Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190609

Bridging Information Science and Deep Learning: Transformer Models for Isolated Saudi Sign Language Recognition

*Soukeina Elhassen and Lama Al Khuzayem**

Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Received: 2 Jun. 2025, Revised: 2 Aug. 2025, Accepted: 12 Sep. 2025
Published online: 1 Nov. 2025

Abstract: Sign Language (SL) is the primary communication method for deaf and hard-of-hearing individuals. This underscores the need for advanced technologies that bridge the communication gap between SL users and the hearing community. Saudi Sign Language (SSL), the main SL used in Saudi Arabia, lacks large-scale isolated datasets, posing challenges in developing recognition models that perform well on small to medium-sized data. Most state-of-the-art approaches for Arabic Sign Language (ArSL) in general, and SSL in particular, rely on Convolutional Neural Networks (CNNs) architectures, which often struggle to capture long-range temporal dependencies. In contrast, this paper establishes a benchmark for isolated SSL recognition using transformer-based video models. Specifically, we evaluate three state-of-the-art architectures—Swin Transformer, VideoMAE, and TimeSformer—on the King Saud University Arabic Sign Language (KSU-ArSL) dataset. All models are pre-trained on the Kinetics-400 dataset and fine-tuned using 16-frame RGB clips. The Swin Transformer achieved the highest accuracy at 97.50%, followed by VideoMAE at 95.25% and TimeSformer at 93.44%. Despite challenges posed by visually similar signs, these results demonstrate the superior effectiveness of transformer networks over CNNs in sign language recognition. Future work will focus on signer-independent evaluation and continuous SSL recognition to build more generalizable systems and improve accessibility for the Saudi deaf community.

Keywords: Sign Language Recognition, Isolated Sign Language, Saudi Sign Language, Video Transformers, Swin, TimeSformer, VideoMAE, Transfer Learning, KSU-ArSL Dataset.

1 Introduction

Sign Language (SL) serves as the primary means of communication for the deaf and hard-of-hearing community worldwide, highlighting its global significance [1]. In Saudi Arabia, Saudi Sign Language (SSL) is the main SL used by this community, featuring a unique cultural vocabulary, while also sharing some elements with Arabic Sign Language (ArSL) [2]. SSL inherits several linguistic challenges from Arabic, such as multiple terms for a single concept and complex sentence structures, making it a significant area of study [2, 3].

An estimated 720,000 individuals in Saudi Arabia have a hearing disability [3], emphasizing the urgent need for advanced technologies to facilitate communication between the deaf and hearing communities. These technologies are critical for enabling equitable access to services in education, healthcare, and social integration.

Despite SSL's importance, research in this domain remains limited compared to ArSL and other sign languages, resulting in a scarcity of SSL-specific datasets [4, 5]. Most existing resources are image-based datasets, which constrain model performance in capturing the dynamic, continuous nature of sign language gestures in video data [1, 6]. Moreover, many existing solutions rely on Convolutional Neural Networks (CNNs), which are inadequate for modeling long-term temporal dependencies in video-based Sign Language Recognition (SLR) tasks [7, 8].

To address this, our research proposes the use of transformer-based models, which have shown promise in capturing global temporal context and improving video understanding [9, 10, 11]. Specifically, this paper investigates the fine-tuning of three pre-trained video transformer architectures — Swin Transformer [12],

* Corresponding author e-mail: lalkhuzayem@kau.edu.sa

VideoMAE [13], and TimeSformer [14] — on the King Saud University Arabic Sign Language (KSU-ArSL) dataset [3] for isolated sign language recognition. These models leverage self-attention mechanisms to effectively capture spatio-temporal features critical for interpreting complex hand and body gestures. VideoMAE utilizes masked autoencoding for robust feature learning, Swin Transformer employs a hierarchical architecture with shifted window attention for multi-scale representation, and TimeSformer implements divided space-time attention to enhance computational efficiency [12, 13, 14, 15].

Given the low-resource nature of SSL datasets, we adopt transfer learning to mitigate data scarcity, utilizing pre-trained weights from large-scale video datasets. This approach enhances model robustness and generalization [16]. Our results demonstrate that transfer learning yields competitive performance across all models, with distinct variations that offer insights into the most effective transformer architectures for sign language tasks.

The contributions of this paper are as follows: (1) Evaluate the accuracy and robustness of Swin, VideoMAE, and TimeSformer for isolated SSL recognition; (2) Establish a benchmark for the KSU-ArSL dataset to support future research on SSL; (3) Analyze the strengths and limitations of transformer-based models and their implications for improving accessibility for the deaf community in Saudi Arabia.

The remainder of this paper is organized as follows. Section 2 reviews related work on SLR and Section 3 provides a summary of the dataset used. Section 4 details the methodology, including the dataset and model descriptions, Section 5 shows pretraining and fine-tuning strategy, and Section 6 presents the experiments and results and discusses the findings, followed by Section 7 which concludes the paper with future research directions.

2 Related Work

SLR has progressed significantly, evolving from early hand-crafted feature extraction methods — such as Histogram of Oriented Gradients (HOG) [6]— to modern deep learning approaches [7, 18]. Traditional SLR systems relied on manual feature engineering, which was time-consuming and poorly adapted to gesture variability and lighting conditions. The advent of deep learning, particularly CNNs, enabled end-to-end learning directly from video data, significantly improving SLR performance [7, 8].

SLR is typically categorized into isolated SLR, which involves word-level classification, and continuous SLR, which focuses on sentence-level gesture recognition [8, 11]. This study focuses on isolated SLR, which, although simpler in structure, remains challenging due to intra-class variability in signs [18]. CNN-based models, including 3D-CNNs, have achieved promising results on datasets such as American Sign Language (ASL) Lexicon, yet they struggle to capture long-term temporal dependencies required for dynamic gesture modeling [7, 19]. This has led to increasing interest in transformer-based architectures that offer improved spatio-temporal modeling.

