Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 9 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i09.60875

PAPER

Self-Attention-Based VGG16 Approach for Sign Language Gesture Recognition in Inclusive Education

Naoufal El-Marzouki(✉), Imane Lasri, Fatim Ezzahrae Dorhmi, Anouar Riadsolh, Mourad Elbelkacemi

Mohammed V University in Rabat, Rabat, Morocco

naoufal_elmarzouki@um5.ac.ma

ABSTRACT

Sign language recognition (SLR) plays an important role in improving communication between deaf or hard-of-hearing individuals and the hearing community and has promising applications in inclusive education. However, conventional convolutional neural network (CNN)-based models mainly focus on local feature extraction and may fail to effectively capture global spatial dependencies, especially when recognizing visually similar static gestures. To address this limitation, this paper proposes a self-attention-based VGG16 framework for static sign language gesture recognition of alphabetic (A–Z) and numeric (0–9) hand signs. The proposed approach integrates a multi-head self-attention (MHSA) mechanism into a pre-trained VGG16 architecture in order to enhance global feature representation while preserving effective local feature extraction. Experiments conducted on a 37-class static sign language dataset show that the proposed model outperforms the baseline VGG16 architecture, achieving a test accuracy of 99.87%. The obtained results confirm the effectiveness of self-attention in improving recognition performance and prediction stability for static sign language gestures. These findings suggest that the proposed framework can serve as a promising building block for intelligent assistive technologies supporting inclusive educational environments.

KEYWORDS

sign language recognition (SLR), self-attention, VGG16, inclusive education

1 INTRODUCTION

Sign language recognition (SLR) has attracted significant research interest over the past decades as a key technology for improving communication between deaf or hard-of-hearing individuals and the hearing community. According to the World Health Organization (WHO), more than 430 million people worldwide live with disabling hearing loss, a number expected to rise to 700 million by 2050 [1]. Sign languages are natural and fully developed languages with their own grammatical structures, relying primarily on hand gestures, finger movements, and spatial configurations to convey meaning. However, the limited proficiency in sign language

El-Marzouki, N., Lasri, I., Dorhmi, F. E., Riadsolh, A., Elbelkacemi, M. (2026). Self-Attention-Based VGG16 Approach for Sign Language Gesture Recognition in Inclusive Education. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(9), pp. 172–185. https://doi.org/10.3991/ijoe.v22i09.60875

Article submitted 2026-02-03. Revision uploaded 2026-04-26. Final acceptance 2026-04-27.

among hearing individuals creates a major communication barrier that restricts social interaction, education, and access to essential services.

This challenge becomes even more critical in the context of inclusive education, where effective communication is fundamental for ensuring equal learning opportunities. Studies indicate that over 90% of deaf children are born to hearing parents who do not master sign language, which often leads to early communication gaps and learning difficulties [2, 3]. In addition, many educational institutions suffer from a shortage of qualified sign language interpreters, particularly in developing regions, which reduces the effective participation of deaf students in classroom activities [4]. These limitations hinder academic integration and highlight the urgent need for technological solutions capable of supporting inclusive and accessible educational environments.

Automatic SLR systems aim to address this challenge by translating visual sign language gestures into textual or spoken representations. Such systems can facilitate communication in multiple domains, including education, healthcare, and human–computer interaction. Early SLR approaches mainly relied on handcrafted features and conventional machine learning methods. More recently, however, deep learning–based models have become dominant because of their ability to automatically learn discriminative representations from visual data. For instance, Saleh and Issa [5] demonstrated that fine-tuned deep neural networks can effectively improve Arabic SLR performance, while more recent CNN-based image recognition strategies have also shown promising results for static hand gesture classification [6].

In one of the earliest comprehensive reviews, Safeel et al. [7] provided a systematic analysis of SLR techniques, categorizing recognition methods into image-based approaches with or without gloves and describing the complete recognition pipeline, including segmentation, feature extraction, dimensionality reduction, and classification. Classical models such as Hidden Markov Models (HMMs), k-NN, Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) were discussed, alongside the emerging role of convolutional neural networks (CNNs). Their survey established a useful foundation for understanding the strengths and limitations of different SLR methodologies.

Beyond algorithmic performance, the practical success of SLR systems also depends on usability and user acceptance. Kurt and Erden [8] investigated the perceptions of pre-service special education teachers toward assistive technologies, including SLR systems. Their findings emphasized that trust, confidence, and ease of use are crucial factors in the adoption of such technologies in educational settings, thereby underlining the importance of robust and reliable recognition systems.

Recent studies have confirmed the effectiveness of deep learning methods [9], particularly CNN-based architectures, in improving SLR performance. El-Marzouki et al. (2024) [10] proposed a CNN-based framework for American Sign Language recognition and demonstrated the capacity of convolutional layers to extract discriminative spatial features. Similarly, Reddy et al. [11] introduced a real-time SLR system combining MediaPipe-based hand landmark extraction with a Random Forest classifier, highlighting the practicality of lightweight and accessible recognition pipelines.

