Recognizing Gesture images with ViT and Spatial Attention Regularization

Zahraa Thamer[1], Noor S.Sagheer[1], Ashwan A.Abdulmunem[1], Hawraa Thamer[1] and Og˘uz Ata[2]
[1]*Department of Information Technology, Department of Computer Science, University of Karbala, 56001 Karbala, Iraq*
[2]*Department of Information Technology, College of Computer Engineering, Altınbaş University, Dilmenler Str. 26, 34218 Istanbul, Turkey*
{Zahraa.th, noor.sabah, Ashwan.a,hawraa.thamer}@uokerbala.edu.iq,oguz.ata@altinbas.edu.tr

Keywords: Computer Vision, ViT, Gesture Images, Spatial Attention Regularization, Image Analysis, Hand Gesture Recognition D.

Abstract: One important area of Human-Computer Interaction (HCI) is image-based gesture recognition. Despite tremendous advancements, it is still very difficult to achieve reliable and accurate gesture recognition in unrestricted, real-world settings. Conventional techniques frequently find it difficult to handle changes in lighting, background noise, occlusions, size variations, and the innate similarity between various gestures. To enhance the discriminative ability of the Vision Transformer (ViT) model for intricate hand gestures, this work presents a carefully planned fine-tuning methodology. Encourage ViT to concentrate on salient gesture regions while remaining resilient to environmental noise; the proposed method combines an adaptive learning rate scheduling system with a novel spatial attention regulator during fine-tuning. Experiments on a challenging and varied gesture dataset demonstrate that the proposed approach significantly performs better than state-of-the-art methods, attaining superior accuracy reaching 100% and demonstrating generalization capabilities. This study opens the door for more user-friendly human-computer interaction systems by providing a highly effective and flexible framework for sophisticated image-based gesture recognition systems.

1 INTRODUCTION

More organic and fluid interactive systems are quickly taking over our world. In this regard, Hand Gesture Recognition (HGR) shows up as a useful substitute interface that allows users to attain natural and intuitive control in a variety of applications, ranging from autonomous driving and medical rehabilitation to smart environments and augmented reality. Conventional techniques frequently find it difficult to handle lighting variations, intricate backgrounds, occlusions, scale changes, and the innate similarities among various gestures [1]. Deep Learning (DL) has emerged as a key component of computer vision tasks, such as gesture recognition, which is essential for assistive technologies, virtual reality, and human-computer interaction. Because Convolutional Neural Networks (CNNs) can automatically extract hierarchical spatial features from images and videos, they have demonstrated remarkable success among DL techniques [2]. CNN-based models are appropriate for identifying both static and dynamic gestures because they can efficiently capture local patterns as shapes, edges, and motion cues. Nevertheless, CNNs frequently have trouble simulating long-range dependencies and need sizable datasets to generalize effectively [3]. ViTs have recently become a potent substitute. The self-attention mechanism, which ViTs rely on in contrast to CNNs, allows the model to recognize global relationships among image patches. This makes it possible for ViTs to perform better in complicated gesture recognition scenarios, particularly when contextual or temporal dependencies are crucial. According to studies, ViTs are more flexible when modeling sequential gesture dynamics and can outperform CNNs on large datasets [4]. Image recognition has been transformed by the introduction of deep learning, especially CNNs [5]. More recently, ViTs initially established for natural language processing, have shown superior results on image tasks by using self-attention mechanisms to capture long-term dependencies. Pre-trained ViTs significantly improve general image classification,

but specific adaptation is needed for gesture recognition due to its fine-grained nature, where exact spatial and temporal features are crucial [6]. In this paper, the authors suggest a method for fine-tuning pre-trained ViT models that is especially designed for reliable image gesture recognition. Our main contribution is the development of an adaptive fine-tuning protocol that improves ViTs' discriminative ability to detect minute gesture variations in a variety of scenarios while also leveraging their powerful feature extraction capability. Experiments were conducted on a unique, challenging, and realistic gesture dataset [7]. To validate our methodology. Thousands of images from various gesture categories are included in this dataset. The images were taken in a variety of settings, such as with varying lighting, indoor and outdoor backgrounds, subject distances, and minor overlaps. The dataset's diversity is essential for assessing the proposed method's resilience and generalizability. Every picture has a precise label indicating which gesture category it belongs to.

