Available *online* at : http://jnte.ft.unand.ac.id/

Jurnal Nasional Teknik Elektro

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

Vision Reformer Model To Enhance Transformer Efficiency for Face Recognition

Muhammad Dedi Saputra, Khairun Saddami, Nasaruddin Nasaruddin

Master of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia

ARTICLE INFORMATION

Received: June 2, 2025
Revised: April 26, 2026
Accepted: July 30, 2026
Available online: July 31, 2026

KEYWORDS

Face recognition, vision transformer, vision reformer, memory efficiency, accuracy

CORRESPONDENCE

Phone: +62 821 82164470
E-mail: nasaruddin@usk.ac.id

A B S T R A C T

The rapid advancement of digital technology has transformed attendance systems in educational environments. Conventional manual attendance methods are vulnerable to manipulation, while fingerprint-based biometric systems often experience limitations due to certain physical conditions. Facial recognition has therefore emerged as a reliable and non-contact biometric solution for attendance verification. However, existing deep learning architectures such as the Vision Transformer (ViT) require substantial computational resources and memory during training. This study proposes the Vision Reformer (ViR) model to enhance the computational efficiency of the ViT architecture for facial recognition. The proposed approach integrates the memory-efficient attention mechanism of the Reformer architecture into the ViT framework to improve training efficiency while maintaining high recognition performance. Experiments were conducted using the Indonesian Muslim Student Face Dataset (IMSFD) with learning rates of 1×10^{-3} and 1×10^{-4}. The experimental results show that the ViR model with a learning rate of 1×10^{-3} and 1×10^{-4} achieved the best performance, with accuracy of 94.63%, F1-score of 94.77%, precision of 95.89%, and recall of 94.63%, while requiring only 41.076 seconds of training time. In comparison, the ViT model achieved accuracy of 94.11% and required 82.368 seconds under the same configuration. Additionally, GPU memory analysis indicates that ViR reduces memory consumption by 78 MB and 39 MB at batch sizes of 16 and 32, respectively, while maintaining a stable and linear memory usage pattern. These results demonstrate that the proposed Vision Reformer model improves computational efficiency and memory utilization while maintaining competitive accuracy, making it a practical solution for facial recognition-based attendance systems in educational settings.

INTRODUCTION

Digital technology has profoundly influenced various sectors, including education [1]. In this context, attendance systems are essential for supporting effective teaching and learning processes [2]. While manual attendance remains susceptible to manipulation, such as proxy sign-ins, biometric methods like fingerprint recognition have minimized fraud due to the uniqueness of fingerprints [2]. However, fingerprint systems can be less effective for individuals with physical impairments or damaged fingerprints from daily activities. As a more reliable alternative, facial recognition technology leverages the uniqueness of facial features to overcome the limitations of fingerprint-based systems.

The accuracy of facial recognition technology has even reached levels comparable to fingerprint systems in certain cases [3], and its implementation in educational settings aligns with the need to improve efficiency and ensure compliance with regulations such as the Ministry of Education and Culture Regulation No. 15 of 2018 and the Circular Letter of the Ministry of Administrative and Bureaucratic Reform No. 16 of 2022 [4], [5]. Facial recognition systems leverage advanced algorithms to identify individuals based on facial patterns [6], and their application in education can enhance attendance efficiency and reduce fraudulent behavior.

Facial recognition research has become a key focus within the field of Computer Vision, where Deep Neural Networks (DNNs) have replaced manual feature-based methods and traditional machine learning approaches due to their superior accuracy and scalability [7], [8]. In educational institutions with large student populations, real-time processing capability in facial recognition systems is critical [9]. However, DNN models also face challenges regarding speed and memory usage when processing large-scale data.

The transformation in sequential data processing began with Long Short-Term Memory (LSTM) models equipped with attention mechanisms [9], [10], leading to the development of the Transformer architecture by Vaswani et al., which enables parallel processing through multi-head self-attention [11]. The Transformer architecture has become a dominant approach in modeling sequential and visual data since its introduction in Attention Is All You Need [11]. Although it excels at capturing global dependencies through self-attention, the model exhibits

https://doi.org/10.25077/jnte.v15n2.1350.2026

quadratic computational complexity $O(n^2)$), which limits its efficiency on large-scale data.

The development of the Vision Transformer extends its application to the computer vision domain, where images are divided into uniform patches and used as input tokens [12]. ViT is competitive even in complex image classification tasks; however, it still inherits efficiency limitations [11]. Various approaches have been proposed to improve efficiency, such as the Sparse Transformer [13], which employs restricted attention patterns but lacks adaptability, and Linformer [14], which assumes a low-rank structure to achieve linear complexity, though it may incur information loss.

Thus, there remains a need for an architecture that can integrate computational efficiency, global representation, and adaptability to data. Reformer offers a solution through the Locality-Sensitive Hashing (LSH) mechanism with O(n log n) complexity [15], which optimizes memory usage by replacing dot-product attention with locality-sensitive hashing (LSH), while also improving efficiency in the activation and feed-forward layers [15], and can be adapted into a Vision Reformer (ViR) for the image domain.

Although Reformer has demonstrated high efficiency in text-based tasks [15], empirical exploration of Vision Reformer in computer vision remains limited. Therefore, this study aims to develop a Vision Reformer (ViR) model that integrates the memory efficiency of the Reformer architecture with the visual representation capability of the Vision Transformer (ViT) in a face recognition system. The focus is on improving data-matching speed, analyzing its impact on accuracy, and evaluating the performance of this hybrid model in terms of processing speed, memory efficiency, and accuracy compared to the Vision Transformer model.

