Results are distributed into four subplots to maintain visual clarity over the 50-class sign vocabulary: (**a**) F1-scores for Classes 1–13, (**b**) F1-scores for Classes 14–26, (**c**) F1-scores for Classes 27–39, and (**d**) F1-scores for Classes 40–50. This figure indicates the superiority of this hybrid model for all sign classes, subsequently furthers a comprehensive comparison with unimodal baselines seen in Figs. 5 and 6

Table 5: Experimental results using hybrid framework for recognition of ArSL

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
1	0.91	0.95	0.91	0.93	37	3	2	1777
2	0.98	0.97	0.98	0.97	35	0	1	1783
3	0.95	0.97	0.95	0.96	36	2	1	1780
4	0.99	0.97	0.99	0.98	35	0	1	1784
5	0.99	0.96	0.99	0.97	40	0	1	1778
6	0.99	0.98	0.99	0.98	35	0	0	1784

(Continued)

Table 5 (continued)

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
7	0.99	0.96	0.99	0.98	39	0	1	1779
8	0.95	0.98	0.95	0.96	37	2	0	1780
9	0.98	0.96	0.98	0.97	34	0	1	1784
10	0.96	0.97	0.96	0.97	34	1	1	1783
11	0.98	0.98	0.98	0.98	30	0	0	1789
12	0.95	0.97	0.95	0.96	39	2	1	1777
13	0.99	0.96	0.99	0.97	30	0	1	1788
14	0.95	0.97	0.95	0.96	30	1	1	1788
15	0.96	0.97	0.96	0.97	34	1	1	1784
16	0.98	0.97	0.98	0.98	32	0	0	1787
17	0.93	0.95	0.93	0.94	34	2	1	1782
18	0.96	0.94	0.96	0.95	34	1	2	1782
19	0.96	0.96	0.96	0.96	34	1	1	1783
20	0.97	0.97	0.97	0.97	32	1	1	1786
21	0.97	0.97	0.97	0.97	33	1	1	1784
22	0.97	0.98	0.97	0.97	34	1	0	1784
23	0.99	0.96	0.99	0.97	35	0	1	1783
24	0.97	0.98	0.97	0.98	34	1	0	1784
25	0.99	0.97	0.99	0.98	35	0	1	1783
26	0.97	0.96	0.97	0.96	35	1	1	1783
27	0.96	0.97	0.96	0.97	40	1	1	1777
28	0.98	0.97	0.98	0.97	47	0	1	1771
29	0.98	0.98	0.98	0.98	35	0	0	1784
30	0.92	0.97	0.92	0.95	33	2	0	1784
31	0.98	0.99	0.98	0.98	35	0	0	1784
32	0.97	0.98	0.97	0.97	35	1	0	1784
33	0.95	0.95	0.95	0.95	34	1	1	1783
34	0.99	0.98	0.99	0.98	34	0	0	1785
35	0.97	0.97	0.97	0.97	35	1	1	1784
36	0.99	0.98	0.99	0.98	35	0	0	1784
37	0.97	0.99	0.97	0.98	34	1	0	1784
38	0.97	1.00	0.97	0.98	35	1	0	1784
39	0.97	0.99	0.97	0.98	34	1	0	1784
40	0.99	0.98	0.99	0.99	35	0	0	1784
41	0.99	0.98	0.99	0.98	35	0	0	1784
42	0.98	0.99	0.98	0.99	34	0	0	1785
43	0.98	0.99	0.98	0.98	35	0	0	1784
44	0.98	0.97	0.98	0.98	34	0	0	1785
45	0.99	0.99	0.99	0.99	35	0	0	1784
46	0.98	0.96	0.98	0.97	35	0	1	1783
47	0.98	0.99	0.98	0.98	35	0	0	1784

(Continued)

Table 5 (continued)

Classes	Acc.	Pre.	Rec.	F1	TPs	FNs	FPs	TNs
48	0.99	0.99	0.99	0.99	35	0	0	1784
49	0.99	0.98	0.99	0.98	35	0	0	1784
50	0.99	0.99	0.99	0.99	35	0	0	1784
Average	0.97	0.97	0.97	0.97				

The true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) are also better in this light. The TP values are high for most classes, which justifies the proposed approach in precisely classifying the target gestures. Similarly, small FN and FP counts (zero in many cases) point to a strong misclassification reduction capability of the framework. The TN values are consistently high, further emphasizing the framework's ability to correctly classify non-target gestures.