Despite these advances, an important limitation remains: most CNN-based SLR models primarily focus on local feature extraction and do not explicitly model global spatial dependencies between different hand regions. This issue becomes particularly challenging when recognizing visually similar static alphabetic and numeric gestures, where subtle differences in global hand configuration may determine the target class. Therefore, the research problem addressed in this work is how to improve the representation of global contextual relationships in static sign language images while preserving the strong local feature extraction capability of convolution neural networks.

To address this problem, this paper proposes a Self-Attention-Based VGG16 approach for sign language gesture recognition, focusing on static alphabetic and numeric hand gestures. The proposed model integrates a self-attention module into a pre-trained VGG16 [12] backbone in order to enhance global feature representation and improve class discrimination. The main objective of this work is to investigate whether the inclusion of self-attention can improve the recognition performance and stability of a CNN-based static SLR system compared with a fine-tuned VGG16 baseline.

The main contributions of this work are as follows:

- the integration of a self-attention mechanism into the VGG16 architecture for static sign language gesture recognition;
- a comparative experimental evaluation between the baseline VGG16 model and the proposed self-attention-enhanced model;
- an in-depth analysis of training dynamics, confusion matrices, and performance gains for alphabetic and numeric gesture classification.

The remainder of this paper is organized as follows: Section 2 presents the proposed methodology for static sign language gesture recognition, including dataset preparation, the baseline VGG16 architecture, and the integration of the self-attention mechanism. Section 3 reports and discusses the experimental results, providing a comparative analysis between the baseline model and the attention-enhanced approach. Finally, Section 4 concludes the paper and outlines directions for future research.

2 PROPOSED METHODOLOGY

The proposed methodology aims to develop an accurate and robust deep learning–based system for static sign language gesture recognition, with a particular focus on alphabetic and numeric signs. To address the limitations of conventional CNNs in capturing global spatial dependencies, the framework combines a transfer learning strategy based on the VGG16 architecture with a self-attention mechanism. This hybrid design leverages the strong local feature extraction capability of convolutional layers while enhancing global contextual representation through attention. Figure 1 illustrates the overall architecture of the proposed framework. The overall methodology is structured as a sequential pipeline that includes dataset preprocessing, feature extraction using a pre-trained VGG16 backbone, self-attention–based feature enhancement, and final gesture classification. A comparable deep learning workflow based on pre-trained CNNs has also been reported in iJOE for image-based classification applications [9].

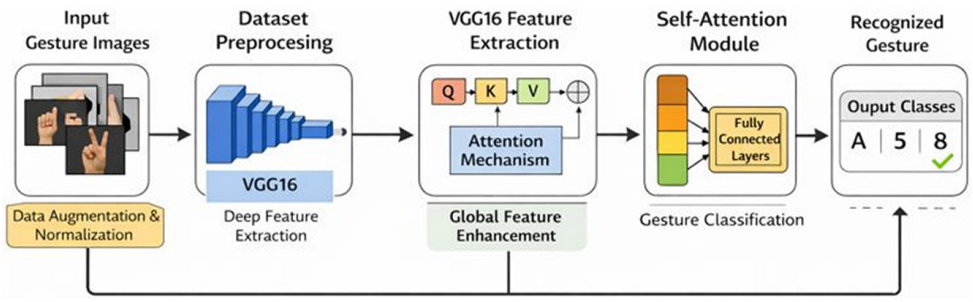

Fig. 1. Proposed static sign language recognition framework using VGG16 with self-attention

2.1 Dataset description

The dataset [14] used in this study is a comprehensive collection of static hand-gesture images representing symbols commonly used in sign language communication. It consists of 37 distinct gesture classes, including 26 alphabetic gestures (A–Z), 10 numeric gestures (0–9), and one "space" gesture used to denote separation between words or letters. The dataset is organized into two main directories: Gesture Image Data, which contains colored hand-gesture images, and Gesture Image Pre-Processed Data, which provides thresholded binary images optimized for machine learning applications.

Each gesture class contains 1,500 samples with a spatial resolution of 50×50 pixels, resulting in a total of 55,500 images per directory. The pre-processed dataset enhances gesture boundary clarity and reduces background noise, thereby facilitating more efficient feature extraction by CNNs. Owing to its large scale, balanced class distribution, and uniform image structure, this dataset is highly suitable for training and evaluating deep learning models for static SLR. Figure 2 provides representative examples from several gesture classes within the dataset, showing the variability in hand-shape, orientation and visual appearance across alphabetic, numeric, and space gestures.

Fig. 2. Visual samples of the 37-class static sign language gesture dataset

2.2 VGG16 architecture

VGG16 is a deep CNN developed by the Visual Geometry Group (VGG) at the University of Oxford. It is widely recognized for its simplicity and strong performance, relying exclusively on small 3 × 3 convolution filters and 2 × 2 max-pooling layers arranged in a uniform and highly structured architecture. The standard VGG16 model is composed of 13 convolutional layers followed by three fully connected layers, resulting in a total of 16 trainable layers—hence the name VGG16.