METHODS

A. Model Architecture

This study is experimental research aimed at developing and evaluating a facial recognition system based on the Vision Transformer (ViT) architecture [12], which is adapted using the Reformer algorithm [15], resulting in a new framework called Vision Reformer. ViT, which originates from the Transformer architecture [11] in Natural Language Processing (NLP), processes images by dividing them into uniformly sized patches as input tokens [12]. This approach enables effective spatial feature extraction for image classification [12]. Meanwhile, Reformer [15] is an advancement of the Transformer architecture [11], designed to improve memory and computational efficiency in processing large-scale sequential data [15]. Reformer integrates several key techniques: a. Locality-Sensitive Hashing (LSH) Attention, which reduces the attention computation complexity from $O(L^2)$ [9] to $O(L\ log\ L)$ [15], b. Reversible Residual Layers [16], which enable forward and backward propagation without storing the entire network activation, thereby conserving memory usage during training, and c. Chunking, a method of processing inputs in smaller segments, allowing the model to handle long sequences with lower memory overhead [15].

The study begins with a literature review on the Transformer architecture [11], Vision Transformer (ViT) [12], and Reformer [15]. ViT has previously been applied to image data by dividing images into patches [12], while Reformer has been predominantly used in text processing and has been only minimally explored in visual data, primarily focusing on performance evaluations related to speed and memory efficiency [15]. This research is conducted through several stages, including the selection of hardware and software, image data collection and preprocessing, training of the Vision Reformer model, and performance evaluation of the model on facial recognition tasks.

Figure 1 presents the flow diagram of a face recognition system based on the Vision Reformer architecture, which integrates the Vision Transformer (ViT) and Reformer for image processing in facial recognition tasks.

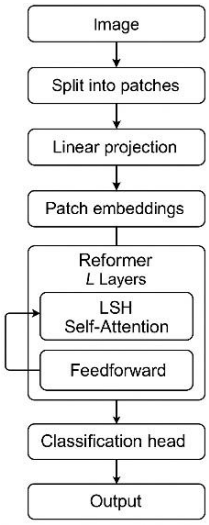

Figure 1. Model Workflow

The Vision Transformer (ViT) architecture [11] integrated with Reformer [15] processes fixed-size input images (224×224 pixels) by first dividing the image into smaller patches of (16×16 pixels). These patches are then flattened into vectors and linearly projected into a feature space through patch and positional embedding. The arranged patch embeddings are subsequently fed into the Reformer Encoder, which employs locality-sensitive hashing (LSH) [15] to replace the conventional self-attention mechanism [11] and simplifies CNN-based approaches through the use of reversible layers [16]. This approach aims to accelerate embedding matching while improving memory efficiency. After passing through the transformer encoder, the output tokens are extracted and passed to the MLP Head, a simple neural network responsible for classification. The MLP Head determines the final class of the image, such as class “A,” “B,” or “C.”

B. Experimental Setup

The image preprocessing stage in this study involves normalization and resizing of images to a fixed dimension of 224×224 pixels, applied to both the Vision Transformer (ViT) and Vision Reformer architectures. This process aims to standardize image size and pixel value scale to align with the input format required by Transformer-based models. Images originally varying in resolution were resized to a uniform size, as illustrated in Figure 2. This resizing step is essential to ensure

https://doi.org/10.25077/jnte.v15n2.1350.2026

input data consistency, computational efficiency, and compatibility with deep learning model architectures that process data in fixed structures and batches.

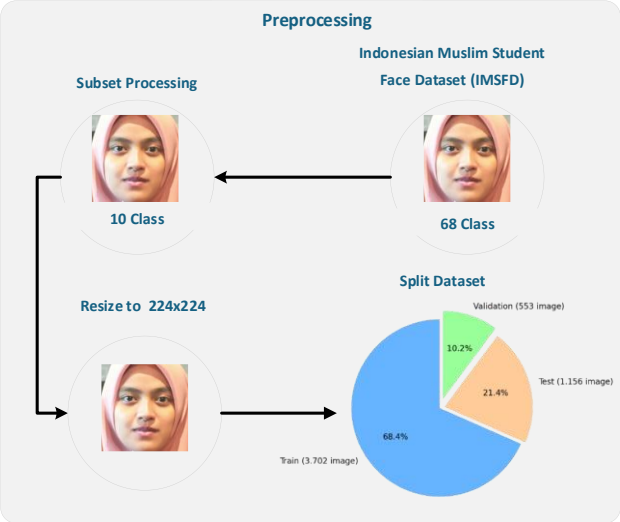

Figure 2. Dataset Workflow [17]

The dataset utilized is the validated Indonesian Muslim Student Face Dataset (IMSFD) [17], consisting of 21,048 labeled facial images. The dataset includes 68 classes, 47 male and 21 female, each representing an individual with approximately 400-600 facial images [17]. For this research, only 5 male classes (2,894 images) and 5 female classes (2517 images) were selected, based on the highest number of images per class (Table 1) [17]. This selection was made because each dataset class contains a varying number of samples and requires extensive data preparation and computational time. Quality versus quantity trade-offs necessitate strategic decision-making, where small data often leads to quicker, more accurate, and cost-effective insights [18], [19]. This is consistent with previous studies on ethnicity and race classification, which utilized 4 ethnic classes (African, Asian, Caucasian, Indian) [20], 7 feature-based classes (nose, skin, hair, eyes, eyebrows, back, mouth) [21], 10 student-level classes [22], 5 classes of bank employee facial images [23], and 10 classes of color facial datasets, each with 20 images showing different backgrounds and expressions [24].

Table 1. Subset of the Dataset.