This indicates that the hybrid framework is highly accurate and reliable for the recognition of ArSL with very low error rates for all classes. Moreover, the excellent performance over a large number of gestures not only proves its scalable but also demonstrates its potential for practical application in gestures recognition systems.

The performance of our hybrid model is depicted in Table 5. These representations combine skeletal and temporal features, with almost all the classes showing huge improvements, verified further against previous methodologies in Table 6, respectively.

Table 6: Summary of Arabic Sign Language (ArSL) recognition methods and their reported performance on respective datasets. Note: These comparisons are provided for contextual understanding. Due to differences in datasets, class definitions, signer variability, and metrics, these results are not directly comparable

Study	Methodology	Dataset	Accuracy
Abdul et al. [47]	CNN-LSTM	KSArSL	85.60%
Sidig et al. [48]	CNN-LSTM	KArSL-150	43.62% SI
Suliman et al. [49]	KNN	KArSL-100	58% SI
Aly et al. [50]	CSOM, BiLSTMs	KSArSL	89.59% SI
Proposed	Hybrid (Skeleton + I3D), Ensemble classifier	ArSL (6 signers, 50 sentences)	97.20%

An analysis was performed to show the benefits of an ensemble classifier by comparing it with the individual models. With the same fused feature set, XGBoost produced an average F1-score of 95.5%, Random Forest produced 94.7%, and SVM produced 93.9%. In contrast, the ensemble classifier registered a higher F1-score of 97.0%, thereby reflecting a gain of 1.5% over the best individual classifier (XGBoost). Such a gain provides an indication of improved generalization and a mitigation of the limitations posed by individual models in cases of classes that share overlapping or ambiguous gestures.

5.1 Discussion

The bar graph shown in Fig. 8 shows the average F1-scores achieved by ArSL recognition from I3D-Flow features, skeleton Features, and a hybrid Framework. A gradual performance increase is observed as more feature types are combined. The I3D-Flow features, while effectively capturing temporal and motion-based information, achieved an average F1-score of 86.8%. Notably, skeleton features alone surpassed this performance, reaching an average F1-score of 92.6% per gesture, suggesting that the structurally representative dynamics of hand and body movements have a much better discrimination power for the sign language gestures.

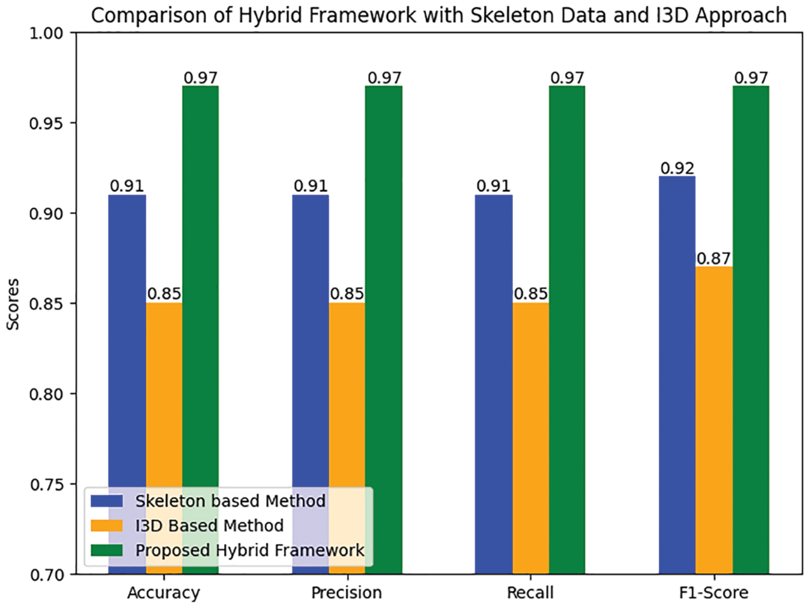


Figure 8: Comaprison of hybird framework with skeleton data approach and optical flow featues (I3D) approach