In the context of ArSL and SSL recognition, research is growing, driven by the availability of datasets such as KArSL [4] and KSU-ArSL [3]. These datasets capture regional linguistic variation and provide isolated signs for experimentation. Specifically, KArSL includes 502 signs collected from multiple signers, providing a heterogeneous foundation for recognition tasks [4]. Most models applied to ArSL/SSL, however, have employed CNNs or a combination of CNNs with Recurrent Neural Networks (RNNs), which are effective for simple or static signs but insufficient for modeling complex dynamic gestures [9, 20]. For example, CNN models on KArSL performed well for basic gestures but failed to capture temporal relationships in complex sequences [4], emphasizing the need for architectures that model long-range temporal dependencies [15, 21].

KSU-ArSL, which focuses on SL-specific features, remains underutilized with regard to advanced architectures. Transformer-based models, leveraging self-attention to capture spatio-temporal dependencies, have shown success on large-scale datasets like WLASL [19], surpassing CNN performance for isolated sign recognition [21]. These gains are attributed to transformers' ability to model long-term dependencies crucial for dynamic signs [23, 24]. However, transformer applications in SSL remain rare, motivating this work's evaluation of Swin, VideoMAE, and TimeSformer for SSL in low-resource settings.

Transfer learning has become a key strategy in SLR to address data scarcity, particularly in low-resource languages such as SSL [16, 25]. Pre-training on large-scale video datasets like Kinetics-400 provides robust initial weights that can be fine-tuned on smaller SLR datasets [22, 26]. Studies on WLASL and similar corpora have demonstrated that transfer learning significantly improves accuracy by leveraging generalizable video representations [16, 19]. However, limited research has applied pre-trained transformer models to ArSL or SSL datasets [17, 21]. This study aims to bridge this gap by applying Swin, VideoMAE, and TimeSformer to the KSU-ArSL dataset, establishing a performance benchmark for transformer-based models in SSL recognition.

3 Dataset

KSU-ArSL dataset [3] is an isolated SL dataset. This dataset comprises 16,000 videos covering 80 isolated signs, including the Arabic alphabet, numbers, and common daily-use signs, performed by 40 signers with five repetitions each. The signs are categorized into static

Fig. 1: Sample of KSU-ArSL Dataset

signs (e.g., numbers and most letters) and dynamic signs (e.g., words requiring continuous hand movements), as detailed in the ground truth shown in Table 1. The dataset was recorded using Microsoft Kinect V1, Kinect V2, and Sony handheld cameras. The dataset includes RGB, depth, and skeleton data, capturing diverse modalities under varied conditions, such as different lighting, distances (1–2 meters), and signer attire. A sample frame from the dataset, illustrating a signer performing a static sign with annotated hand and body keypoints, is shown in Figure 1. The comprehensive coverage of the KSU-ArSL dataset of 80 signs and diverse recording modalities makes it a robust benchmark for evaluating deep learning models, supporting both signer-dependent and signer-independent experiments.

4 Methodology

This section describes how three transformer-based video architectures (Swin Transformer, VideoMAE, and TimeSformer) are fine-tuned on the KSU-ArSL dataset to perform isolated Saudi Sign Language (SSL) recognition. This can be broken down into four main parts: (1) problem formulation, (2) training objective, (3) transfer learning and preprocessing, (4) transformer architectures, and (5) training setup.

4.1 Problem Formulation

The recognition task is treated as a video classification problem. Each sign language video is uniformly sampled into $T = 16$ frames, where each frame is a 224×224 RGB image:

$$x = \{x_1, x_2, \ldots, x_T\}, \quad x_i \in \mathbb{R}^{224\times224\times3} \tag{1}$$

The model f_θ, parameterized by weights θ, takes this sequence and projects it into a feature space that is then mapped to a probability distribution over $K = 80$ gesture classes:

$$f_\theta : \mathbb{R}^{T\times H\times W\times C} \longrightarrow \Delta^{K-1} \tag{2}$$

where Δ^{K-1} is the probability simplex over 80 classes, and H is the frame's height, W is width, and C is the number of channels. The final predicted class is chosen by the maximum likelihood criterion:

$$\hat{y} = \arg \max_{j\in\{0,\ldots,K-1\}} p_\theta(y_j \mid x) \tag{3}$$

4.2 Training Objective

To train the models, the paper uses the cross-entropy loss function, which penalizes the difference between predicted probabilities and the ground truth class label:

$$L_{\text{CE}} = -\sum_{j=1}^{K} y_j \log \hat{y}_j \tag{4}$$

where y_j is a one-hot encoded true label. To improve generalization and reduce overfitting, an L_2-norm regularization term (weight decay) is added:

$$L_{\text{total}} = L_{\text{CE}} + \lambda \|\theta\|_2^2 \tag{5}$$

The optimization process minimizes this objective over all training samples:

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} L_{\text{total}}\left(f_\theta(x^{(i)}), y^{(i)}\right) \tag{6}$$

This ensures the learned parameters are those that best generalize across the dataset.

4.3 Transfer Learning and Preprocessing

To address the low-resource nature of SSL datasets, all models were pretrained on the large-scale Kinetics-400 action recognition dataset (240,000 clips) to acquire robust spatiotemporal representations before fine-tuning on KSU-ArSL. During adaptation, the early layers responsible for generic motion feature extraction were frozen, while higher layers including the classification head were retrained to capture SSL-specific patterns. Preprocessing steps ensured data consistency by normalizing videos to 16 frames per second, resizing frames to 224×224, and applying augmentation techniques such as random cropping, horizontal flipping with a probability of 0.5, and temporal subsampling. The dataset was split into 80% training, 10% validation, and 10% testing, a configuration that balances learning and evaluation while mitigating overfitting and enhancing model generalization across different signers, environments, and lighting conditions.