Facial Image Labeling	Class
13020220341	0
13020220335	1
13020220334	2
13020220326	3
13020220321	4
13020220195	5
13020220191	6
13020220179	7
13020220175	8
13020220157	9

The Vision Transformer (ViT) model was trained using the original ViT architecture, while the Vision Reformer (ViR) model was developed by integrating the Reformer algorithm into the Vision Transformer framework to leverage its advantages in memory efficiency and localized attention mechanisms. Both models were trained using identical hyperparameter settings, with learning rates of 1×10^{-3} and 1×10^{-4}, to ensure a fair and consistent comparison of their performance in facial recognition classification tasks. Experimental evaluation was conducted on the Indonesian Muslim Student Face Dataset (IMSFD) [17], consisting of 10 image classes standardized to a resolution of 224 × 224 pixels. This standardization was implemented to ensure uniform input data, improve computational efficiency, and maintain compatibility with both the Vision Transformer and Reformer architectures, which require fixed input structures and batch sizes during the training process.

Table 2. Hyper-parameter.

Hyper-parameter	Value
Epochs	300-400
Optimizer	Adam
Patch Size	8
Batch Size	16
Learning rate	$1x10^{-3}$, $1x10^{-4}$
Image Size	224x224

The training process was conducted over 300 to 400 epochs with a batch size of 16. The optimizer employed was Adam, with learning rates set at 1×10^{-3} and 1×10^{-4} , as summarized in Table 2. Model training was performed using Amazon Web Services (AWS), while performance evaluation was carried out on the Google Colab platform. The use of these computational resources was intended to ensure adequate computing capacity and to support efficient training and evaluation of the model.

C. Tools

The hardware used in this study consists of a Lenovo ThinkPad T40S laptop equipped with an Intel(R) Core(TM) i5-5300U processor, CPU @2.30GHz 2.3 GHz, and 8192MB of RAM. The software components include the Windows 10 Pro 64-bit operating system and the Python programming language. Model training was conducted using Amazon Web Services (AWS) cloud infrastructure with an ml.p3.2×large virtual machine configuration, which includes 8 vCPUs, 61 GB of RAM, and an NVIDIA TESLA V100 GPU (16 GB VRAM). The trained models were automatically stored in Amazon S3 (Simple Storage Service) and later saved to local storage. Subsequently, model performance evaluation and analysis were conducted using Google Colab with a T4 instance. The virtual machine configuration included 4 vCPUs, 16 GB of RAM, and an NVIDIA T4 GPU (16 GB VRAM). The trained models were automatically saved to Google Drive and then transferred to local storage.

D. Evaluation Metrics

The evaluation was conducted to assess the effectiveness of the model in recognizing faces within the Indonesian Muslim Student Face Dataset (IMSFD). Model performance was assessed using the metrics of Precision (1), Recall (2), F1-Score (3), and Accuracy (4). Precision measures the proportion of relevant positive predictions, while Recall indicates the proportion of relevant instances correctly identified by the model. The F1-Score represents the balance between Precision and Recall, whereas Accuracy reflects the closeness between predicted and actual values. A confusion matrix analysis was employed to determine

the model's average accuracy, thereby assessing how effectively the model can recognize facial characteristics.

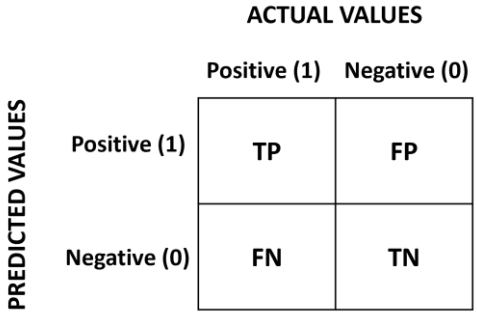

Figure 3. Confusion Matrix

$$Precision_{weighted} = \sum_{i=1}^{K} \left(\frac{n_i}{N} \times \frac{TPi}{TPi+FPi}\right) \times 100\% \quad (1)$$

$$Recall_{weighted} = \sum_{i=1}^{K} \left(\frac{n_i}{N} \times \frac{TPi}{TPi+FNi}\right) \times 100\% \quad (2)$$

$$F1 - Score_{weighted} = \sum_{i=1}^{K} \left(\frac{n_i}{N} \times \frac{2\ x\ precision_i\ x\ recall_i}{precision_i + recall_i}\right) \quad (3)$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (4)$$

The interpretation of the confusion matrix is essential for evaluating model performance. Its key components include True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which collectively provide a comprehensive overview of the model's accuracy and effectiveness in image recognition tasks.

RESULTS AND DISCUSSION

A. Optimizer Learning Rate 0,001 and 0,0001

Model validation was performed concurrently during the training process to comprehensively monitor model performance. The evaluation involved four primary metrics: train accuracy (1), validation accuracy (2), train loss (3), and validation loss (4). These metrics serve as primary indicators for evaluating the model's ability to learn patterns and generalize, as well as for detecting overfitting or underfitting to assess the effectiveness of the hyperparameter configuration.

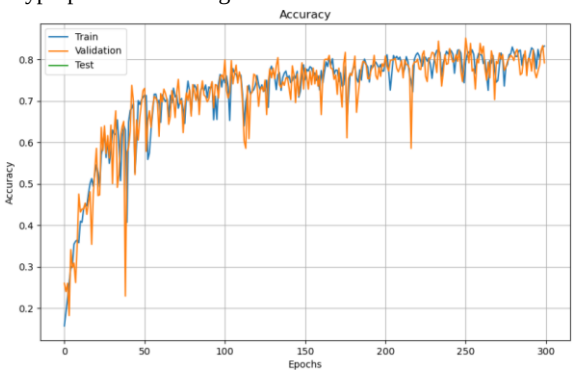

Figure 4. Validation Accuracy of the ViR Models

Figure 4 presents the validation accuracy of the ViR models at a learning rate of 1×10^{-3}. The ViR model achieved a validation accuracy of 79.20%.

Table 3. Training Results Using Learning Rate of 1×10^{-3}.