As a result, with the fusion of I3D-Flow and skeleton features, the hybrid framework achieved the highest performance, a mean F1-score of 97.3%. This substantial improvement demonstrates the complementary nature of the combined features, as the Hybrid Framework effectively leverages both motion and structural information for robust sign language recognition. The results underscore the significance of multi-modal feature fusion in achieving superior accuracy and reliability in complex gesture recognition tasks. This fusion approach not only enhances classification performance but also demonstrates the potential for real-world applicability in sign language translation systems.

One important thing to keep in mind is that while our primary focus is on Arabic Sign Language (ArSL), the design of our skeleton and temporal deep feature based multimodal fusion model is generally applicable to different varieties of sign languages. The challenges that our model addresses, gesture variability, occlusion, and differences in the characteristics of signers, are common to all sign language systems. Thus, the proposed framework will serve as a basis for adapting other languages with minimal modifications, and this cross linguistic extension remains an active focus of our research going forward.

Table 6 illustrates different methods for Arabic Sign Language (ArSL) recognition, showing the better performance of our Hybrid (Skeleton + I3D) XGBoost method. Conventional methods, such as CNN-LSTM, yielded mixed results with accuracy between 43.62% and 85.60%, being unable to handle dataset changes and signer independence. KNN, at 58% accuracy, did not have temporal modeling, whereas Convolutional Self-Organizing Map with Bidirectional Long Short-Term Memory (CSOM-BiLSTMs) enhanced performance to 89.59% but did not use multimodal data.

Most sign language recognition systems use the CNN-LSTM or BiLSTM architecture, but they often fail to perform well in either signer independent or low-resource conditions. Our hybrid model, which combines deep I3D-based features with a lightweight ensemble classifier, is better generalized and more robust for deployment in real settings. While Table 6 covers an overview of reported performances in various studies, methods differ in dataset and experimental protocols, which calls for doing this comparison in the realm of context-building advancement in ArSL recognition.

Our proposed method far surpasses all existing methods, with 97.20% accuracy on an ArSL dataset of six signers and fifty sentences. By combining skeleton-based joint features with I3D motion features and applying XGBoost with ten-fold cross-validation, our method improves spatial and temporal perception. In contrast to existing methods, it effectively addresses signer variability, enhances robustness in classification, and is more scalable for practical use. This sets our model as a new standard for ArSL recognition with higher accuracy and generalization.

As illustrated in Table 6, the existing traditional CNN, LSTM and KNN methods only achieve a fair performance level, which falls in a range from 43.6% to 85.6%, especially under signer independence conditions. Whereas, the hybridization method could significantly improve the gain up to an accuracy value of 97.2% on the ArabSign dataset. This improvement indicates the impressive power of multimodal feature fusion as well as ensemble classification in trying to overcome gesture complexity and signer variability effects.

Despite the ArabSign dataset encompassing pertinent sentence-level information with synchronized RGB and skeletal modalities, it is afflicted by the limitation of only six unique signers. Such restricted diversity in signers might hinder the generalization of the models to user populations characterized by larger variations in physical traits, signing styles, and articulation speeds. Our stratified 10-fold cross-validation with multimodal feature fusion would reduce the impact of overfitting and allow for some degree of robustness against variability in signers. Therefore, performance on signers who have not been trained could emerge as an important consideration. In the future, validation frameworks will need to explore more diverse and larger-scale ArSL datasets to include a higher number of signers from varied demographic backgrounds to further test the scalability and signer independence of the model within real-life scenarios.