Table 1: Sample of ground truth classes for the KSU-ArSL dataset [3]

Sign Type	Category Type	Classes
Static	Numbers (11)	0, 1, 2, 3, ..., 8, 9, 10
Static	Letters (28)	Alf, Ba, Taah, Daah, Gim, Haa, ..., Waw, Ya, Ha
Static	Common Words (13)	Father, Feel, Hospital, King, Sorry, ..., University, Where
Dynamic	Common Words (28)	Alslam Alikom, Arabic Language, ..., Sign Language, Evening, Morning

4.4 Transformer Architectures

The three transformer architectures contribute uniquely to SSL recognition by capturing spatiotemporal features in different ways.

Swin Transformer: Employs Shifted Window Multi-Head Self-Attention (SW-MSA), where attention is computed locally within windows of size M, reducing computational complexity as shown in Equation (7):

$$\mathcal{O}(N^2 d) \quad \rightarrow \quad \mathcal{O}(M^2 d) \tag{7}$$

thus enabling efficient learning of both fine-grained hand shapes and broader gesture context.

VideoMAE: Adopts a masked autoencoding pretraining strategy in which approximately 75% of video patches are randomly masked, forcing the model to reconstruct missing patches:

$$x_m[i] = \begin{cases} x_e[i], & i \in M \\ 0, & i \notin M \end{cases} \tag{8}$$

This encourages the capture of robust spatiotemporal dependencies. During fine-tuning, the decoder is replaced with a classification head.

TimeSformer: Introduces divided space-time attention by applying spatial attention within individual frames and temporal attention across frames:

$$A_{\text{spatial}} = \text{Softmax}\left(\frac{Q_s K_s^\top}{d_h}\right) V_s \tag{9}$$

$$A_{\text{temporal}} = \text{Softmax}\left(\frac{Q_t K_t^\top}{d_h}\right) V_t \tag{10}$$

which reduces computational cost as shown in Equation (11):

$$\mathcal{O}((THW)^2) \quad \rightarrow \quad \mathcal{O}(T^2 HW + TH^2 W) \tag{11}$$

Collectively, these designs allow the models to efficiently balance local feature extraction with long-range temporal modeling, making them well-suited for sign language recognition.

4.5 Training Setup

All models are fine-tuned on the KSU-ArSL dataset using a transfer learning optimization strategy with hyperparameters carefully selected for stability and generalization. The learning rate is initialized at $\eta_{\max} = 3 \times 10^{-5}$, with a batch size of 4 due to GPU memory limitations, and a weight decay parameter $\lambda = 0.01$ is applied to penalize large weights and improve generalization. The total loss, as defined in Equation (5), combines cross-entropy with weight decay where L_{CE} is the cross-entropy loss and $\|\theta\|_2^2$ is the squared Euclidean norm of the parameters.

A cosine annealing learning rate scheduler with warmup is employed to gradually adjust the step size during training. For iteration t, the learning rate is computed as:

$$\eta_t = \eta_{\min} + \tfrac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\frac{(t - T_{\text{warmup}})\pi}{T - T_{\text{warmup}}}\right)\right] \tag{12}$$

where $\eta_{\min}$ is the minimum learning rate, $\eta_{\max}$ is the peak learning rate, T is the total number of training steps, and T_{warmup} is the number of warmup steps. This schedule ensures that the learning rate starts small, gradually increases during warmup to stabilize convergence, and then decreases smoothly, preventing overshooting in later epochs.

To enhance training efficiency and model robustness, early stopping is applied: if the validation loss does not improve for three consecutive epochs, training is halted. Formally, if

$$L_{\text{val}}(e) \geq \min_{j<e} L_{\text{val}}(j) \quad \text{for three consecutive epochs } e \tag{13}$$

then optimization is stopped and the best-performing model checkpoint is retained. This prevents unnecessary computation and mitigates overfitting to the training set.

In addition, gradient-based optimization is carried out with the AdamW optimizer, which decouples weight

decay from the gradient update. For model parameters θ, the update at step t is:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \eta_t \lambda \theta_t \tag{14}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates of the gradients, and ε is a small constant for numerical stability.

This training setup—combining transfer learning, regularization, adaptive learning rate scheduling, and early stopping—ensures stable convergence, efficient use of computational resources, and strong generalization to unseen signers and signing conditions.

5 Pretraining and Fine-Tuning Strategy

This section outlines the complete pipeline used to adapt transformer-based video models for isolated SSL recognition. The process begins with dataset preprocessing to standardize inputs and apply augmentation for robustness. Next, models are pretrained on the large-scale Kinetics-400 dataset to learn transferable spatiotemporal representations. Finally, the pretrained models are fine-tuned on the KSU-ArSL dataset, adapting them to SSL-specific gesture dynamics.

The following subsections describe each stage in detail, along with model-specific architectural considerations.

5.1 Dataset Preprocessing

The KSU-ArSL dataset contains videos of varying lengths and signer styles, requiring preprocessing for consistency and robustness. Each video is resampled to a fixed frame rate of 16 FPS and resized to 224×224 pixels. For a video of duration D, the total number of sampled frames is

$$T = \lfloor 16 \cdot D \rfloor \tag{15}$$

from which exactly 16 frames are uniformly selected, as defined in Equation (1). The dataset is split into 80% training (12,800 clips), 10% validation (1,600 clips), and 10% testing (1,600 clips), ensuring generalization to unseen data. Augmentation techniques are applied, including random cropping, horizontal flipping with probability $p = 0.5$, and temporal subsampling, to enhance variability and prevent overfitting.

Each frame is also normalized per channel:

$$x'_{i,j,c} = \frac{x_{i,j,c} - \mu_c}{\sigma_c} \tag{16}$$

where μ_c and σ_c denote the mean and standard deviation of RGB channels.