Lr	Model	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
1x10^{-3}	ViT	34.49%	37.43%	1.7990	1.7141
	ViR	83.25%	79.20%	0.4923	0.6738

Table 3 presents a comparison of the hyperparameter configurations used in training the ViT and ViR models, indicating that ViR significantly outperforms ViT at a learning rate of 1×10^{-3}. The ViR model achieved a training accuracy of 83.25% and a validation accuracy of 79.20%, along with low loss values of 0.4923 for training and 0.6738 for validation. In contrast, the ViT model exhibited suboptimal performance, with training and validation accuracies of only 34.49% and 37.43%, respectively, and loss values exceeding 1.7141. These findings indicate that integrating the Reformer into the ViR model enhances training efficiency, stability, and generalization capability, even with identical hyperparameter settings.

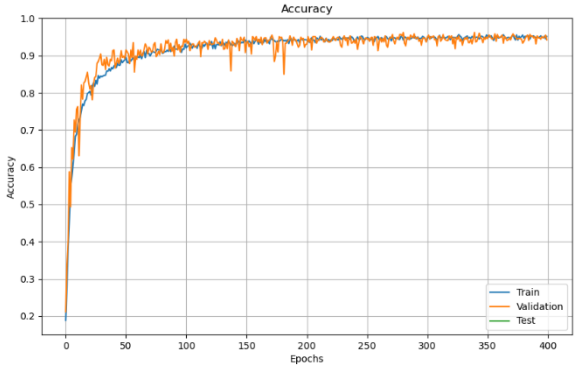

Figure 5. Validation Accuracy of the ViR Models

Figure 5 presents the validation accuracy of the ViR models at a learning rate of 1×10^{-4}. The ViR model achieved a validation accuracy of 94.21%.

Table 4. Training Results Using Learning Rate of 1×10^{-4}.

Lr	Model	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
1x10^{-4}	ViT	95.92%	91.32%	0.0989	0.3226
	ViR	95.16%	94.21%	0.1224	0.2016

Table 4 presents a performance comparison between the ViT and ViR models at a learning rate of 1×10^{-4}. The results indicate that ViR demonstrates superior generalization performance compared to ViT. This is reflected in ViR's higher validation accuracy of 94.21%, compared to 91.32% for ViT, and its lower validation loss (0.2016 versus 0.3226). Although ViT achieved a higher training accuracy of 95.92% and lower training loss (0.0989), the substantial discrepancy between training and validation results suggests potential overfitting. In contrast, ViR exhibited a better balance with a training accuracy of 95.16% and training loss of 0.1224, indicating greater stability and effectiveness in learning and generalizing patterns from the data. These findings reaffirm the superiority of the ViR architecture in optimizing model performance, even when using identical hyperparameter configurations.

https://doi.org/10.25077/jnte.v15n2.1350.2026

B. *Test and Evaluation Results*

Evaluation of ViT and ViR Models at a Learning Rate of 1×10^{-3}

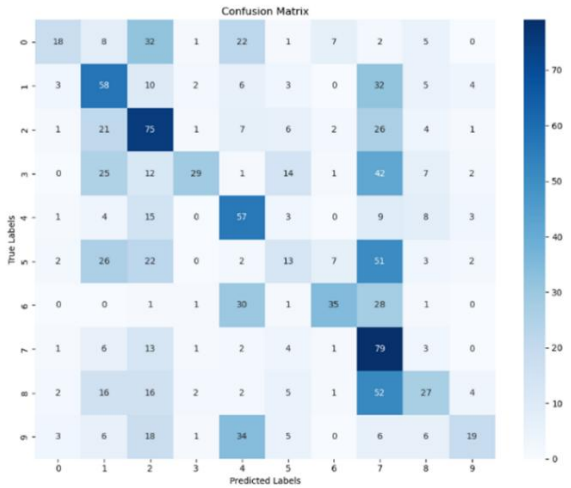

Figure 6. Confusion Matrix of the ViT Model

Based on Figure 6, the ViT model tested with a learning rate of 1×10^{-3} demonstrated its highest performance in class 2 (75 correct predictions) and class 7 (79 correct predictions), while its lowest performance was observed in classes 0 and 9, characterized by high misclassification rates and unfocused prediction distributions. The model exhibited a prediction bias toward class 7, which frequently appeared as the predicted label for various other classes. Significant misclassifications among classes 0, 2, and 3 suggest suboptimal feature representations and overlap between class characteristics. Overall, the confusion matrix indicates that the model struggles to effectively distinguish between certain classes. Table 5 presents the evaluation results of the ViT model tested with a learning rate of 1×10^{-3}

Table 5. Testing Performance Metrics of the ViT Model Using a Learning Rate of 1×10^{-3}.

Accuracy	F1-Score	Precision	Recall
35.46%	33.79%	43.87%	35.46%

Based on the test results presented in Table 5, the model's performance is still considered relatively low, with an accuracy of 35.46% and an F1-score of 33.79%, indicating an imbalance between precision (43.87%) and recall (35.46%). These results suggest that the model has not yet achieved optimal capability in detecting and classifying data, and further indicate a potential issue of underfitting.

Figure 7 presents the confusion matrix for the ViR model tested with a learning rate of 1×10^{-3}. Based on Figure 7, the ViR model demonstrates superior classification performance compared to ViT, achieving high accuracy in several classes, such as class 6 (132 correct predictions), class 4 (118), class 3 (107), and class 1 (105). The strong diagonal dominance in the confusion matrix reflects the model's effectiveness in distinguishing between class-specific features.