2 RELATED WORKS

Hand gesture recognition is vital for natural HCI applications such as VR, robotics, and sign language. Earlier studies focused on traditional machine learning methods like K-NN, SVM, and ANN, which showed limited performance on static gesture datasets [7]-[9] as in Table 1. Deep neural networks, particularly CNNs, significantly improved accuracy; for example, CNN models like YOLOv8n achieved up to 93.61% accuracy in gesture recognition [8], [9], though still below state-of-the-art. Recently, Vision Transformers (ViT) have outperformed CNNs, with models like HGR-ViT showing very high accuracy [10]. Hybrid models combining CNN and ViT, such as FasterViT, demonstrated superior accuracy and speed [11]. Additionally, studies using sEMG with deep learning achieved >95% accuracy for prosthetic control, with optimal hyperparameters (learning rate 0.0001-0.001, 80-100 epochs) enhancing performance [12], [13].

3 METHODOLOGIES

The suggested model design methodology was implemented in two main phases. The first phase involves data preprocessing, while the second phase focuses on building and configuring the model according to the specified requirements.

3.1 Dataset

The Hand Gesture Recognition Dataset [7], available on Kaggle by Arya Rishabh, was used as the cornerstone of this research. The images are classified into 20 specific gestures as showed in Table 2. The images vary in lighting, camera angle, left and right-hand orientation, and backgrounds. The images are organized into separate folders for each category, making them easy to load and process using libraries like tensorflow and keras.

Figure 1 illustrates the image preprocessing steps before feeding them into the proposed model. The process begins by resizing the image to (224 * 224) and then converting the grayscale images with dimensions (H, W) to RGB color images with dimensions (H, W, 3) by iterating the values across the three channels.

Pixel values are then normalized from the range [0, 255] to the range [-1, 1] to improve training performance using ViTImageProcessor.from pretrained("google/vit-base-patch16-224", do_rescale=False) according to the snippet. Finally, the images are re-formatted into tensors with dimensions (224 * 224 * 3) in preparation for feeding them into the deep learning model.

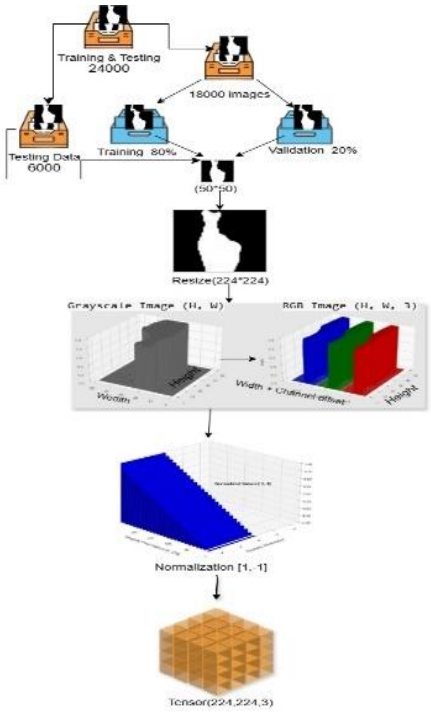

Figure 1: Preprocessing phase.

Table 1: Summarization of the related works.

Re.	Methods	Dataset	Results
[10]	HGR-ViT	National University of Singapore	99.85%
		Digits with American Sign Language (ASL)	99.36%
		ASL	99.98%
[14]	ResNet50, MobileNet V3 small, YOLO-Nas, YOLOv8	HaGRID	F1-score is 98.3%, 86.4%, 99.37%, 99.21%, respectively
		TQU-HG	F1-score is 99.36%, 99.27%, 98.98%, 98.99%, respectively
[11]	FasterViT	ASL	99.48% accuracy
		Indian Sign Language (ISL)	100% accuracy
[15]	CNN	Tanzanian Sign Language	96% accuracy
	SVM		95% accuracy
[16]	XGBoost	Sign MNIST	81.35% accuracy
	Random forest		84.43% accuracy
	CNN		91.41% accuracy
	Stochastic Gradient Descent Classifier		59.80% accuracy
	Naïve Bayes Gaussian		38.9% accuracy
	Naïve Bayes Multinomial		46.85% accuracy
	logistic regression		68.21% accuracy
	KNN		80.46% accuracy
	SVC		66.85% accuracy
[9]	KNN	ASL	92.71% accuracy
	SVM	Hand gesture consisting of 14,000 samples from 35 individuals.	97.1% accuracy
	Random-Forest	sourced from the UCI machine learning	81.69% accuracy
	CNN	20000 images of hand gesture	93.61% accuracy
	Naïve Bayes Algorithm	various images of hand gestures	90.6% accuracy
	RNN	To determine the 21 hand features across utilizing Media Pipe	99.28% accuracy
[12]	CNN	sEMG	>95% accuracy
[8]	YOLOv8n model	Self-collected hand gesture, including 12,000 images	99.3% accuracy
[13]	CNN	sEMG data	Accuracy (>95%)

Table 2: Hand gesture 20 classes [7].