These preprocessing steps expose the model to diverse visual variations, enabling improved classification accuracy and robustness during testing.

5.2 Pretraining Phase

To address the scarcity of annotated SSL data, all models are pretrained on Kinetics-400, a large-scale dataset containing 240,000 video clips across 400 human action classes. This pretraining step provides strong spatiotemporal representations transferable to SSL recognition.

Formally, given a video input $x \in \mathbb{R}^{T \times H \times W \times C}$ and its corresponding action label $y \in \{1, \ldots, 400\}$, the pretraining objective is defined using the cross-entropy loss:

$$L_{\text{pretrain}} = -\sum_{j=1}^{400} y_j \log \hat{y}_j \tag{17}$$

where y_j is the one-hot encoded ground-truth label, and $\hat{y}_j$ is the predicted probability for class j.

Through minimizing L_{pretrain}, the model learns generalizable spatiotemporal features such as motion patterns and contextual dependencies, which can later be adapted to the SSL domain via fine-tuning.

5.2.1 Swin Transformer

Swin Transformer (Shifted Window) is pretrained in a supervised manner using 16-frame video clips divided into non-overlapping 16×16 patches, producing 3,136 tokens embedded in a 96-dimensional space:

$$x_p = \text{PatchEmbed}(x), \quad x_p \in \mathbb{R}^{3136 \times 96} \tag{18}$$

Its hierarchical structure comprises four stages with feature dimensions (96, 192, 384, 768), and attention is computed locally in 7×7 windows. Shifted windows capture cross-window dependencies efficiently. The attention mechanism is expressed as:

$$\text{SW-MSA}(Q,K,V) = \text{Softmax}\left(\frac{QK^\top}{d_h} + B\right)V \tag{19}$$

5.2.2 VideoMAE

VideoMAE is pretrained in a self-supervised manner using masked autoencoding, where approximately 75% of video patches are masked ($m = 0.75$). Let $M \subset \{1, \ldots, N\}$ denote the indices of unmasked patches, with $|M| = (1-m)N$. The masked input is defined in Equation (8).

The encoder processes unmasked tokens $X \in \mathbb{R}^{|M| \times d}$ through $L = 12$ transformer layers with multi-head self-attention:

$$\text{MHSA}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{20}$$

$$\text{head}_i = \text{Softmax}\left(\frac{Q_i K_i^\top}{d_h}\right)V_i \tag{21}$$

A lightweight decoder reconstructs masked patches, encouraging robust spatiotemporal feature learning.

5.2.3 TimeSformer

TimeSformer adopts supervised pretraining with divided space-time attention. Spatial attention is first computed frame-wise, followed by temporal attention across frames as shown in Equations (9) and (10).

The model contains 12 transformer layers with 768 hidden dimensions and 12 heads, enabling it to capture long-range temporal dependencies critical for dynamic sign gestures.

5.3 Fine-Tuning Phase

After pretraining, all models are fine-tuned on KSU-ArSL to classify 80 isolated signs. Fine-tuning uses a learning rate of 3×10^{-5}, batch size 4, weight decay $\lambda = 0.01$, and a warmup ratio of 0.1. The total loss combines cross-entropy with regularization:

$$L_{\text{total}} = L_{\text{CE}} + \lambda \sum_i w_i^2 \tag{22}$$

where L_{CE} is categorical cross-entropy. Training employs cosine annealing for learning rate scheduling as defined in Equation (12).

Early stopping halts training after three epochs without validation improvement, ensuring efficiency and preventing overfitting. Training is conducted on the Aziz supercomputer of King Abdulaziz University, leveraging parallel GPU resources.

5.3.1 Swin Transformer

The fine-tuned Swin Transformer averages frame-level features temporally before feeding them into a linear classifier, outputting logits over 80 classes. Its hierarchical design enables effective extraction of both local (hand shapes) and global (gesture context) patterns. Swin Transformer Architecture is fully explained in Figure 2.

5.3.2 VideoMAE

For fine-tuning, the pretraining decoder of VideoMAE is replaced by a classification head. Each frame is divided into 196 patches, resulting in 3,136 patches across 16 frames, which are embedded and passed through the transformer encoder. VideoMAE Architecture is illustrated in Figure 3. Outputs are averaged:

$$z_{\text{pool}} = \frac{1}{N} \sum_{i=1}^{N} Z_i \tag{23}$$

and mapped to class logits via a linear layer:

$$\text{logits} = W_c z_{\text{pool}} + b_c, \quad W_c \in \mathbb{R}^{K \times d} \tag{24}$$

The predicted class is:

$$\hat{y} = \arg \max_{j \in \{0,\ldots,79\}} p(y_j \mid x) \tag{25}$$

with training objective:

$$L_{\text{CE}} = -\sum_{j=0}^{K-1} y_j \log \left(\text{Softmax}(\text{logits})_j\right) \tag{26}$$

5.3.3 TimeSformer

TimeSformer processes each video through divided attention to model spatial and temporal dependencies, then pools frame-level features into a linear classification head. Pretrained for 400 actions on Kinetics-400, it is fine-tuned for 80 SSL signs, producing logits for classification as shown in Figure 4.

5.4 Model Comparison

All three models share a common pipeline of dataset preprocessing, Kinetics-400 pretraining, and KSU-ArSL fine-tuning, but their architectural differences influence their SSL performance. Swin Transformer's hierarchical attention excels in multi-scale feature extraction, making it well-suited for complex gestures. VideoMAE, with its self-supervised masked autoencoding, demonstrates strong robustness to noise and signer variability, improving generalization under real-world conditions. TimeSformer, with divided space-time attention, captures long-range temporal dependencies, making it particularly effective for dynamic gestures.

Table 2 summarizes the comparative strengths of the models: Swin is best for complex gestures, VideoMAE for noisy environments, and TimeSformer for temporal modeling. Together, they provide a comprehensive benchmark for transformer-based SSL recognition.