In this study, VGG16 is employed as a transfer-learning backbone, initialized with ImageNet pre-trained weights to take advantage of the rich low-level visual features learned from large-scale natural images. To adapt the model to the SLR task, the first four convolutional blocks are frozen, while only the last convolutional block (block5) is unfrozen and fine-tuned. This strategy allows the network to adjust high-level representations while preserving stable low-level feature extraction.

After feature extraction, the output tensor from the convolutional base is flattened and passed to a custom classification head consisting of:

- a Flatten layer
- a Dense layer with 512 units and ReLU activation
- a Dropout layer (rate = 0.5) to reduce overfitting
- a final Dense layer with 37 softmax units corresponding to the 37 gesture classes.

The model is trained end-to-end using the Adam [13] optimizer (learning rate = 1×10^{-4}), the categorical cross-entropy loss function, a batch size of 32, and 50 training epochs. This fine-tuning configuration enables VGG16 to effectively learn discriminative spatial patterns in static sign-language gestures while ensuring fast and stable convergence.

2.3 Self-attention mechanism

The self-attention mechanism is a deep learning technique that enables a model to identify the most relevant regions of an input by dynamically computing relationships between all of its spatial or sequential elements. Originally introduced in natural language processing, self-attention has recently become a powerful tool in computer vision due to its ability to capture long-range dependencies that traditional CNNs may fail to model effectively.

Unlike convolutional filters, which operate on local receptive fields and extract features progressively, self-attention computes interactions between every pair of positions in the feature map. Given an input feature matrix, three learned projections—Query (Q), Key (K), and Value (V)—are generated. The attention mechanism then measures the similarity between each query and all keys using dot-product scores, applies a softmax normalization, and uses the resulting weights to aggregate the value vectors, as defined in Equation (1):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

This operation allows the network to focus on the most informative spatial regions of the image, regardless of their physical location. As a result, the model can better distinguish between gestures that share similar local patterns but differ

in global spatial configuration. Modern architectures frequently employ multi-head self-attention (MHSA), where multiple attention heads operate in parallel, enabling the network to learn diverse contextual relationships. This mechanism enhances global feature interaction, improves robustness to spatial variations, and strengthens the network's ability to model complex gesture structures. To better clarify this process, Figure 3 illustrates the self-attention mechanism, where the similarity between queries and keys is computed and used to weight the value features across the entire input space.

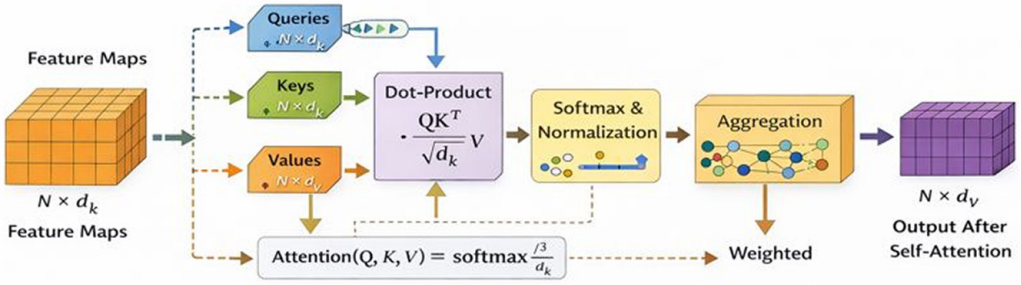

Fig. 3. Self-attention mechanism

2.4 Model evaluation

The performance of the proposed models was evaluated using standard classification metrics commonly adopted in SLR and computer vision tasks. These metrics provide a comprehensive assessment of both overall accuracy and class-wise recognition behavior.

The primary evaluation metric is classification accuracy, which measures the proportion of correctly classified samples over the total number of samples and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively. To further analyze class-wise performance, precision, recall, and F1-score were computed for each gesture class. Precision measures the correctness of positive predictions and is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall evaluates the model's ability to correctly identify all relevant samples:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The F1-score, which represents the harmonic mean of precision and recall, provides a balanced evaluation metric:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

In addition to these quantitative metrics, a confusion matrix was employed to visually analyze classification results across the 37 gesture classes. The diagonal

elements represent correctly classified gestures, while off-diagonal entries indicate misclassifications, particularly among visually similar alphabetic and numeric signs. Furthermore, training and validation accuracy and loss curves were monitored throughout the learning process to evaluate convergence behavior and detect possible overfitting or underfitting. These curves provide insights into the stability and generalization capability of both the baseline VGG16 model and the proposed self-attention-enhanced architecture.

All evaluation metrics were computed on the validation set, ensuring a fair and consistent comparison between the evaluated models.

3 EXPERIMENTAL RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained from evaluating the proposed deep learning models for static sign language gesture recognition. The experiments were designed to assess the effectiveness of the baseline VGG16 architecture and the proposed VGG16 enhanced with a self-attention mechanism under identical training and evaluation conditions.