5.2 Time and Performance Analysis: Hybrid Framework vs. Skeleton Approach vs. Flow Feature Approach

The bar graph in Fig. 9 compares three methods for Arabic Sign Language Recognition in terms of inference time and F1-score, showing the trade-offs between computational efficiency and recognition performance. The skeleton-based method has the shortest inference time of 0.5 seconds, which makes it suitable for real-time applications while maintaining a high F1-score of 0.92, indicating reliable accuracy. Instead, the I3D-based method requires a whopping 1.2 s to perform an inference, more than twice what is needed for the skeleton-based method, and still yields a slightly lower F1-score of 0.87, revealing that more complex spatiotemporal features are not of benefit in any way. Although the hybrid framework incurs a maximum inference time of 1.5 s, it yields the highest F1-score of 0.97, proving its high accuracy and robustness in sign language gesture recognition. The result suggests that a multi-technique or multi-modal hybrid combination can effectively deliver the best recognition performance at the expense of processing speed. Lastly, results have been observed in the nature of the trade-off between efficiency and accuracy; that is, skeleton-based is perfect for applications requiring rapid results, and hybrid is well-suited when the emphasis of the application is towards accuracy.

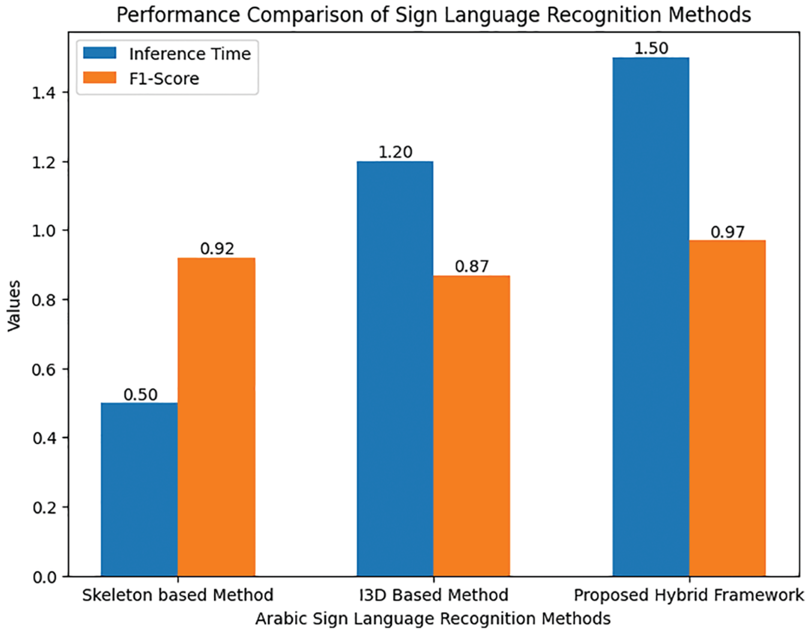


Figure 9: Time and Performance Analysis: skeleton based method vs. I3D based method vs. proposed hybrid framework (Unit Time is Second)

Proposed systems are designed to be modular and adaptable. The skeleton-only approach performs well with virtually no latency in real-time applications and will not use GPU acceleration. It is a little slower than hybrid models but achieves state-of-the-art recognition accuracy, which is best suited for high-precision offline applications. Such flexibility allows the systems to be configured and deployed according to the different hardware platforms.

5.3 Comparison with Existing Methods on the ArabSign Dataset

In conducting a comprehensive and principled evaluation, we make a sincere attempt at a head-to-head comparison involving our proposed hybrid approach and some previously reported ones tested on the ArabSign dataset. A notable mention is the encoder-decoder system reported in [47], which obtained the word error rate of 0.50 on this dataset. However, the said authors basically focused on sentence-level translation without coming out to report standard classification measures F1-score or accuracy. Whereas our model achieved an F1-score of 97.3% using a 10-fold cross-validation strategy that respects signer independence conditions, markedly improving recognition accuracy and reliability. Also, while the methods reported in literature often traffic in complex encoder-decoder architectures or require heavy GPU usage, our method achieves high performance using a lightweight ensemble classifier tuned for real-time deployment. To our knowledge, no other recent methods have reported per-class or aggregated F1-scores over the entire 50-class sentence-level ArSL dataset. We encourage future benchmarking on ArabSign to more uniformly validate and compare methods under controlled experimental settings.