Gesture	Illustration	Gesture	Illustration	Gesture	Illustration
0		1		2	
3		4		5	
6		7		8	
9		10		11	
12		13		14	
15		16		17	
18		19			

3.3 Vision Transformer (ViT) Model Structure

An AI model based on the Transformer architecture, originally developed for natural language processing, but adapted for image processing. ViT divides an image into small patches, then transforms each patch into a numerical representation (embedding) and adds information about its position in the image (position encoding). It then utilizes self-attention to comprehend the relationships between different parts of the image, allowing it to capture visual patterns and features with high accuracy.

ViT is superior to convolutional neural networks (CNNs) in its ability to learn from large, complex images, sometimes achieving better performance when large training data are available. However, it requires significant computational resources [18].

This framework is based on the ViT architecture, specifically the google/vit-base-patch16-224 model, as shown in Figure 2.

Preprocessed the image by dividing it into pieces (Image Patching), the input is an RGB image of size (H, W, 3). The image is divided into square patches of fixed size (e.g., 16x16 pixels), (1), [17], [18]. $P \times P$. The resulting patches number is calculated as follows:

$$\mathrm{N} = \frac{\mathrm{H} \times \mathrm{W}}{\mathrm{P}^2} \ . \tag{1}$$

Each patch contains p2.3 elements (since the image is in color). The patch is flattened and then linearly projected into the model space (D Dimension) via a learnable weight matrix (2) [18]:

$$x_i E, \quad E \in R^{(P^2 \cdot 3) \times D}. \tag{2}$$

Output: A sequence of N vectors representing the patches.

Then, positional embeddings and class tokens were added. The class token is added at the beginning of the sequence as a special token x_{cls} and represents the final summary of the information.

Positional Embeddings are also added, which define the model by ordering the patches. The final equation for representing the input is (3) [18]:

$$\begin{aligned} X_0 &= x_{\mathrm{cs};x_1 E;\ldots;x_N E} + E0\mathrm{ps} \quad . \\ X_0 &\in R^{(N+1) \times D} \end{aligned} \tag{3}$$

Then, the Transformer Encoder was applied. It consists of several recursive blocks, each containing: After that applied Query, Key, and Value Projections (Q, K, V) (4) [18]:

$$\begin{aligned} Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \\ W_Q, W_K, W_V \in R^{D \times d_k} \end{aligned} \tag{4}$$

Then, Scaled Dot-Product Self-Attention (SDPSA) was applied the (5) [18]:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

The divider $\sqrt{d_k}$ prevents large values and maintains training stability [18]. Then, the layer-norm and residual connections are printed using the Pre-LayerNorm strategy (6), (7)[18]:

$$Y = X + \mathrm{MHSA}\big(\mathrm{LN}(X)\big) \tag{6}$$

$$Z = Y + \mathrm{MLP}\big(\mathrm{LN}(Y)\big) \tag{7}$$

Then, a Feed Forward Network (FFN) is applied (8) [18]:

$$\mathrm{FFN}(X) = \max(0, XW_1 + b_1)\, W_2 + b_2 \tag{8}$$

3.4 Proposed Fine-Tuning Strategy for Gesture Recognition

Our novel fine-tuning strategy for image gesture recognition on a ViT model includes two main components:

Adaptive Learning Rate Schedulingю A cosine decay learning rate with a warm-up phase is used to ensure stable fine-tuning of large pre-trained models, preventing early overfitting and improving convergence. We adopted a comprehensive Fine-Tuning strategy for ViT-B/16 layers with a learning scheduler that uses Warmup and Cosine Decay as shown by the (9)

$$LR(t) = \begin{cases} LR_{peak} \cdot \frac{t}{T_{warmup}} & \text{if } t < T_{warmup} \\ \frac{1}{2} LR_{peak} \cdot \left(1 + \cos\left(\pi \cdot \frac{t - T_{warmup}}{T_{total} - T_{warmup}}\right)\right) & \text{if } t \geq T_{warmup} \end{cases} \tag{9}$$

Spatial Attention Regularizationю A spatial attention regularizer guides ViT to focus on gesture-relevant regions by penalizing dispersed attention. It reduces attention map entropy, producing sharper focus on the hand and gesture areas for more accurate classification. To enhance the model's focus on motor features, we added a spatial attention regulator(R_{sar}) which encourages the spread (sparsity) in the spatial attention vectors emanating from the classification code$CLS\ token$ Thus, the total loss function is defined as follows (10).