6 Results and Discussion

The transformer-based models were evaluated on the KSU-ArSL test subset using multiple performance metrics, including overall accuracy, precision, recall, F1-score, and class-wise results derived from the confusion matrix. These metrics are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{28}$$

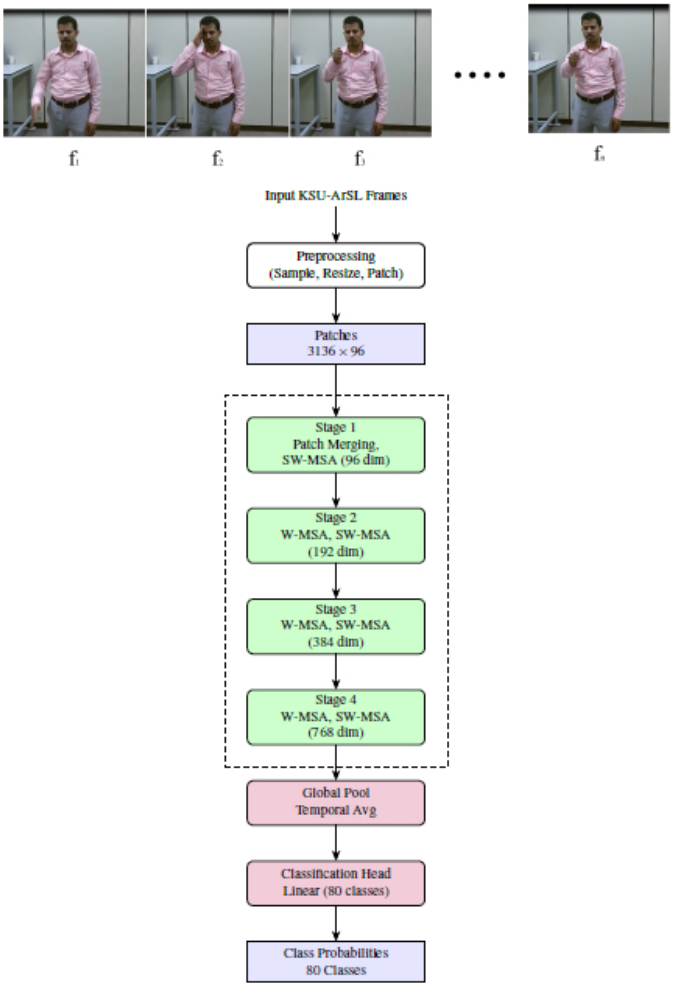


Fig. 2: Swin Transformer Architecture for Isolated SLR. The pipeline processes KSU-ArSL video frames through hierarchical stages with SW-MSA, followed by temporal averaging and classification.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{29}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{30}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

This evaluation framework enables not only overall performance assessment but also identification of class-specific weaknesses and generalization ability across signer variation. Precision and recall scores ranged from 85%–100% across most signs, though certain gestures such as “fa” and “kha” showed reduced accuracy (62–65%), likely due to limited training samples and strong visual similarity with other signs. Table 3 shows accuracy comparison per class on the three models.

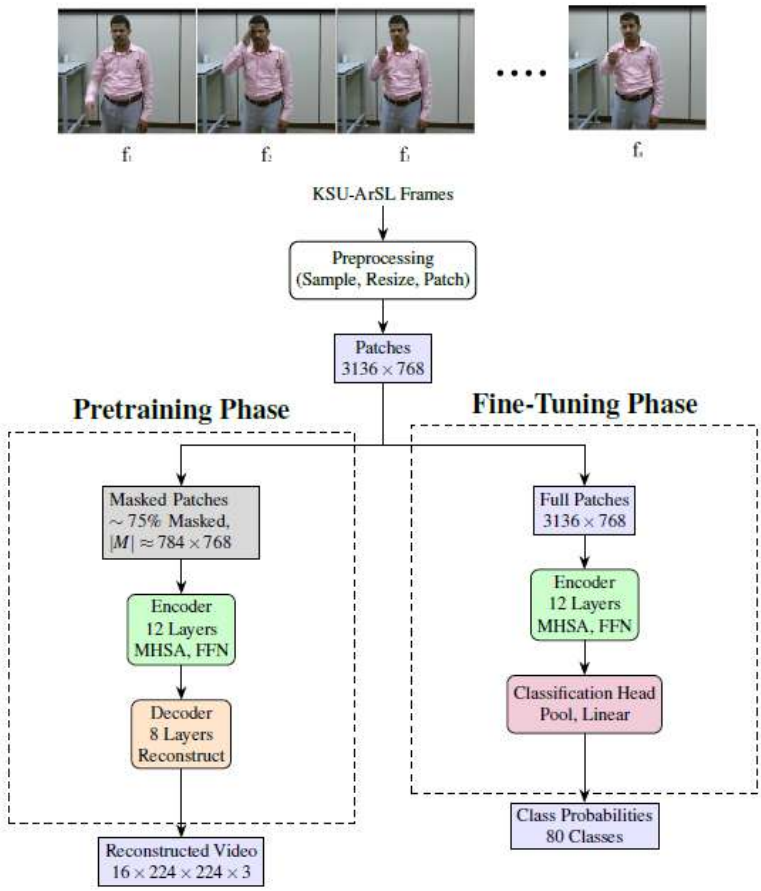


Fig. 3: VideoMAE Architecture for Isolated SL Recognition. The pipeline starts with KSU-ArSL dataset frames, followed by preprocessing. *Pretraining* uses masked autoencoding to reconstruct videos, while *fine-tuning* classifies 80 isolated SL gestures.