The performance analysis focuses on both quantitative metrics and qualitative observations. Standard evaluation measures, including accuracy, precision, recall, F1-score, confusion matrices, and learning curves, are used to provide a comprehensive assessment of the models' recognition capabilities across the 37 gesture classes. In addition, training and validation behavior is analyzed to examine convergence stability and generalization performance. A comparative discussion is then conducted to highlight the impact of integrating self-attention into the VGG16 architecture. Particular attention is given to improvements in classification consistency, reduction of misclassification between visually similar gestures, and overall robustness of the proposed approach.

3.1 Experimental setup

All experiments were conducted to evaluate the performance of the proposed static SLR framework based on deep transfer learning. The implementation was carried out in Python using the TensorFlow and Keras deep learning frameworks, and all models were trained in a GPU-enabled environment to ensure computational efficiency.

The experiments were performed on a large-scale static sign language dataset comprising 37 gesture classes, including alphabetic and numeric signs as described in Section 2.1. The dataset was split into 70% for training and 30% for validation using the built-in validation_split mechanism, ensuring a balanced representation of all gesture classes in both subsets.

Prior to training, all images were resized to 224 × 224 pixels to match the input requirements of the VGG16 architecture. Pixel values were normalized using the ImageNet preprocessing function. To improve generalization and reduce overfitting, data augmentation techniques were applied during training, including horizontal flipping, width and height shifting (±0.2), and zooming (0.1).

Two deep learning models were evaluated under identical conditions: a baseline VGG16 model and an enhanced VGG16 model incorporating a self-attention mechanism. In both cases, the convolutional backbone was initialized with ImageNet pre-trained weights. The first four convolutional blocks were frozen, while the last convolutional block (block5) was fine-tuned to adapt high-level feature representations to the SLR task.

For both architectures, a custom classification head was added on top of the feature extractor. Training was performed using the Adam optimizer with a learning rate of 1×10^{-4} and the categorical cross-entropy loss function. All models were trained for 50 epochs with a batch size of 32.

During training, accuracy and loss were monitored on both the training and validation sets. After training, model performance was evaluated using standard metrics, including overall accuracy, confusion matrices, precision, recall, and F1-score, to provide a detailed analysis of recognition performance across all gesture classes. For reproducibility and fair comparison, Table 1 reports the main experimental setup and hyperparameter configuration used for training and evaluating all models.

Table 1. Experimental setup and training configuration

Hyperparameter	Value
Input image dimension	(224, 224, 3)
Activation function	Softmax
Loss function	categorical cross-entropy
Data augmentation	Horizontal flip, width/height shift (±0.2), zoom (0.1)
Optimizer	Adam
Learning rate	0.0001
Momentum	0.9
Decay	1e-4
Batch size	32
Batch normalization	Yes
Number of epochs	50

3.2 Experimental results

This subsection presents the experimental results obtained from training and evaluating the baseline VGG16 model and the proposed VGG16 enhanced with a self-attention mechanism. The models were assessed under the same experimental conditions described in Section 3.1 in order to ensure a fair comparison. Performance is analyzed using learning curves, confusion matrices, and standard classification metrics such as accuracy, precision, recall, and F1-score. The obtained results highlight the effectiveness of deep transfer learning for static SLR and demonstrate the impact of integrating self-attention on classification stability and overall recognition accuracy.

Results of VGG16. After 50 training epochs, the fine-tuned VGG16 model achieved a training accuracy of 99.86%, while maintaining a validation accuracy of 99.80%, confirming the effectiveness of transfer learning in extracting discriminative features from static hand gesture images. The final evaluation on the validation set resulted in a test accuracy of 99.80% with a test loss of 0.18, indicating strong generalization across the 37 gesture classes.

Figure 4 shows the learning curves of accuracy and loss for both training and validation sets. The model converges rapidly during the early epochs and reaches near-saturation performance, with training and validation curves remaining closely aligned. This behavior suggests stable optimization and limited overfitting. Although slight fluctuations appear in the validation loss, they remain within a low range, reflecting robust learning under the applied data augmentation strategy.

Figure 5 presents the confusion matrix obtained on the validation set. The matrix exhibits clear diagonal dominance, demonstrating highly accurate class-wise recognition. Only minor misclassifications occur between a small number of visually similar gestures, which can be attributed to subtle variations in finger positioning and hand orientation. Overall, these results confirm that the fine-tuned VGG16 model provides a strong baseline for static sign language gesture recognition.