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{R}_{SAR} \tag{10}$$

The overall fine-tuning process includes:

- Data preprocessing. This ensures consistency with the expected ViT input format.
- Classification header. A task-specific classification header (a simple dense layer with Softmax activation for multi-class classification) is added over the [CLS] code output from the fine-tuned ViT encoder.
- Loss function and optimizer. We use the class cross-propagation loss and the AdamW optimizer, known for its effectiveness in transformer training.

4 RESULTS AND DISCUSSION

The improved performance of our ViT-based approach for image gesture recognition can be attributed to the synergistic effect of leveraging a robust pre-trained vision transformer and applying a specialized fine-tuning strategy. Adaptive learning rate scheduling ensures stable and efficient convergence. At the same time, the spatial attention regulator plays a crucial role in fine-tuning the model's focus, making it more robust to realistic variations.which goes beyond simple transfer learning. By explicitly directing ViT's attention mechanisms toward salient gesture features, we enable the model to handle better the complexities of gesture recognition, where subtle visual cues are of paramount importance. This approach reduces the model's susceptibility to noise and background distractions, a common drawback in real-world computer vision applications.

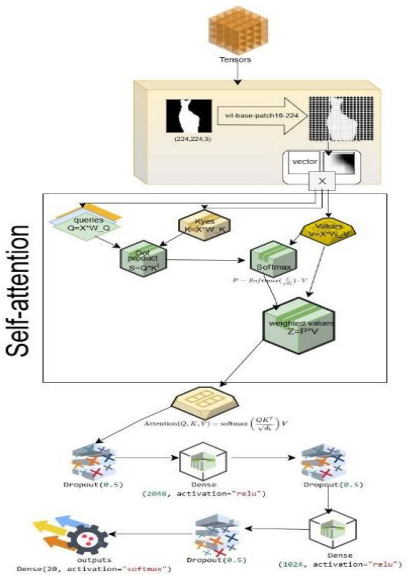

Figure 2: Architecture of ViT model.

The model was implemented utilizing Python language on Google Colab, T4 GPU. It was fine-tuned for 100 epochs on a custom gesture dataset, with a batch size of 32. We used an initial learning rate of 5e-5, applying adaptive learning rate scheduling and a spatial attention regularizer. Performance was evaluated using standard metrics such as precision (Pr), accuracy (Acc), recall (Re), and F1-score (F1s).

We compared our approach to several baselines, including:

- ResNet-50. A widely used CNN architecture trained from scratch and fine-tuned on ImageNet.
- ViT. Standard Fine-Tuning: The google/vit-base-patch16-224 model fine-tuned with standard learning rate scheduling and without a spatial attention regularize-r.
- Other CNN-based. Gesture Recognition Models: Selected contemporary models specifically designed for gesture recognition.
- The results illustrate as Table 3, the enhanced efficacy of the proposed ViT-based approach with the new fine-tuning strategy.

Table 3: Comparison of the Proposed Model's Performance with State-of-the-Art Methods.

Model	Recall	Accuracy	Precision	F1-score
ResNet50 [16]	99.42 %	98.92%	99.25%	99.33 %
CNN [19]	97%	96%	96%	96%
FastVit [18]	70.40 %	99.74%	81.99%	73.47 %
Proposed Model	100%	100%	100%	100%

The proposed method consistently achieved higher accuracy and better F1 scores compared to all baselines. The significant performance gains over standard ViT fine-tuning highlight the effectiveness of our adaptive learning rate scheduling and spatial attention scheduler in guiding the model to learn more discriminative features for gesture recognition.

In Figure 3 shows the training and validation accuracy across the number of epochs during training of the deep learning model. Both accuracies reached a near-perfect value of nearly 100%, and the matching of the blue and yellow lines indicates that the model does not suffer from the overfitting problem.

Figure 4 showing the change in the loss value for both training and validation over the epochs' number. We observe that the loss reduced rapidly from the beginning to values close to zero and stabilized at a

very low level, with a clear match between the training curve and the validation curve, indicating that the model learned effectively without any overfitting or under generalization issues. Figure 5 shows the ROC curves for all classes in the validation data using the One-vs-Rest method, indicating perfect classification accuracy and no errors. This demonstrates excellent performance. Figure 6 shows identical ROC curves with an AUC of 1.0 for all classes, indicating perfect classification.