6.1 Model Performance

Among the three transformer models, **Swin Transformer** achieved the highest performance, reaching **97.50% accuracy**, 97.69% precision, and 97.50% recall. Its hierarchical architecture and Shifted Window Multi-Head Self-Attention (SW-MSA) enabled strong multi-scale representation learning, excelling at both static and dynamic gestures. Table 4 summarizes the performance of Swin, VideoMAE, and TimeSformer.

VideoMAE followed with 95.25% accuracy, benefiting from its masked autoencoding strategy, which enhances robustness to signer variation (across 40 signers in the dataset) and noisy frames, though at higher computational cost.

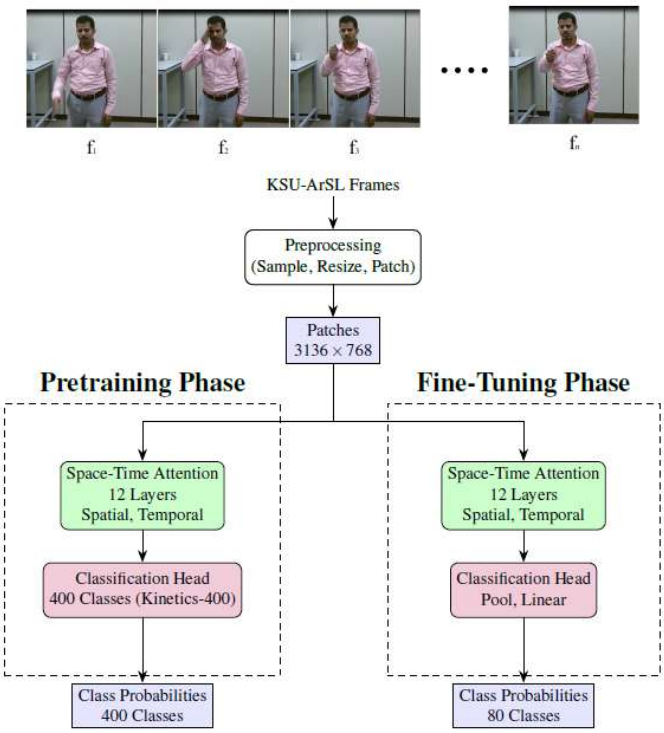


Fig. 4: TimeSformer Architecture

TimeSformer achieved 93.44% accuracy, leveraging divided space-time attention to capture long-range dependencies efficiently, though it performed slightly worse on static signs compared to Swin.

Training and validation loss curves, as depicted in Figure 5, showed smooth convergence across models, confirming stable optimization with minimal overfitting. This validates the effectiveness of the pretraining–fine-tuning strategy, where pretrained spatiotemporal representations from Kinetics-400 were successfully adapted to SSL recognition.

6.2 Comparison with Prior Work

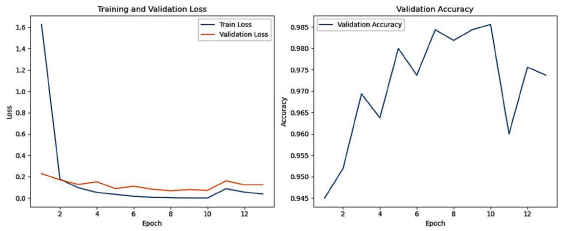
Table 5 presents a comparison between transformer-based approaches and previous CNN-based baselines on the KSU-ArSL dataset. Swin Transformer achieved 97.50% accuracy, surpassing Al Khuzayem *et al.* [27]'s CNN-BiLSTM (94.46%) and Bencherif *et al.* [3]'s 3D CNN + Point CNN (89.62%), both of which were more sensitive to signer dependency and gesture variability.

VideoMAE also outperformed CNN-based models, demonstrating the effectiveness of self-supervised pretraining in low-resource sign language tasks. TimeSformer, while slightly lower in accuracy, still exceeded prior CNN-based approaches due to its efficient

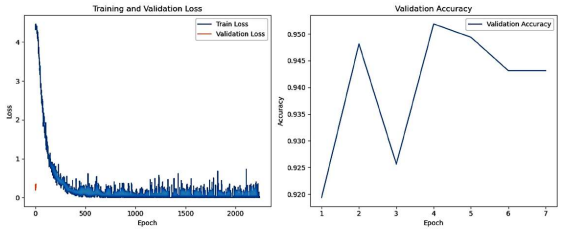
Table 2: Comparison of Models for Isolated SL Recognition

	Swin Transformer	VideoMAE	TimeSformer
Architecture	Hierarchical, 4 stages (96, 192, 384, 768 dim)	Encoder-decoder, 12 layers (768 dim)	12 layers (768 dim, 12 heads)
Pretraining Strategy	Supervised (Kinetics-400)	Self-supervised (Masked Autoencoding)	Supervised (Kinetics-400)
Attention Mechanism	Shifted Window (SW-MSA)	Multi-Head Self-Attention (MHSA)	Divided Space-Time Attention
SL Suitability	Multi-scale feature extraction for complex gestures	Noise-robust generalization for real-world SL	Long-range temporal dependencies for dynamic gestures

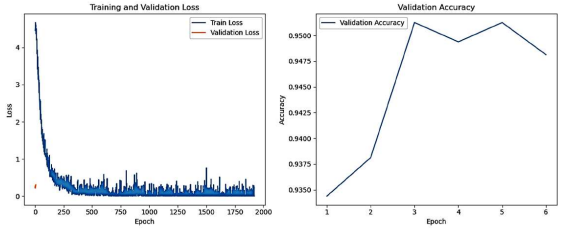
modeling of temporal dependencies. Figure 5 illustrates the training and validation loss curves for all three models, confirming convergence and stable learning dynamics across epochs.



(a) Swin Transformer model



(b) VideoMAE Transformer model



(c) TimeSformer Transformer model

Fig. 5: Training and validation loss curves for Swin, VideoMAE, and TimeSformer on KSU-ArSL.