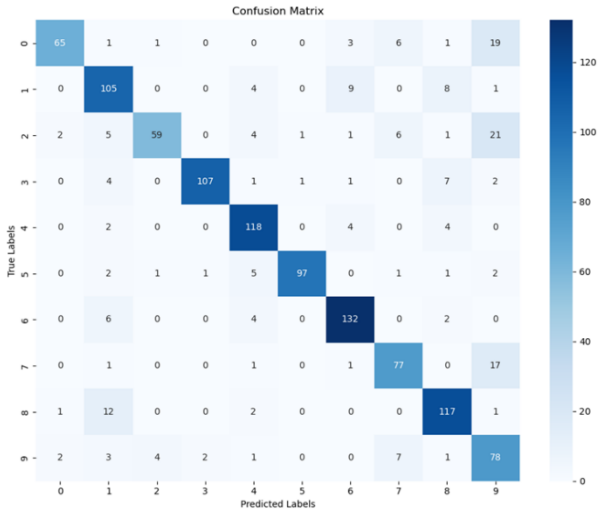

Figure 7. Confusion Matrix of the ViR Model

However, misclassifications are still present, particularly in class 0 and class 2, which show lower correct prediction counts and mispredictions dispersed across other classes, such as classes 7 and 9. These findings suggest that while the model performs well overall, challenges remain in accurately distinguishing classes with similar feature representations. Table 6 presents the evaluation results of the ViR model tested with a learning rate of 1×10^{-3}. This summary highlights the key evaluation metrics, reflecting the overall performance of the model based on the conducted testing.

Table 6. Testing Performance Metrics of the ViR Model Using a Learning Rate of 1×10^{-3}.

Accuracy	F1-Score	Precision	Recall
82.61%	82.78%	84.47%	82.61%

Based on Table 6, the ViR model demonstrates strong classification performance, with an accuracy of 82.61%, an F1-score of 82.78%, a precision of 84.47%, and a recall of 82.61%. These results indicate a balanced capability in detecting relevant instances while maintaining prediction reliability. Compared to the ViT model, ViR yields a higher number of correct predictions, lower classification errors, and a more concentrated distribution along the diagonal of the confusion matrix. Therefore, ViR can be considered more reliable for classification tasks within the context of this study.

Evaluation of ViT and ViR Models at a Learning Rate of 1×10^{-4}

The confusion matrix for the ViT model with a learning rate of 1×10^{-4} is presented in Figure 8. Based on Figure 8, the ViT model demonstrates high accuracy with a strong dominance of correct predictions along the diagonal, particularly for classes 0, 2, 3, 5, 6, 7, and 9. Classification errors are relatively low and dispersed; however, there is a notable tendency for significant misclassification in class 8, which is frequently predicted as class 7. Additionally, classes 1 and 4 also exhibit prediction shifts toward class 7. Overall, the confusion matrix indicates strong classification performance with some concentrated misclassification in specific classes. Table 4.5 presents the evaluation results of the ViT model with a learning rate of 1×10^{-4} based on the obtained testing metrics.

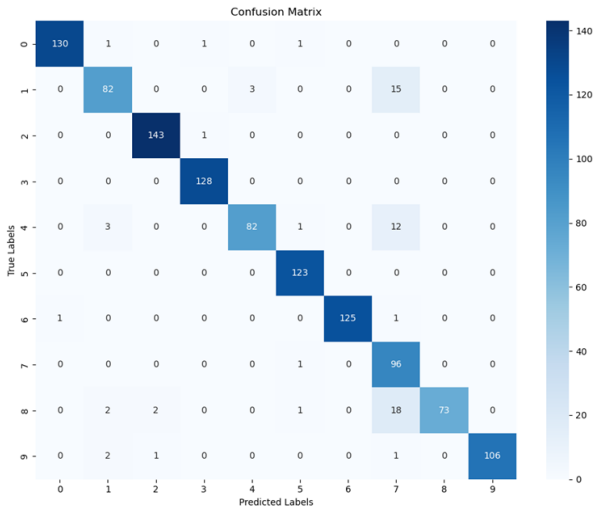

Figure 8. Confusion Matrix of the ViT Model

Table 7. Testing Performance Metrics of the ViT Model Using a Learning Rate of 1×10^{-4}.

Accuracy	F1-Score	Precision	Recall
94.11%	94.26%	95.32%	94.11%

Based on Table 7, the ViT model demonstrates superior performance with an accuracy of 94.11%, F1-score of 94.26%, precision of 95.32%, and recall of 94.11%, indicating a highly effective capability in facial recognition prediction.

The confusion matrix for the ViR model tested with a learning rate of 1×10^{-4} is presented in Figure 9.

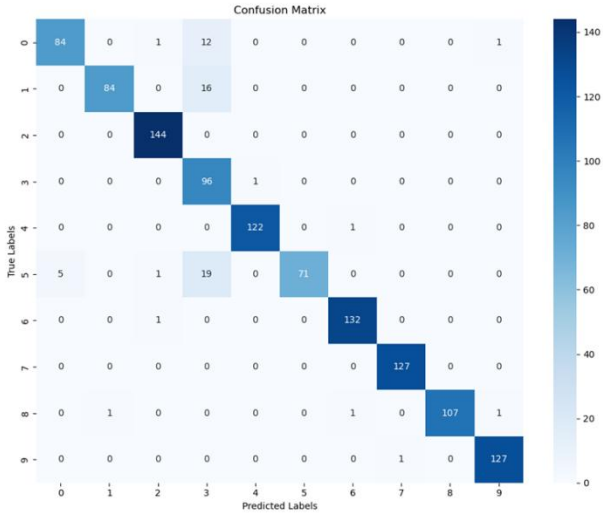

Figure 9. Confusion Matrix of the ViR Model

Based on Figure 9, the ViR model demonstrates strong facial recognition performance, as indicated by the high diagonal values in the confusion matrix, including perfect classification in class 2. Classes 6 and 7 also exhibit high accuracy, with 132 and 127 correct predictions, respectively. However, notable misclassifications remain, particularly with classes 0 and 1 frequently being predicted as class 3, and class 5 showing lower performance due to a significant number of instances being misclassified, primarily as class 3. Minor errors are also observed in several other classes. These findings suggest that while the model performs effectively overall, challenges persist in distinguishing between certain class-specific features.