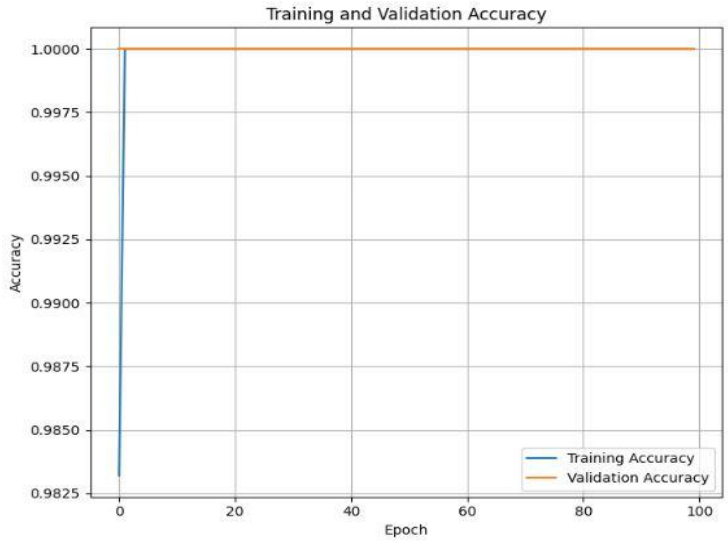

Figure 3: Accuracy graph.

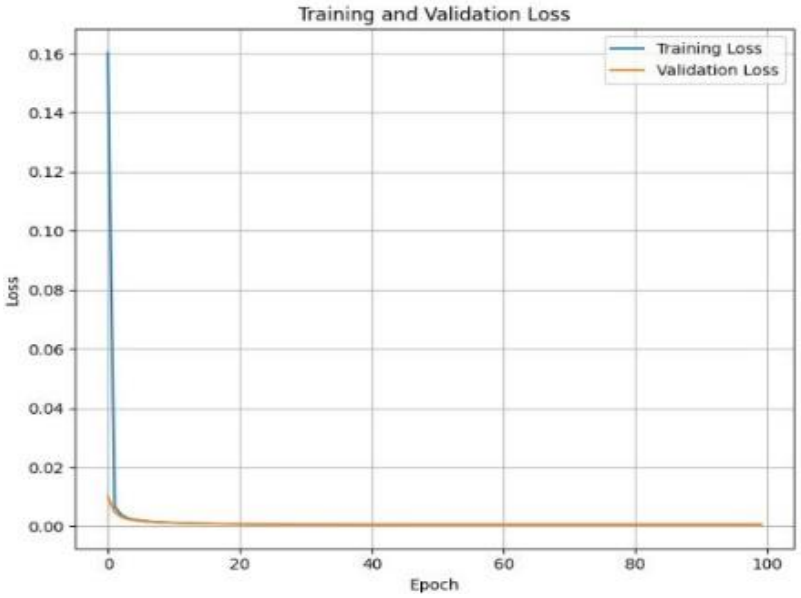

Figure 4: Loss graph.

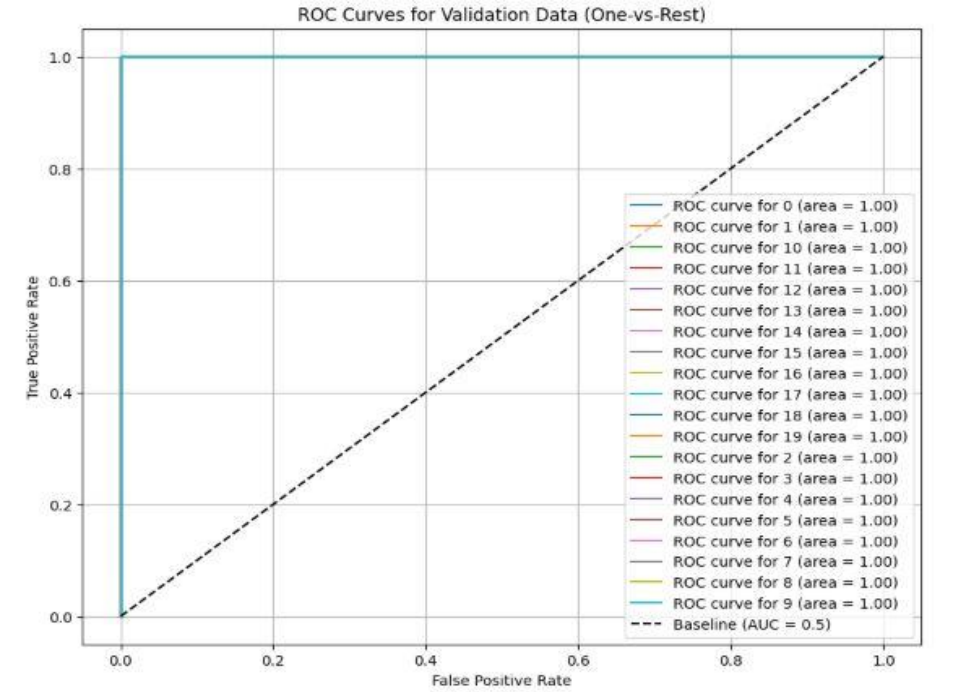

Figure 5: ROC curves for validation classes.