These findings confirm that transformer architectures outperform traditional CNN-based methods for isolated sign language recognition. The success of Swin Transformer is attributed to its hierarchical design and supervised pretraining on large-scale video datasets, while VideoMAE benefits from masked autoencoding, which improves resilience to occlusion and signer variability. In contrast, CNN-based approaches lack the ability to model long-range temporal dependencies and complex spatial features as effectively as transformers.

6.3 Discussion

The results highlight that transformer-based SSL recognition models generalize well across unseen signers and environments, making them strong candidates for real-world deployment. However, class imbalance and visual similarity remain challenges, as shown by reduced performance on certain underrepresented or visually overlapping signs. Addressing these limitations may involve class-balancing strategies, focal loss functions, or synthetic data augmentation to strengthen representation of difficult signs.

Looking forward, we plan to conduct signer-independent evaluations to further test generalizability, and to explore cross-lingual transfer learning from other sign language datasets such as ASL, British Sign Language (BSL), and Chinese Sign Language (CSL), leveraging shared spatiotemporal features across sign languages. Another promising avenue is to investigate hybrid architectures combining CNNs and RNNs with transformers model, to capture complementary strengths: transformers for global dependencies and BiLSTMs for fine-grained temporal modeling. Ultimately, these advances aim to contribute toward real-time, robust SSL recognition systems, supporting accessible communication technologies for the deaf and hard-of-hearing communities.

7 Conclusion and Future Work

This study establishes a benchmark for isolated Saudi Sign Language (SSL) recognition using three transformer-based architectures (Swin Transformer, VideoMAE, and TimeSformer) applied to the KSU-ArSL dataset. By adopting a transfer learning approach, the models were able to deliver high recognition accuracy despite the limited availability of SSL-specific training data. The Swin Transformer emerged as the strongest performer, achieving **97.50% accuracy**, followed by VideoMAE at 95.25%, and TimeSformer at 93.44%. These results demonstrate the advantage of transformer architectures in modeling complex spatiotemporal patterns and underline the importance of pretraining on

Table 3: Per-class Accuracy Comparison of Swin Transformer, VideoMAE, and TimeSformer Models on the KSU-ArSL Dataset

Class	Swin Transformer	VideoMAE	TimeSformer
0	80%	90%	70%
1	100%	95%	90%
10	100%	85%	75%
2	100%	90%	100%
3	90%	80%	100%
4	95%	100%	80%
5	100%	95%	85%
6	100%	95%	100%
7	100%	90%	95%
8	95%	90%	90%
9	95%	85%	85%
ain	100%	100%	100%
alf	100%	100%	100%
alslam-aliukom	100%	100%	100%
arabic-language	100%	80%	100%
ba	100%	90%	95%
brother	95%	95%	100%
cold	100%	100%	100%
come-in	100%	100%	100%
daah	95%	95%	85%
dal	95%	100%	90%
deaf	100%	100%	100%
death	95%	95%	100%
doctor	95%	100%	100%
english-language	95%	100%	100%
evening	95%	100%	65%
fa	95%	90%	55%
family	100%	100%	100%
father	100%	100%	100%
feel	100%	100%	95%
file	100%	100%	100%
gim	100%	90%	100%
gin	100%	95%	90%
ha	95%	80%	60%
haa	100%	95%	100%
hello	100%	100%	100%
hospital	100%	100%	100%
hot	100%	100%	100%
how-are-you	95%	100%	85%
job	80%	100%	100%
kaf	95%	100%	90%
kha	90%	65%	80%
king	100%	100%	90%
lam	100%	100%	100%
manager	95%	100%	100%
medication	100%	100%	100%
meeting	100%	100%	100%
mem	95%	95%	100%
morning	95%	100%	100%
mosque	100%	100%	100%
mother	100%	100%	95%
name	95%	100%	95%
non	100%	100%	65%
pain	95%	100%	100%
pharmacy	100%	100%	100%
prayer	95%	100%	100%
qaf	100%	65%	100%
ra	100%	95%	85%
reason	100%	100%	100%
sad	100%	65%	75%
saud	100%	100%	100%
shin	100%	100%	100%
sign-language	100%	100%	100%
sin	95%	90%	95%
sister	95%	100%	100%
sorry	100%	100%	100%
surgery	100%	100%	100%
ta	100%	90%	95%
taah	95%	80%	75%
tha	100%	100%	90%
thad	100%	100%	95%
thal	95%	100%	100%
thank	100%	100%	100%
tired	95%	95%	85%
university	95%	100%	95%
vacation	100%	100%	100%
waw	100%	95%	100%
where	100%	95%	100%
ya	100%	100%	95%
zai	90%	90%	85%

large-scale action datasets to address the challenges of data scarcity in sign language research.

Beyond their technical performance, these findings represent an important step toward advancing accessible communication technologies for the deaf and hard-of-hearing community in Saudi Arabia. Automatic SSL recognition systems can be integrated into real-time translation tools, inclusive educational platforms, healthcare services, and public service interfaces, reducing barriers between signers and the wider hearing population. In educational contexts, such systems can support bilingual learning environments and provide greater access to instructional content. Likewise, integration into broadcasting and public communication systems can improve accessibility through automated captioning and sign-to-text translation. These contributions align closely with the United Nations **Sustainable Development Goal 10 (Reduced Inequalities)**, reinforcing the societal relevance of this research in promoting inclusivity and equal access to information.

Table 4: Comparison of Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)
Swin	97.50	97.69	97.50
VideoMAE	95.25	95.71	95.25
TimeSformer	93.44	94.34	93.44

Looking ahead, future research will focus on several directions. First, extending recognition from isolated signs to continuous signing remains a critical challenge, as it involves handling gesture coarticulation and segmentation. Second, there is a growing need to optimize transformer architectures for real-time and resource-constrained environments, particularly for deployment on mobile and wearable devices. Third, multimodal fusion approaches, which combine visual input with skeletal, depth, or motion data, offer a promising path to enhanced robustness. Finally, cross-lingual transfer learning, leveraging resources from sign languages such as ASL, BSL, and CSL, may provide further improvements in generalization and adaptability. Collectively, these avenues will not only refine SSL recognition systems but also contribute to building practical, scalable, and inclusive technologies that empower diverse signing populations.