Table 8. Testing Performance Metrics of the ViR Model Using a Learning Rate of 1×10^{-4}.

Accuracy	F1-Score	Precision	Recall
94.63%	94.77%	95.89%	94.63%

Based on Table 8, the ViR model demonstrates excellent performance, with an accuracy of 94.63%, an F1-score of 94.77%, a precision of 95.89%, and a recall of 94.63%. These results reflect the model's accuracy, consistency, and its strong ability to identify the majority of positive cases in a balanced and reliable manner.

The confusion matrix shows that ViR performs better and more consistently than ViT, with more correct predictions along the diagonal. In contrast, ViT has more misclassifications across classes. Therefore, ViR is the more reliable model in this study.

Table 9. Comparison of Previous Research.

Method	Amount of Data	Dataset	Accuracy
S-ViT [19]	90.000 images	AI Hub	98.57%
S-ViT [19]	507.645 images	CASIA/LFW	92.70%
S-ViT [19]	8.300 images	PubFig83	86.53%
ViT-B/16/S [25]	1.036.480 images	FER-2013, AffectNet, CK+48	52.25%
ViT-B/16/SG [25]	1.036.480 images	FER-2013, AffectNet, CK+48	52.42%
ViT-B/16/SAM [25]	1.036.480 images	FER-2013, AffectNet, CK+48	53.10%
ViT 1×10^{-3}	5.411 images	IMSFD	35.46%
ViR 1×10^{-3}	5.411 images	IMSFD	82.61%
ViT 1×10^{-4}	5.411 images	IMSFD	94.11%
ViR 1×10^{-4}	5.411 images	IMSFD	94.63%

Table 9 compares multiple methods across varying datasets and data scales. While S-ViT achieves high accuracy on large-scale datasets, its performance declines on smaller or more complex datasets. Similarly, ViT-based models trained on very large datasets exhibit relatively low accuracy, indicating limitations in generalization despite abundant data.

In contrast, ViR demonstrates consistently superior performance, even with a significantly smaller dataset (5,411 images). Under identical experimental conditions on the IMSFD dataset, ViR outperforms ViT at both learning rates (82.61% vs. 35.46% at 1×10^{-3} and 94.63% vs. 94.11% at 1×10^{-4}. These results highlight ViR's stronger feature representation, learning stability, and robustness, making it a more reliable and effective model, particularly in limited data scenarios.

C. Speed and Memory Performance Results

Profiling was conducted to evaluate model performance using PyTorch Profiler [26] and TensorBoard [27], with tracking of time and memory usage during training, as well as comparison of CPU and GPU resource utilization [27], [28]. The tests were

https://doi.org/10.25077/jnte.v15n2.1350.2026

performed on the training data using batch sizes ranging from 1 to 16 and patch sizes of 4, 8, 16, and 32 for both ViT and ViR models.

Batch Time Performance of ViT and ViR Models

The batch time analysis presents an evaluation of data processing efficiency, measured in samples per millisecond, based on batch size and patch size configurations.

Table 10. Batch Time Performance Results.

Batch Size	ViR Patch 4	ViT Patch 4	ViR Patch 8	ViT Patch 8	ViR Patch 16	ViT Patch 16	ViR Patch 32	ViT Patch 32
2	26.1	57.8	22.4	57.0	25.0	57.4	40.4	57.5
4	39.4	104.6	38.8	104.8	40.5	104.5	34.5	95.5
6	39.3	104.2	33.2	108.8	41.2	128.0	35.4	114.8
8	40.9	142.4	44.1	165.9	47.0	174.5	39.7	158.3
10	50.2	183.2	51.4	199.6	50.5	197.4	51.2	204.4
12	60.3	221.8	63.0	225.2	61.3	236.1	66.3	246.8
14	72.8	258.0	75.0	263.8	73.7	274.4	79.9	289.0
16	81.5	304.1	82.0	302.4	84.6	314.7	88.1	329.6

The results presented in Table 10, Batch Time, indicate that the ViR model consistently demonstrates superior processing time efficiency compared to the ViT model across various patch size and batch size configurations. ViT exhibits a significant and linear increase in processing time as batch size increases, whereas ViR shows more stable and lower processing times, particularly after overcoming initial overhead. This trend is clearly observed across all patch size variations (4, 8, 16, and 32), with ViR maintaining superior time efficiency. These findings highlight the computational advantage of the Vision Reformer (ViR) architecture over the Vision Transformer (ViT), making it more suitable for real-time biometric applications such as facial recognition in educational settings.

Throughput Performance of ViT and ViR Models

The following throughput analysis presents an evaluation of the amount of data that a system can process in samples per second.

Table 11. Throughput Performance Results.