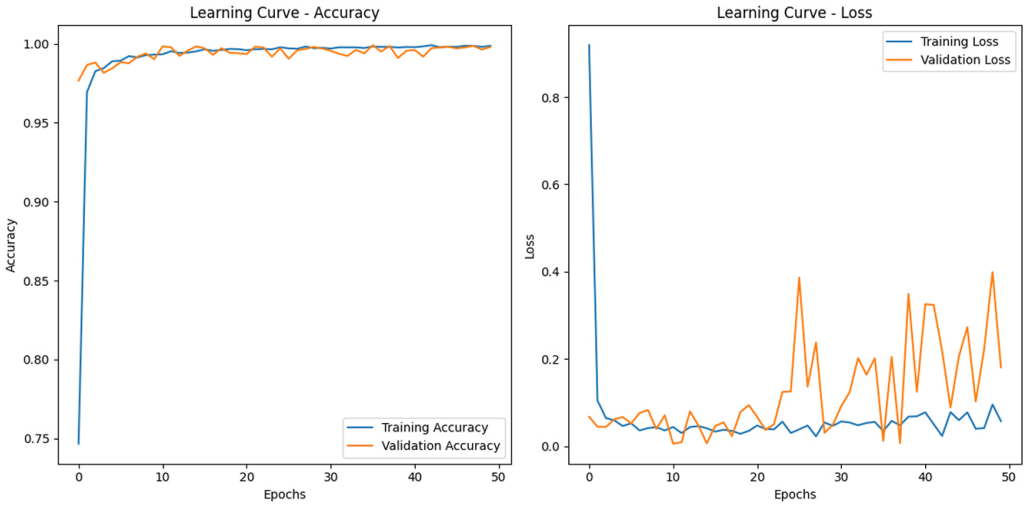

Fig. 4. Training and validation accuracy/loss curves of the fine-tuned VGG16 model over 50 epochs

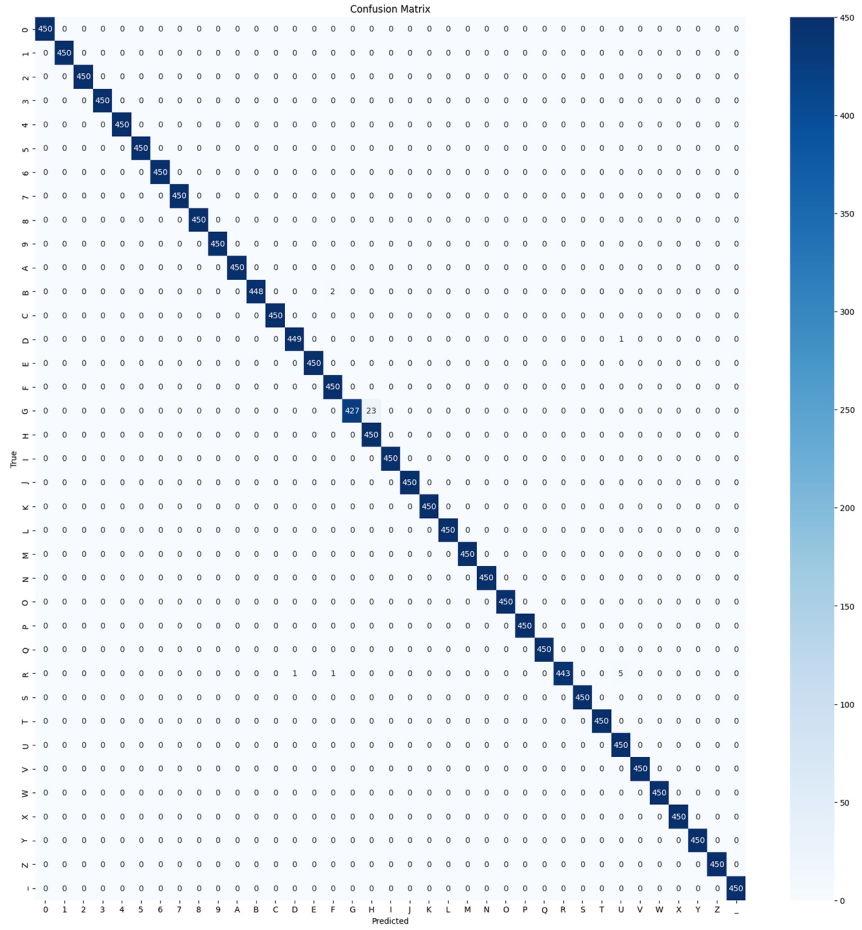

Fig. 5. Confusion matrix of the VGG16 model on the validation set across 37 gesture classes

Results of VGG16 + self-attention. The proposed VGG16 enhanced with a self-attention mechanism achieved excellent performance on the static sign language gesture recognition task. After 50 epochs, the model reached a training accuracy of 99.86% and a validation accuracy of 99.87%, demonstrating high learning stability and strong generalization across the 37 gesture classes. The final evaluation on the validation set yielded a test accuracy of 99.87% with a notably low test loss of 0.0049, indicating highly confident predictions and improved optimization compared to the baseline model.

Figure 6 illustrates the training and validation learning curves of the VGG16 + Self-Attention model. The model converges rapidly within the first few epochs, and both accuracy curves remain closely aligned throughout training, suggesting consistent learning behavior and limited overfitting. Although validation loss shows occasional fluctuations, the overall trend remains low, confirming robust performance under the applied data augmentation strategy.

To further evaluate the classification behavior, the confusion matrix shown in Figure 7 demonstrates strong diagonal dominance, confirming that the model correctly recognizes the vast majority of gesture classes. Misclassifications are minimal and mostly occur among visually similar gestures, which remain challenging due to subtle variations in hand posture and finger positioning. Overall, these results validate the effectiveness of integrating self-attention into the VGG16 backbone and highlight its contribution to enhancing global feature representation for static sign language gesture recognition.