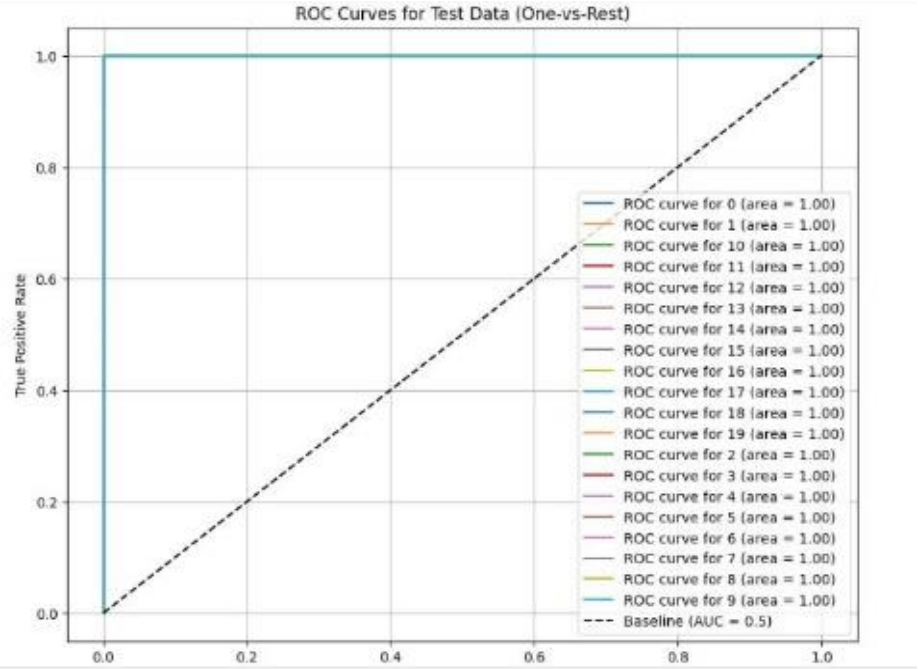

Figure 6: AUC performance curves.

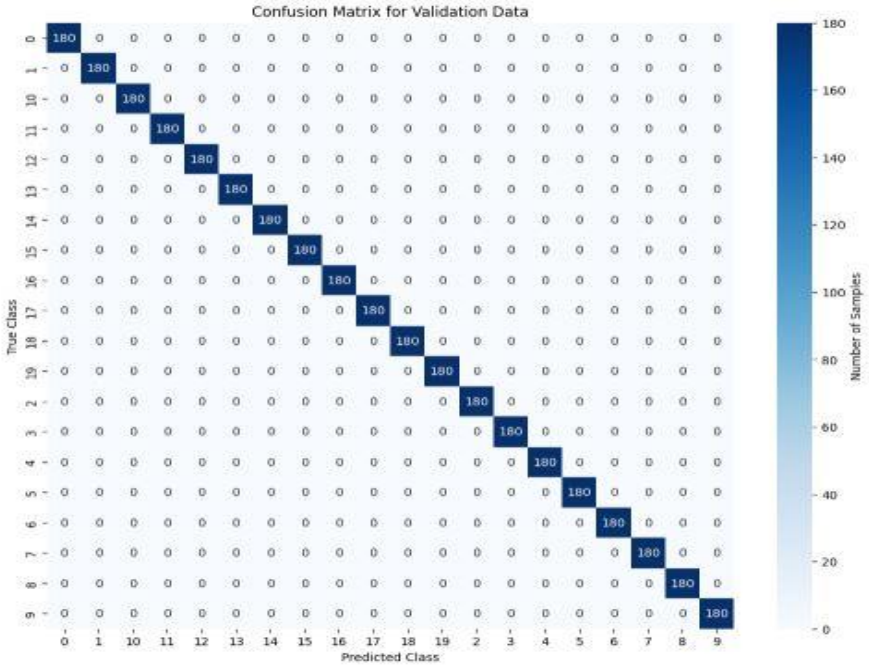

Figure 7: Confusion matrix.