Batch Size	ViR Patch 4	ViT Patch 4	ViR Patch 8	ViT Patch 8	ViR Patch 16	ViT Patch 16	ViR Patch 32	ViT Patch 32
2	76.5	34.5	89.1	35.0	79.9	34.8	49.4	34.7
4	101.3	38.2	102.8	38.1	98.7	38.2	115.8	41.8
6	152.6	57.5	180.3	55.1	145.5	46.6	169.0	52.2
8	195.1	56.1	181.3	48.2	170.1	45.8	201.0	50.5
10	199.0	54.5	194.2	50.0	197.7	50.6	195.0	48.9
12	198.7	54.0	190.2	53.2	195.5	50.8	180.7	48.6
14	192.1	54.2	186.6	53.0	189.9	51.0	175.1	48.4
16	196.0	52.6	195.0	52.9	189.0	50.8	181.5	48.5

Table 11 presents the throughput analysis, which shows that the ViR model consistently outperforms the ViT model across various patch size and batch size configurations./. At patch sizes 4, 8, and 16, ViR achieves optimal throughput exceeding 190–199 samples per second starting from mid-range batch sizes and remains stable up to batch size 16. At patch size 32 and batch size 8, ViR reaches a peak throughput of 201 samples per second. In contrast, ViT shows a slow and stagnant increase in throughput, remaining in the range of 50–58 samples per second. Although ViR experiences a slight decline in throughput at patch size 32, its performance remains significantly superior to that of ViT. Based on the analysis results, the throughput table reinforces the findings from the previous batch time graph, confirming that the Vision Reformer (ViR) is significantly more efficient and scalable than the Vision Transformer (ViT).

GPU Memory Performance of ViT and ViR Models

The following GPU analysis presents the maximum GPU memory usage of the ViT and ViR models across batch sizes ranging from 1 to 16 and four different patch size configurations (4, 8, 16, and 32).

Table 12. GPU Performance Results.

Batch Size	ViR Patch 4	ViT Patch 4	ViR Patch 8	ViT Patch 8	ViR Patch 16	ViT Patch 16	ViR Patch 32	ViT Patch 32
2	314	1556	252	367	237	289	237	276
4	395	2837	273	463	243	293	240	279
6	476	4120	294	560	249	305	243	283
8	558	5401	315	656	256	314	247	286
10	639	6685	337	752	262	325	250	289
12	720	7966	358	849	268	336	253	293
14	802	9247	379	945	274	347	256	296
16	883	10531	400	1041	280	358	260	299

Table 12 presents a comparison of GPU memory usage between the ViT and ViR models across various patch size configurations. At patch size 4, ViT consumes significantly more memory, increasing from 915 MB to 10,531 MB, while ViR rises modestly from 271 MB to 883 MB. A similar pattern is observed at patch size 8, with ViT reaching 1,041 MB compared to only 400 MB for ViR. The difference begins to narrow at patch sizes 16 and 32, where the memory usage gap between the two models is reduced to 78 MB and 39 MB, respectively. Additionally, memory growth in ViR remains stable and linear, indicating consistent resource allocation.

CPU Memory Performance of ViT and ViR Models

The following CPU analysis presents the maximum CPU memory usage of the ViT and ViR models across batch sizes ranging from 1 to 16 and four different patch size configurations (4, 8, 16, and 32).

Table 13. CPU Performance Results.

Batch Size	ViR Patch 4	ViT Patch 4	ViR Patch 8	ViT Patch 8	ViR Patch 16	ViT Patch 16	ViR Patch 32	ViT Patch 32
2	1009	1024	1191	1191	1197	1197	1201	1201
4	1049	1049	1191	1191	1197	1197	1201	1201
6	1055	1055	1191	1191	1197	1197	1201	1201
8	1056	1057	1191	1191	1197	1197	1201	1201
10	1076	1088	1191	1191	1197	1169	1201	1201
12	1124	1138	1191	1192	1177	1177	1209	1226
14	1164	1172	1192	1192	1170	1177	1226	1226
16	1180	1191	1192	1196	1187	1196	1227	1229

Table 13 presents the measurement results of CPU memory usage for the ViR and ViT models, which generally increase with larger batch sizes, with efficiency depending on the patch size. At a patch size of 4, ViR demonstrates greater efficiency, with memory usage ranging from 947 to 1180 MB, compared to 996 to 1191 MB for ViT. At a patch size of 8, memory consumption remains relatively stable and similar between the two models: 1191 to 1192 MB for ViR and 1191 to 1196 MB for ViT, indicating a nearly balanced CPU workload. At a patch size of 16, a slight decrease in memory usage is observed, with ViR decreasing from 1197 MB to 1187 MB, and ViT from 1197 MB to 1196 MB, suggesting more efficient memory management by ViR. Meanwhile, at a patch size of 32, ViR's memory usage increases from 1201 MB to 1227 MB, while ViT rises from 1201 MB to 1229 MB. Overall, ViR exhibits greater CPU memory efficiency compared to ViT, particularly under small and large patch size configurations, highlighting ViR's advantage in computational resource management for inference scenarios.

Comparison of Training Time and Output Field Size

Table 14 presents a comparison of training time and model field size between the Vision Transformer (ViT) and Vision Reformer (ViR), evaluated at two different learning rates (1×10^{-3} and 1×10^{-4}). The assessment includes the total training duration (in seconds) and the resulting model field size (in megabytes) after training for a specified number of epochs.

Table 14. Training Time and Training Field Size.

Lr	Model	Number of Epochs	Training Time/Second	Size Field/MB
$1x10^{-3}$	ViT	300	79344	81.46
	ViR	300	30600	34.18
$1x10^{-4}$	ViT	300	82368	81.46
	ViR	400	41076	34.18

The analysis indicates that the Vision Reformer (ViR) is more efficient than the Vision Transformer (ViT) in terms of training time and model field size. At a learning rate of 1×10^{-3}, ViR requires 30,600 seconds, significantly faster than ViT, which takes 79,344 seconds. ViR's efficiency is also maintained at a learning rate of 1×10^{-4} with an increased epoch count of up to 400, demonstrating ViR's advantage in time and resource utilization.

CONCLUSIONS

This study proposes the Vision Reformer (ViR), integrating Reformer into Vision Transformer (ViT) to improve efficiency in facial recognition. The main contribution is a Transformer-based model that reduces computational cost and memory usage while maintaining competitive accuracy. Results show that ViR outperforms ViT in accuracy, training time, and resource efficiency, with stable performance across configurations. Practically, ViR is suitable for real-world applications such as attendance systems in resource-limited educational settings. Its efficiency enables reliable and scalable deployment. Future work includes exploring other architectures, applying image enhancement, and developing lighter models for real-time implementation.