6 Conclusion and Future Directions

A major challenge for Arabic Sign Language (ArSL) recognition is the lack of large annotated datasets, restricting the training of generalizable models. Moreover, the accuracy of recognition is also affected by gesture variability, occlusions, and environmental noise. Finally, such gaps stifle the creation of strong, generalizable models in various real-world environments. To tackle these issues, we introduce an ensemble approach that combines pairwise joint distances and angles from 3D skeletal data with optical flow information. This method achieves complementarity between skeleton-based temporal features, focusing on detailed appearance and motion information, and deep learning-based temporal features, learning holistic gesture features through the use of higher-level insights. Finally, we used XGBoost for classification, which reduces model training time by modeling feature interactions well. We find that our model surpasses the performance of models relying only on subsets of features, such as single feature-based models, where our model achieves a 97.3% average F1 score on the ArabSign dataset. It shows that our method is new as it fills the gap between research-oriented methods and practical methods that can be used on the ground for real solutions. Our model achieves robust recognition through a real-time fusion of multiple feature types, handling occlusions, background noise, and inter- and intra-sign variability. Although our ensemble method performs well, many aspects could be subject to further research. This could greatly alleviate the burden of communication in various situations that require assistive support, where ArSL recognition deployment in edge devices or mobile applications would enable users to communicate assertively and in real-time. And, scaling the system to offer additional gestures and regional dialects of sign language at the same time can make it more useful for a wider audience. Moreover, self-supervised or few-shot learning approaches are open to investigation to minimize reliance on larger annotated datasets, potentially enhancing model generalizability with little labeled data. With these progressions, our work can foster scalable, efficient, and real-time ArSL recognition systems and improve communication accessibility for the Deaf and Hard of Hearing (DHH) community. Although the work being reported was performed on ArSL, the suggested hybrid framework is, by nature, language agnostic and can be adapted to other SLs such as ASL, CSL, and DGS, provided similar RGB and skeleton data are available. In the subsequent step, we plan to evaluate our model on common datasets for ASL and CSL and thereby assess its generalizability with respect to possible differential linguistic and cultural variations in the sign language itself. Our framework for isolated sign recognition is paving the road to continuous ArSL understanding. Owing to its modularity, it can easily be extended to sequence modeling with temporal segmentation and context-aware decoding. Full hybrid model inference time is around 1.5 s, but skeleton-based human-ML-world hybrid frameworks, requiring only 0.5 s for inference, can be made real-time with flexibility in accuracy-speed trade-off.

In terms of ethical constructs, in the future, the deployment of sign language recognition systems should guarantee stringent controls on data privacy. This includes the collection of video or skeletal motion data from real users. The recognition system should also include signers with physical disabilities and those who use signs that are regionally different. This will ensure the inclusivity and equity in the training and evaluation of such technologies, thereby justifying the responsible and fair consumption of such technologies.

Acknowledgement: The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group No. KS-2024-376.

Funding Statement: The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group No. KS-2024-376.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Syed Muhammad Usman, Yazeed Alkhrijah, and Danish Hamid; Methodology, Yazeed Alkhrijah, Shehzad Khalid, and Danish Hamid; Software, Yazeed Alkhrijah, Amina Jameel, and Danish Hamid; Validation, Yazeed Alkhrijah, Shehzad Khalid, Amina Jameel, and Danish Hamid; Formal analysis, Yazeed Alkhrijah, Amina Jameel, and Danish Hamid; Investigation, Yazeed Alkhrijah, Shehzad Khalid, and Danish Hamid; Resources, Syed Muhammad Usman and Shehzad Khalid; Data curation, Amina Jameel, Yazeed Alkhrijah, and Danish Hamid; Writing—original draft preparation, Yazeed Alkhrijah, Amina Jameel, and Danish Hamid; Writing—review and editing, Syed Muhammad Usman, Shehzad Khalid, and

Danish Hamid; Visualization, Amina Jameel, Yazeed Alkhrijah, and Danish Hamid; Supervision, Syed Muhammad Usman and Shehzad Khalid; Project administration, Syed Muhammad Usman; Funding acquisition, Yazeed Alkhrijah, Shehzad Khalid, Syed Muhammad Usman, and Amina Jameel. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data and material will be made available upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest to report regarding the present study.