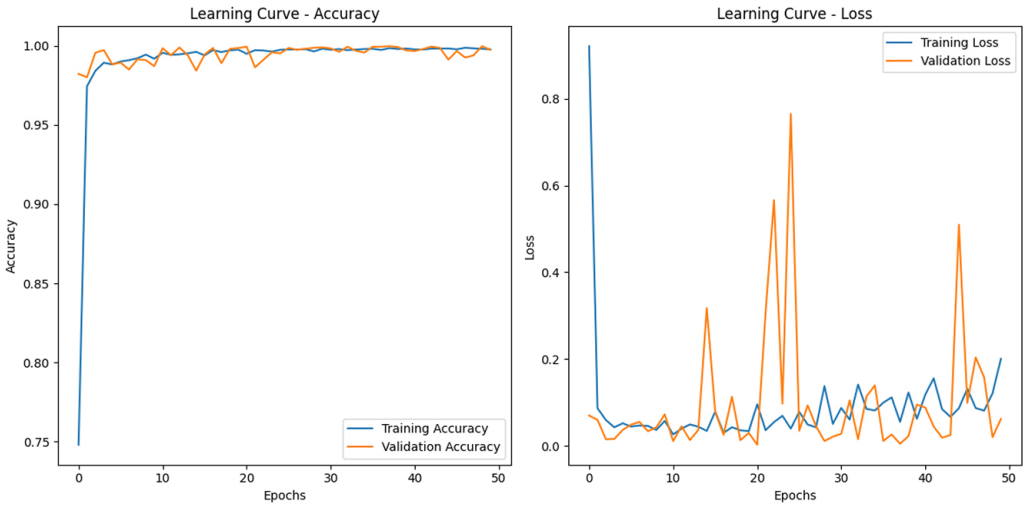

Fig. 6. Training and validation accuracy and loss curves of the VGG16 + Self-Attention model over 50 epochs

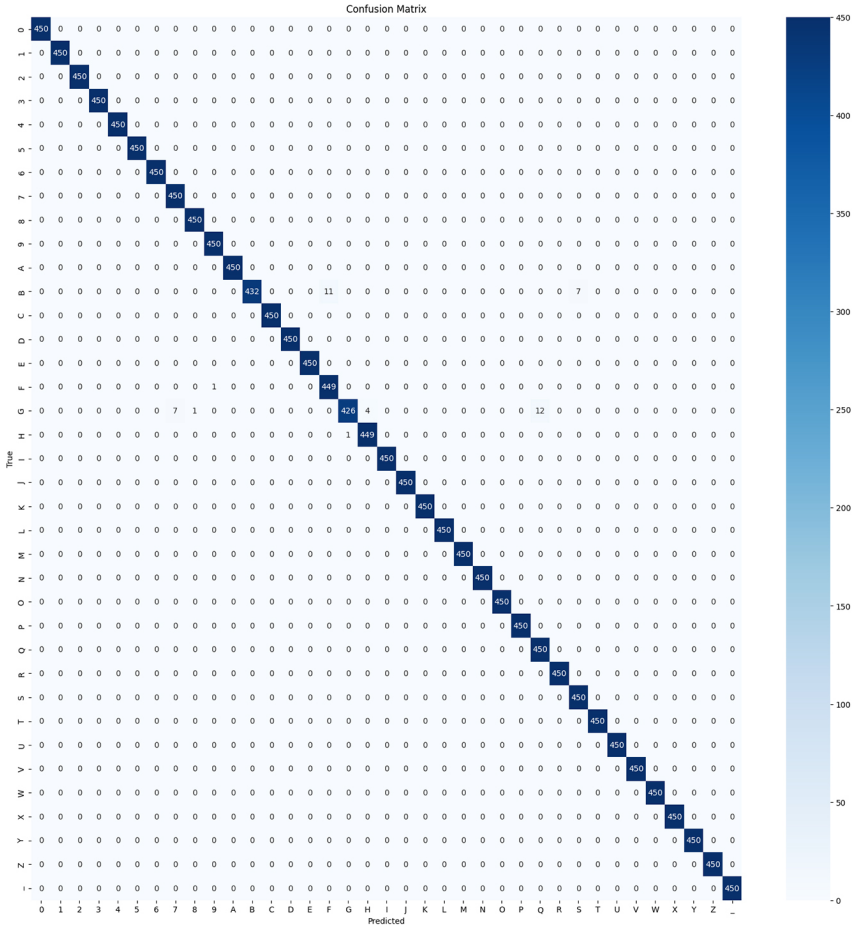

Fig. 7. Confusion matrix of the VGG16 + Self-Attention model on the validation set across 37 gesture classes

Comparative analysis (VGG16 vs VGG16 + self-attention). To further evaluate the impact of integrating self-attention, a comparative analysis was conducted between the baseline VGG16 model and the proposed VGG16 + Self-Attention architecture under identical training conditions. Table 2 summarizes the performance comparison between the two models in terms of training accuracy, validation accuracy, test accuracy, and test loss.