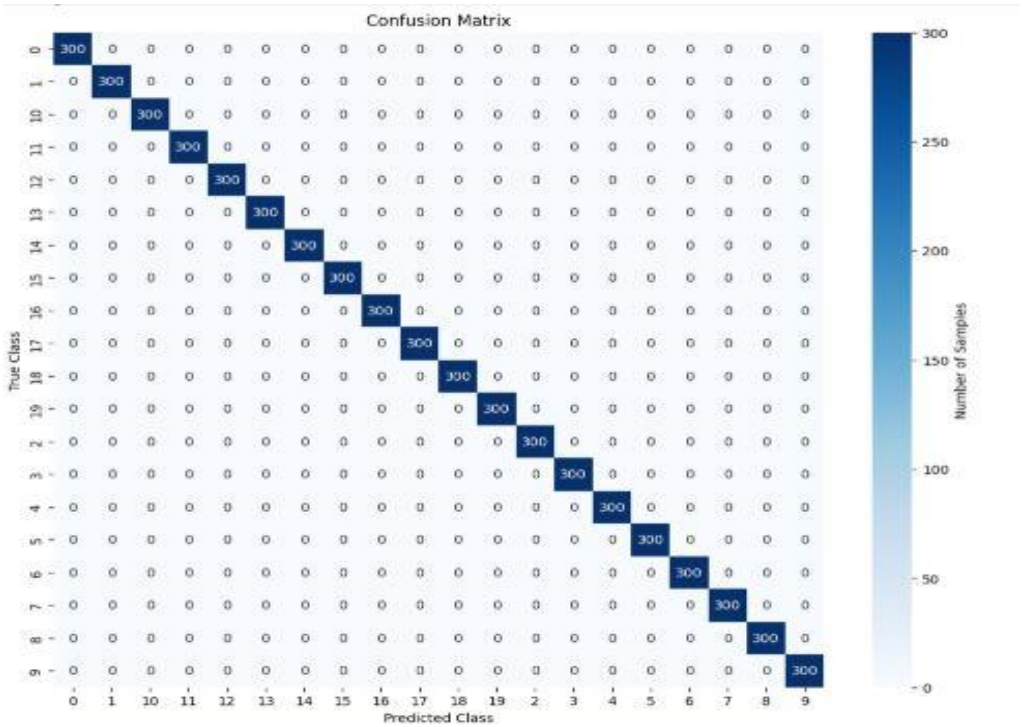

Figure 8: CM for testing data.

Figure 9:The result.

Figure 7 and Figure 8, the confusion matrices for both the validation and test sets confirm flawless high performance, with all predictions correctly classified.

Figure 9 illustrates the resulted image. Qualitative analysis shows that our fine-tuned ViT focuses accurately on hand and gesture regions, unlike standard ViT, which often attends to irrelevant areas. This confirms the spatial attention regulator's role in improving feature learning and performance.

The novelty of this work lies in an advanced fine-tuning method that directs ViT attention to key gesture features, improving recognition of subtle cues and reducing sensitivity to noise and background distractions.

5 CONCLUSIONS

This research presents an advanced approach based on the Vision Transformer (ViT) to develop a highly efficient hand gesture recognition system from images. The proposed model demonstrated outstanding ability to process complex gestures under realistic shooting conditions, benefiting from a carefully designed fine-tuning strategy that enhanced its sensitivity to prominent areas in the image.

The methodology is based on integrating an adaptive learning rate scheduling system with an innovative spatial attention regulator to improve the model's ability to identify the most significant gesture areas. This integration enabled the Vision Transformer to focus its learning resources on the most important features, thereby avoiding distractions from irrelevant details and directly impacting performance and generalization.

The experimental results showed clear superiority, with the model achieving an exceptional accuracy of 100% on the tested dataset, confirming the effectiveness of the fine-tuning method and its ability to extract strong and reliable visual representations.

6 FUTURE WORKS

The experiments also indicate that the model possesses high generalization capabilities, making it suitable for applications in diverse environments that require accurate and rapid responses. This study opens new horizons in the field of human-computer interaction (HCI), providing an advanced framework characterized by efficiency, flexibility, and adaptability to changing gestures and surrounding conditions.

This framework paves the way for more user-friendly and intelligent interactive systems based on visual recognition technologies supported by modern transducers, enabling the development of innovative applications in motion control, assistive systems, and smart interfaces.