References

1. Ibrahim NB, Selim MM, Zayed HH. An automatic Arabic sign language recognition system (ArSLRS). J King Saud Univ Comput Inf Sci. 2018;30(4):470–7. doi:10.1016/j.jksuci.2017.09.007.
2. Neupane RB, Li K, Boka TF. A survey on deep 3D human pose estimation. Artif Intell Rev. 2024;58(1):24. doi:10.1007/s10462-024-11019-3.
3. Koller O, Zargaran S, Ney H, Bowden R. Deep sign: hybrid CNN-HMM for continuous sign language recognition. In: British Conference on Machine Vision (BMVC); 2016 Sep 19–22; York, UK. p. 1–12. doi:10.5244/C.30.136.
4. Luqman H. ArabSign: a multi-modality dataset and benchmark for continuous Arabic sign language recognition. In: Proceedings of the 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG); 2023 Jan 5–8; Waikoloa Beach, HI, USA. p. 1–8.
5. Ye L, Ye S. Deep learning for skeleton-based action recognition. J Phys Conf Ser. 2021;1883:012174. doi:10.1088/1742-6596/1883/1/012174.
6. Wu M. Gesture recognition based on deep learning: a review. EAI Endorsed Trans E-Learning. 2024;10:1–8 doi:10.4108/eetel.5191.
7. Halabi M, Harkouss Y. Real-time Arabic sign language recognition system using sensory glove and machine learning. Neural Comput Appl. 2025;37(9):6977–93. doi:10.1007/s00521-025-11010-1.
8. Noor TH, Noor A, Alharbi AF, Faisal A, Alrashidi R, Alsaedi AS, et al. Real-time Arabic sign language recognition using a hybrid deep learning model. Sensors. 2024;24(11):3683. doi:10.3390/s24113683.
9. Dulloo R, Suryawanshi V, Kamble Y, Mishra S, Patil P. A study on real-time sign language identification using KNN. Human Soc Sci. 2024;83:111–20.
10. De Souza C, Pizzolato E. Sign language recognition with support vector machines and hidden conditional random fields: going from fingerspelling to natural articulated words. In: Proceedings of the Lecture Notes in Computer Science; 2013. Vol. 7988, p. 84–98. doi:10.1007/978-3-642-39712-7_7.
11. Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 1110–8.
12. Du Y, Fu Y, Wang L. Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR); 2015 Nov 3–6; Kuala Lumpur, Malaysia. p. 579–83. doi:10.1109/ACPR.2015.7486569.
13. Liao S, Lyons T, Yang W, Schlegel K, Ni H. Logsig-RNN: a novel network for robust and efficient skeleton-based action recognition. arXiv:2110.13008. 2021.
14. Amir Liu J, Ng TT, Wang G. NTU RGB+D: a large-scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 1010–9.
15. Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the Kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6299–308.
16. Liu Y, Cheng M-M, Hu X, Wang K, Bai X. Richer convolutional features for edge detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 5872–81.