Table 2. Performance comparison between VGG16 and VGG16 + Self-Attention

Model	Training Accuracy	Validation Accuracy	Test Accuracy	Test Loss
VGG16	99.86%	99.80%	99.80%	0.1807
VGG16 + Self-Attention	99.86%	99.87%	99.87%	0.0049

The comparative results indicate that both models achieve near-perfect classification accuracy. However, integrating the self-attention mechanism leads to a noticeable improvement in optimization behavior, reflected by a significantly lower test loss. Although the gain in accuracy is numerically small, the substantial reduction in loss suggests that the attention-enhanced model produces more confident and stable predictions. This behavior can be attributed to the ability of self-attention to capture global spatial dependencies between different hand gesture regions, thereby improving discrimination between visually similar signs.

Static single-frame SLR systems are particularly useful in scenarios such as alphabet and digit learning, beginner-level educational tools, and assistive communication interfaces in constrained environments where isolated gestures are sufficient. These use cases are especially relevant for inclusive learning settings, where lightweight and efficient recognition systems can provide practical support. It should also be noted that recent multimodal SLR systems often rely on richer inputs such as video sequences, hand landmarks, or textual context, whereas the proposed framework focuses on static image-based recognition. Therefore, direct comparison remains limited due to differences in input modality and evaluation settings. Nevertheless, the proposed VGG16 + Self-Attention model provides a stronger and more reliable framework for static sign language recognition.

Table 3 reports a comparative evaluation between the proposed models and selected image-based SLR approaches. The 2D-CNN combined with a modified LSTM method achieved an accuracy of 89.50%, while the augmented CapsNet model reached 95.08%, confirming the effectiveness of deep learning models for static gesture classification. In contrast, the proposed fine-tuned VGG16 model provides a substantial improvement, achieving 99.80% accuracy on the considered dataset. Moreover, integrating a self-attention module into the VGG16 architecture further enhances performance to 99.87%, suggesting that self-attention improves global feature representation and helps reduce confusion between visually similar gestures. Although direct comparisons remain limited due to differences in datasets and experimental protocols, the results demonstrate that the proposed self-attention–based approach achieves highly competitive performance for static sign language gesture recognition.

Table 3. Comparison with selected related works and the proposed models

Method/Model	Dataset Type	Accuracy (%)
2D-CNN + Modified LSTM [15]	Images	89.50
CapsNet (Augmented) [16]	Images	95.08
VGG16 (fine-tuned)	Images	99.80
VGG16 + Self-Attention	Images	99.87

4 CONCLUSION AND FUTURE WORK

This paper proposed an effective deep learning framework for static sign language gesture recognition based on transfer learning. Two architectures were investigated: a fine-tuned VGG16 baseline and an enhanced VGG16 model integrating a MHSA module. Experiments were conducted on a 37-class dataset comprising alphabetic gestures (A–Z), numeric gestures (0–9), and a space gesture. The baseline VGG16 achieved a high recognition accuracy of 99.80%, confirming the suitability of convolutional feature extraction for static gesture classification. Moreover, incorporating self-attention further improved performance to 99.87%, while producing a substantially lower loss, which reflects more confident and stable predictions. These results demonstrate that self-attention enhances global feature representation and contributes to more robust discrimination between visually similar signs.

Despite the strong performance obtained on the considered dataset, the present study is limited to static gestures and evaluation on a single dataset under controlled conditions. Therefore, the proposed framework should be regarded as a promising

step toward intelligent assistive technologies rather than a fully validated deployment in inclusive educational environments. Future work will focus on extending the approach to dynamic SLR by integrating temporal modeling techniques such as 3D-CNNs, recurrent neural networks, or Transformer-based video architectures. In addition, further evaluation under real-world conditions, including background clutter, illumination changes, occlusions, and inter-subject variability, would provide a more comprehensive assessment of generalization and practical applicability. Finally, lightweight optimization and real-time deployment strategies may be explored to support accessible and efficient assistive systems.

5 REFERENCES

[1] World Health Organization (WHO), "World report on hearing," 2021. https://www.who.int/publications/i/item/world-report-on-hearing

[2] R. E. Mitchel and M. A. Karchmer, "Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States," *Sign Language Studies*, vol. 4, no. 2, pp. 138–163, 2004. https://doi.org/10.1353/sls.2004.0005

[3] N. El-Marzouki, I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time sign language recognition using parallel multi-scale CNN to enhance inclusive education for deaf and hard of hearing students," *Multimedia Tools and Applications*, vol. 84, pp. 38741–38760, 2025. https://doi.org/10.1007/s11042-025-20692-7

[4] UNESCO, "Global education monitoring report 2020: Inclusion and education: All means all," 2020. https://unesdoc.unesco.org/ark:/48223/pf0000373718

[5] Y. Saleh and G. F. Issa, "Arabic sign language recognition through deep neural networks fine-tuning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 5, pp. 71–83, 2020. https://doi.org/10.3991/ijoe.v16i05.13087