Data Availability Statement

The KSU-ArSL dataset was obtained from King Saud University with authors' permission. Interested researchers may contact Mansour Alsulaiman (msuliman@ksu.edu.sa) regarding data access.

Acknowledgment

The authors acknowledge the support given by King Saud University in providing them with the KSU-ArSL dataset.

Table 5: Accuracy Comparison of Models for Isolated Saudi Sign Language Recognition on KSU-ArSL Dataset

Approach	Architecture	Pretraining Strategy	Accuracy (%)
Bencherif et al. [3]	3D CNN + Point CNN	Supervised learning	89.62
Al Khuzayem et al. [27]	CNN-BiLSTM	Supervised learning with augmentation	94.46
TimeSformer (Ours)	Transformer	Supervised video classification	93.44
VideoMAE (Ours)	Transformer	Self-supervised video reconstruction	95.25
Swin Transformer (Ours)	Transformer	Supervised video classification	97.50

References

[1] A. Voskou, K. Papoutsakis, and P. Daras, "Deep learning approaches for sign language recognition: A review," *Proc. Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3428–3437.

[2] Alzohairi R., Alghonaim, R., Alshehri W., and Aloqeely S., "Image based Arabic sign language recognition system" *nternational Journal of Advanced Computer Science and Applications*, vol. 9(3) , 2018.

[3] M. A. Bencherif, M. Algabri, M. A. Mekhtiche, M. Faisal, M. Alsulaiman, H. Mathkour, and H. Ghaleb, "Arabic sign language recognition system using 2D hands and body skeleton data," *IEEE Access*, vol. 9, pp. 59612–59627, 2021.

[4] Sidig, A. A. I., Luqman, H., Mahmoud, S., and Mohandes, M., "KArSL: Arabic sign language database," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1), 1-19, 2021.

[5] Madhiarasan, M., and Partha Pratim Roy, "A comprehensive review of sign language recognition: Different types, modalities, and datasets," *arXiv preprint arXiv*, 2022, 2204.03328.

[6] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, p. 113794, 2021.

[7] Kudrinko, K., Flavin, E., Zhu, X., and Li, Q, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, 2020, p. 82-97.

[8] Khan, A., Jin, S., Lee, G. H., Arzu, G. E., Nguyen, T. N., Dang, L. M., ... and Moon, H., "Deep learning approaches for continuous sign language recognition: A comprehensive review," *IEEE Access*, 2025.

[9] S. Al Ahmadi, F. Mohammad, and H. Al Dawsari, "Efficient YOLO-based deep learning model for Arabic sign language recognition," *Journal of Disability Research*, vol. 3, no. 4, p. 20240051, 2024.

[10] Shabaninia, E., Nezamabadi-pour, H., Shafizadegan, F., "Transformers in action recognition: A review on temporal modeling," *arXiv preprint*, arXiv:2302.01921., 2022.

[11] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10023–10033.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[13] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 10078–10093.

[14] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 813–824.

[15] Ulhaq, A., Akhtar, N., Pogrebna, G., and Mian, A., "Vision transformers for action recognition: A survey," *arXiv preprint arXiv:2209.05700*, 2022.

[16] Holmes, R., Rushe, E., De Coster, M., Bonnaerens, M., Satoh, S. I., Sugimoto, A., and Ventresque, A, "From scarcity to understanding: Transfer learning for the extremely low resource irish sign language,"*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2008-2017.

[17] Alnabih, A. F., Maghari, A. Y., "Arabic sign language letters recognition using Vision Transformer," *Multimedia Tools and Applications*, 2024, 81725-81739.

[18] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. Papadopoulos, D. Kosmopoulos, P. Daras, et al., "A comprehensive study on sign language recognition methods," *arXiv preprint arXiv:2107.11427*, 2021.

[19] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 1459–1469.

[20] M. Al-Hammadi, G. Muhammad, W. Abdul, and M. Alsulaiman, "Arabic sign language recognition using convolutional neural networks," *Proc. Int. Conf. Comput. Vis. Workshops*, 2020, pp. 3441–3449.

[21] Saproo, V., Aggarwal, R. K., " A Transformer Based Indian Signed Language Recognition," *In 2024 First International Conference on Pioneering Developments*

in Computer Science Digital Technologies (IC2SDT), 2024, pp. 170-174.
[22] H. R. V. Joze and O. Koller, "MS-ASL: A large-scale dataset and benchmark for understanding American sign language," *arXiv preprint arXiv:2012.01035*, 2020.
[23] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 6836–6846.
[24] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 6824–6835.
[25] H. Hu, W. Zhao, W. Zhou, and H. Li, "SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11221–11239, 2023.
[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[27] L. Al Khuzayem, S. Shafi, S. Aljahdali, R. Alkhamesie, and O. Alzamzami, "Efhamni: A deep learning-based Saudi sign language recognition application," *Sensors*, vol. 24, no. 10, p. 3112, 2024.

Soukeina Elhassen is currently pursuing her Master's degree in computer science at King Abdulaziz University, Jeddah, Saudi Arabia. She received her Bachelors degree in computer science from Umm Al- Qura University, Makkah, Saudi Arabia, in 2019 with first-class honors. Her research focuses on machine learning and computer vision, particularly in sign language recognition. She has worked as a quality management expert in virtual reality applications.

Lama Al Khuzayem is an Assistant Professor in the Computer Science Department at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Dr. Al Khuzayem's research interests include data integration, artificial intelligence, deep learning, semantic web, and computer vision.