17. Feng M, Meunier J. Skeleton graph-neural-network-based human action recognition: a survey. Sensors. 2022;22(6):2091. doi:10.3390/s22062091.
18. Shu Y, Li W, Li D, Gao K, Jie B. Multi-scale dilated attention graph convolutional network for skeleton-based action recognition. In: Pattern Recognition and Computer Vision (PRCV 2023). Singapore: Springer; 2023. p. 16–28. doi:10.1007/978-981-99-8429-9_2.
19. Mohamed A, Hassan N, Jamil AS. Real-time hand gesture recognition: a comprehensive review of techniques, applications, and challenges. Cybern Inf Technol. 2024;24:163–81. doi:10.2478/cait-2024-0031.
20. Podder KK, Chowdhury MEH, Tahir AM, Mahbub ZB, Khandakar A, Hossain MS, et al. Bangla sign language (BdSL) alphabets and numerals classification using a deep learning model. Sensors. 2022;22(2):574 doi:10.3390/s22020574.
21. Podder KK, Tabassum S, Khan LE, Salam KMA, Maruf RI, Ahmed A. Design of a sign language transformer to enable the participation of persons with disabilities in remote healthcare systems for ensuring universal healthcare coverage. In: Proceedings of the 2021 IEEE Technology Engineering Management Conference-Europe (TEMSCON-EUR); 2021 May 17–20; Dubrovnik, Croatia. p. 1–6. doi:10.1109/TEMSCON-EUR.2021.9507222.
22. Podder KK, Chowdhury M, Mahbub Z, Kadir M. Bangla sign language alphabet recognition using transfer learning based convolutional neural network. Bangladesh J Sci Res. 2020;31–33:20–6.
23. Luqman H, El-Alfy ESM. Towards hybrid multimodal manual and non-manual arabic sign language recognition: mArSL database and pilot study. Electronics. 2021;10(14):1739. doi:10.3390/electronics10141739.
24. Hu H, Zhou W, Pu J, Li H. Global-local enhancement network for NMF-aware sign language recognition. ACM Trans Multimedia Comput Commun Appl. 2021;17(3):1–19. doi:10.1145/3436754.
25. Koller O, Ney H, Bowden R. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Trans Pattern Anal Mach Intell. 2019;42(9):2306–20. doi:10.1109/TPAMI.2019.2899040.
26. Camgoz NC, Koller O, Hadfield S, Bowden R. Sign language transformers: joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10023–33. doi:10.1109/CVPR42600.2020.01004.
27. Al-Jarrah O, Halawani A. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artif Intell. 2001;133:117–38. doi:10.1016/S0004-3702(01)00141-7.
28. Assaleh K, Al-Rousan M. Recognition of Arabic sign language alphabet using polynomial classifiers. EURASIP J Adv Signal Process. 2005;2005:507614. doi:10.1155/ASP.2005.507614.
29. Hemayed EE, Hassanien AS. Edge-based recognizer for Arabic sign language alphabet (ArS2V-Arabic sign to voice). In: Proceedings of the 2010 International Computer Engineering Conference (ICENCO); 2010 Dec 27–28; Cairo, Egypt. p. 121–7.
30. Mohandes M, Deriche M, Johar U, Ilyas S. A signer-independent Arabic sign language recognition system using face detection, geometric features, and a Hidden Markov Model. Comput Electr Eng. 2012;38:422–33. doi:10.1016/j.compeleceng.2011.12.002.
31. Elons AS, Elzorkany MM, Elkilany WS, Elhosseini MA. Arabic sign language recognition using leap motion sensor. In: 2014 9th International Conference on Computer Engineering & Systems (ICCES); 2010 Dec 27–28; Cairo, Egypt. p. 368–73. doi:10.1109/ICCES.2014.7030987.
32. Zemgulys J, Raudonis V, Maskeliunas R, Damasevicius R. Recognition of basketball referee signals from real-time videos. J Ambient Intell Humaniz Comput. 2020;11:979–91. doi:10.1007/s12652-019-01365-3.
33. Vaitkevicius A, Taroza M, Blazauskas T, Damasevivius R, Maskeliunas R, Wozniak M. Recognition of American sign language gestures in a virtual reality using leap motion. Appl Sci. 2019;9(3):445. doi:10.3390/app9030445.
34. Krnoul Z, Hruz M, Campr P. Correlation analysis of facial features and sign gestures. In: Proceedings of the IEEE 10th International Conference on Signal Processing; 2010 Oct 24–28; Beijing, China. p. 732–5.
35. Sabyrov A, Mukushev M, Kimmelman V, Imashev A, Koishybay K, Sandygulova A. Towards real-time sign language interpreting robot: evaluation of non-manual components on recognition accuracy. In: Proceedings of the CVPR Workshops; 2019 Jun 16–20; Long Beach, CA, USA. p. 75–82.