[6] P. G. Veronica, R. K. Mokkapati, L. P. Jagupilla, and C. Santhish, "Static hand gesture recognition using novel CNN-based methods," *International Journal of Online and Biomedical Engineering (iJOE),* vol. 19, no. 9, pp. 131–141, 2023. https://doi.org/10.3991/ijoe.v19i09.39927

[7] M. Safeel, T. Sukumar, K. S. Shashank, M. D. Arman, R. Shashidhar, and S. B. Puneeth, "Sign language recognition techniques—A review," in *Proc. 2020 IEEE Int. Conf. Innov. Technol. (INOCON)*, Bengaluru, India, 2020, pp. 1–9. https://doi.org/10.1109/INOCON50539.2020.9298376

[8] A. Kurt and M. K. Erden, "Investigation of the opinions of pre-service special education teachers on the use of assistive technologies in special education," *Education and Information Technologies*, vol. 29, pp. 51–76, 2024. https://doi.org/10.1007/s10639-023-12278-3

[9] I. Lasri, N. El-Marzouki, A. Riadsolh, and M. Elbelkacemi, "Automated detection of dental caries from oral images using deep convolutional neural networks," *International Journal of Online & Biomedical Engineering*, vol. 19, no. 18, pp. 53–70, 2023. https://doi.org/10.3991/ijoe.v19i18.45133

[10] N. El-Marzouki, I. Lasri, A. Riadsolh, and M. Elbelkacemi, "American sign language recognition with convolutional neural networks: A gateway to enhanced inclusivity," in *International Conference on Digital Technologies and Applications (ICDTA 2024),* in Lecture Notes in Networks and Systems, Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-68660-3_3

[11] S. G. Y. N. Reddy, S. Mondal, P. M. G. B. Asdaque, and K. V. Krishna, "A real-time sign language recognition system using MediaPipe and random forest classifier combining with text-to-speech technique," in *2025 International Conference on Artificial Intelligence and Data Engineering, IEEE*, 2025, pp. 925–931. https://doi.org/10.1109/AIDE64228.2025.10986900

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR),* 2015.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR),* 2015.

[14] A. Khan, "Sign language gesture images dataset," *Kaggle*, 2019. Retrieved from https://www.kaggle.com/datasets/ahmedkhanak1995/sign-language-gesture-images-dataset

[15] C. Aparna and M. Geetha, "CNN and stacked LSTM model for indian sign language recognition," in *Machine Learning and Metaheuristics Algorithms, and Applications*, SoMMA 2019, in Communications in Computer and Information Science, S. Thampi, L. Trajkovic, K. C. Li, S. Das, M. Wozniak, and S. Berretti, Eds., vol. 1203, Springer, Singapore, 2020. https://doi.org/10.1007/978-981-15-4301-2_10

[16] P. Shreeyash, N. Prathamesh, and N. Nayan, "Sign language recognition: A deep convolutional neural network approach for accurate alphabet identification," *Int. Res. J Mod. Eng. Technol. Sci.*, 2023.

6 AUTHORS

Naoufal El-Marzouki is a PhD student at the Faculty of Sciences Rabat, University Mohammed V in Rabat, Morocco. He is a member of the Laboratory of Conception and Systems (Electronics, Signals and Informatics). He received a Master's degree in Mathematics from the Faculty of Sciences Kenitra. His current field of research is artificial intelligence applied to education (E-mail: naoufal_elmarzouki@um5.ac.ma).

Imane Lasri is a Professor of Computer science and artificial intelligence in ENSAM, University Mohammed V in Rabat. She is a member of the Laboratory of Conception and Systems (Electronics, Signals and Informatics). She received a PhD in artificial intelligence and Master's degree in Big Data Engineering from the Faculty of Sciences Rabat. She received the awards of excellence of the major winners from the Mohammed V University in Rabat in 2019. She is interested in artificial neural networks and deep learning. She is the author of many research studies published in international journals and conference proceedings (E-mail: imane_lasri@um5.ac.ma).

Fatim Ezzahrae Dorhmi is a Ph.D. candidate at the Laboratory of Conception and Systems (Electronics, Signals and Informatics), Faculty of Sciences, Mohammed V University in Rabat, Morocco. Her research interests include artificial intelligence, deep learning, computer vision, and sign language recognition, with a particular focus on assistive technologies and inclusive educational application (E-mail: fatimezzahrae.dorhmi@um5r.ac.ma).

Anouar Riadsolh received his PhD in Computer Science from the Faculty of Sciences Rabat (FSR), University Mohammed V in Rabat, Morocco. He is a Professor at the FSR. He is a member of the laboratory of conception and systems (electronics, signals, and Informatics), FSR. His current research interests are focused on data mining, big data, and machine learning (E-mail: a.riadsolh@um5r.ac.ma).

Mourad Elbelkacemi has received PhD in Computer Science. He was the Dean of the Faculty of Sciences Rabat (FSR). He is a Professor at the FSR. He is a member of the laboratory of conception and systems (electronics, signals, and Informatics), FSR, University Mohammed V in Rabat, Morocco. His main research interests are focused on electronics, education, data mining, and big data (E-mail: mourad_prof@yahoo.fr).