REFERENCES

[1] A. Osman Hashi, S. Zaiton Mohd Hashim and A. Bte Asamah, "A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024," IEEE Access, vol. 12, pp. 143599-143626, 2024, [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3421992.

[2] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84-90, May 2017, , [Online]. Available: https://doi.org/10.1145/3065386.

[3] P. Molchanov, S. Gupta, K. Kim and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2015, pp. 1-7, [Online]. Available: https://doi.org/10.1109/CVPRW.2015.7301342.

[4] P. Mittal, B. Sharma and D. P. Yadav, "Comparative Analysis between CNN and ViT using Brain MRI Dataset," in 2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC), Dec. 2024, pp. 290-295, [Online]. Available: https://doi.org/10.1109/PDGC64653.2024.10984339.

[5] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov and N. Pacal, "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," Biomedical Signal Processing and Control, vol. 104, p. 107627, June 2025, , [Online]. Available: https://doi.org/10.1016/j.bspc.2025.107627.

[6] A. Al-Zebari, N. Omar and A. Sengur, “Vision Transformers-based Hand Gesture Classification,” in 2022 3rd International Informatics and Software Engineering Conference (IISEC), Dec. 2022, pp. 1-3, [Online]. Available: https://doi.org/10.1109/IISEC56263.2022.9998295.

[7] T. Kaggle, "Hand gesture recognition dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/tapakah68/hand-gesture-recognition-dataset

[8] T.-H. Nguyen, B.-V. Ngo and T.-N. Nguyen, “Vision-Based Hand Gesture Recognition Using a YOLOv8n Model for the Navigation of a Smart Wheelchair,” Electronics, vol. 14, no. 4, p. 734, Jan. 2025, , [Online]. Available: https://doi.org/10.3390/electronics14040734.

[9] Shivani and S. B. Gupta, "A comprehensive analysis of recognition of hand gestures using machine learning," *Makara Journal of Technology*, vol. 29, no. 1, Art. no. 5, 2025, doi: 10.7454/mst.v29i1.1679.

[10] C. K. Tan, K. M. Lim, R. K. Y. Chang, C. P. Lee and A. Alqahtani, “HGR-ViT: Hand Gesture Recognition with Vision Transformer,” Sensors, vol. 23, no. 12, p. 5555, Jan. 2023, , [Online]. Available: https://doi.org/10.3390/s23125555.

[11] Y. Altaf, “Efficient Hand Sign Recognition with Fine-Tuned Faster Vision Transformers: A Comparative Study on Benchmark image Datasets,” Journal of Electrical Systems, vol. 20, no. 3, pp. 8082-8098, Apr. 2024.

[12] A. R. Asif et al., “Performance Evaluation of Convolutional Neural Network for Hand Gesture Recognition Using EMG,” Sensors, vol. 20, no. 6, p. 1642, Jan. 2020, , [Online]. Available: https://doi.org/10.3390/s20061642.

[13] H. Hellara, R. Barioul, S. Sahnoun, A. Fakhfakh and O. Kanoun, “Comparative Study of sEMG Feature Evaluation Methods Based on the Hand Gesture Classification Performance,” Sensors, vol. 24, no. 11, p. 3638, Jan. 2024, , [Online]. Available: https://doi.org/10.3390/s24113638.

[14] V.-D. Do, V.-H. Le, H.-S. Do, V.-N. Phan and T.-H. Te, “TQU-HG dataset and comparative study for hand gesture recognition of RGB-based images using deep learning,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 34, no. 3, pp. 1603-1617, 2024.

[15] K. Myagila and H. Kilavo, "A comparative study on performance of SVM and CNN in Tanzania sign language translation using image recognition," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2005297, Nov. 2021, doi: 10.1080/08839514.2021.2005297.

[16] S. Bhushan, M. Alshehri, I. Keshta, A. K. Chakraverti, J. Rajpurohit and A. Abugabah, “An Experimental Analysis of Various Machine Learning Algorithms for Hand Gesture Recognition,” Electronics, vol. 11, no. 6, p. 968, Jan. 2022, , [Online]. Available: https://doi.org/10.3390/electronics11060968.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, Oct. 2020.

[19] K. Gupta, A. Singh, S. R. Yeduri, M. B. Srinivas and L. R. Cenkeramaddi, “Hand gestures recognition using edge computing system based on vision transformer and lightweight CNN,” J Ambient Intell Human Comput, vol. 14, no. 3, pp. 2601-2615, Mar. 2023, , [Online]. Available: https://doi.org/10.1007/s12652-022-04506-4.