36. Yang S, Zhu Q. Continuous Chinese sign language recognition with CNN-LSTM. In: Falco CM, Jiang X, editors. Proceedings of the Ninth International Conference on Digital Image Processing (ICDIP 2017). Bellingham, WA, USA: SPIE; 2017.
37. Elhagry A, Elrayes RG. Egyptian sign language recognition using CNN and LSTM. arXiv:2107.13647. 2021. doi:10.48550/arXiv.2107.13647.
38. Basnin N, Nahar L, Hossain MS. An integrated CNN-LSTM model for bangla lexical sign language recognition. In: Advances in intelligent systems and computing. Singapore: Springer; 2020. p. 695–707. doi:10.1007/978-981-15-0929-3_63.
39. Al-Hammadi M, Muhammad G, Abdul W, Alsulaiman M, Bencherif MA, Mekhtiche MA. Hand gesture recognition for sign language using 3DCNN. IEEE Access. 2020;8:79491–509. doi:10.1109/ACCESS.2020.2990687.
40. Bencherif MA, Algabri M, Mekhtiche MA, Faisal M, Alsulaiman M, Mathkour H, et al. Arabic sign language recognition system using 2D hands and body skeleton data. IEEE Access. 2021;9:59612–27. doi:10.1109/ACCESS.2021.3072821.
41. Boukdir A, Benaddy M, Ellahyani A, Meslouhi OE, Kardouchi M. Isolated video-based Arabic sign language recognition using convolutional and recursive neural networks. Arab J Sci Eng. 2021;47:2187–99. doi:10.1007/s13369-021-05725-2.
42. Boukdir A, Benaddy M, Ellahyani A, Meslouhi OE, Kardouchi M. 3D gesture segmentation for word-level Arabic sign language using large-scale RGB video sequences and autoencoder convolutional networks. Signal Image Video Process. 2022;16:2055–62. doi:10.1007/s11760-022-02144-3.
43. Alyami S, Luqman H, Hammoudeh M. Isolated Arabic sign language recognition using a transformer-based model and landmark keypoints. ACM Trans Asian Low-Resour Lang Inf Process. 2024;23(1):3.
44. Ghimire S, Deo RC, Casillas-Perez D, Salcedo-Sanz S, Sharma E, Ali M. Deep learning CNN-LSTM-MLP hybrid fusion model for feature optimizations and daily solar radiation prediction. Measurement. 2022;202:111759. doi:10.1016/j.measurement.2022.111759.
45. Al-Barham M, Alsharkawi A, Al-Yaman M, Al-Fetyani M, Elnagar A, Abu SaAleek A, et al. RGB arabic alphabets sign language dataset. arXiv:2301.11932. 2023.
46. Zhang H, He H, Chen L, Liu Y. A comprehensive review of skeleton-based action recognition. Pattern Recognit Lett. 2021;144:20–34.
47. Abdul W, Alsulaiman M, Amin SU, Faisal M, Muhammad G, Albogamy FR, et al. Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM. Comput Electr Eng. 2021;95:107395. doi:10.1016/j.compeleceng.2021.107395.
48. Sidig AAI, Luqman H, Mahmoud S, Mohandes M. KArSL: Arabic sign language database. ACM Trans Asian Low-Resour Lang Inf Process. 2021;20(1):1–19. doi:10.1145/3423420.
49. Suliman W, Deriche M, Luqman H, Mohandes M. Arabic sign language recognition using deep machine learning. In: Proceedings of the 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT'21); 2021 Dec 6–8; Alkhobar, Saudi Arabia. p. 1–4. doi:10.1109/isaect53699.2021.9668405.
50. Aly S, Aly W. DeepArSLR: a novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition. IEEE Access. 2020;8:83199–212. doi:10.1109/ACCESS.2020.2990699.