ACKNOWLEDGMENT

The author extends sincere gratitude to Universitas Syiah Kuala, particularly to the academic advisors and examiners, as well as to all parties who have contributed and provided support in the completion of this research.

REFERENCES

[1] S. Tools and B. B. Packages, *S E V E N T H E DI T ION and Behavior Learning*.

[2] D. I. Mulyana and Edi, "Penerapan Face Recognition Dengan Algoritma Viola Jones Dalam Sistem Presensi Kehadiran Siswa Dan Guru Pada Sekolah Idn Boarding School Jonggol," *J. Indones. Manaj. Inform. dan Komun.*, vol. 4, no. 3, pp. 1749–1757, 2023, doi: 10.35870/jimik.v4i3.398.

[3] J. Lunter, "Everyday biometrics: can face replace fingerprint recognition?," *Biometric Technol. Today*, vol. 2021, no. 4, pp. 7–10, 2021, doi: 10.1016/S0969-4765(21)00048-5.

[4] Permendikbud, "Peraturan Menteri Pendidikan dan Kebudayaan RI Nomor 15 Tahun 2018 tentang Pemenuhan Beban Kerja Guru, Kepala Sekolah dan Pengawas Sekolah," *Menteri Pendidik. dan Kebud. RI*, vol. 53, no. 9, pp. 1689–1699, 2018.

[5] Menteri PANRB, "Kewajiban Menaati Ketentuan Jam Kerja Bagi Aparatur Sipil Negara," *Menteri PANRB*, 2022.

[6] I. Fibriani, E. M. Yuniarno, R. Mardiyanto, and M. H. Purnomo, "ViTMa: A Novel Hybrid Vision Transformer and Mamba for Kinship Recognition in Indonesian Facial Micro-expressions," *IEEE Access*, vol. 12, no. October, pp. 164002–164017, 2024, doi: 10.1109/ACCESS.2024.3487180.

[7] R. Natsume, K. Inoue, Y. Fukuhara, S. Yamamoto, S. Morishima, and H. Kataoka, "Understanding fake faces," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11131 LNCS, pp. 566–576, 2019, doi: 10.1007/978-3-030-11015-4_42.

[8] D. S. Trigueros, L. Meng, and M. Hartnett, "Face Recognition: From Traditional to Deep Learning Methods," 2018, [Online]. Available: http://arxiv.org/abs/1811.00116

[9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 4, no. January, pp. 3104–3112, 2014.

[10] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," pp. 1–23, 2016, [Online]. Available: http://arxiv.org/abs/1609.08144

[11] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[12] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[13] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating Long Sequences with Sparse Transformers," 2017.

[14] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer : Self-Attention with Linear Complexity," vol. 2048, no. 2019, 2020.

[15] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: the Efficient Transformer," *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–12, 2020.

https://doi.org/10.25077/jnte.v15n2.1350.2026

[16] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, "The reversible residual network: Backpropagation without storing activations," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 2215–2225, 2017.

[17] Z. purnawansyah, purnawansyah; Wibawa, Aji Prasetya; Widyaningtyas, Triyanna; Haviluddin, Haviluddin; Darwis, Herdianti; Azis, Huzain; Ali, "Indonesian Muslim Student Face Dataset (IMSFD)," 2023. doi: 10.17632/f6f3y6ndgw.1.

[18] I. Hussain, R. Sushil, and K. Amir, *Breaking the data barrier : a review of deep learning techniques for democratizing AI with small datasets*, vol. 57, no. 9. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10859-3.

[19] G. Kim, H. M. Group, S. Korea, G. Kim, and G. Park, "S-ViT : Sparse Vision Transformer for Accurate Face Recognition S-ViT : Sparse Vision Transformer for Accurate Face Recognition," no. March 2023, 2026, doi: 10.1145/3555776.3577640.

[20] K. Huri, E. Omid David, and N. S. Netanyahu, "DeepEthnic: Multi-label ethnic classification from face images," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11141 LNCS, no. October, pp. 604–612, 2018, doi: 10.1007/978-3-030-01424-7_59.

[21] K. Khan, J. Ali, I. Uddin, and B.-H. Roh, "A Facial Feature Discovery Framework for Race Classification Using Deep Learning," *Under Rev. Comput. Mater. Contin. 2021*, 2021.

[22] D. P. Andini, Y. G. Sugiarta, T. Y. Putro, and R. D. Setiawan, "Sistem Presensi Kelas Berbasis Pengenalan Wajah Menggunakan Metode CNN," *JTERA (Jurnal Teknol. Rekayasa)*, vol. 7, no. 2, p. 315, 2022, doi: 10.31544/jtera.v7.i2.2022.315-322.

[23] M. Arsal, Bheta, Anggraini, "Face Recognition Untuk Akses Pegawai Bank Menggunakan Deep Learning Dengan Metode CNN," *TEKNOSI.v6i1.2020.55-63*, vol. 01, pp. 413–418, 2021.

[24] M. Athoillah, "Pengenalan Wajah Menggunakan SVM Multi Kernel dengan Pembelajaran yang Bertambah," *J. Online Inform.*, vol. 2, no. 2, p. 84, 2018, doi: 10.15575/join.v2i2.109.

[25] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER : Facial Emotion Recognition with Vision Transformers," 2022.

[26] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.

[27] A. Diwan, E. Choi, and D. Harwath, "When to Use Efficient Self Attention? Profiling Text, Speech and Image Transformer Variants," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2, pp. 1639–1650, 2023, doi: 10.18653/v1/2023.acl-short.141.

[28] D. Gyawali, "Comparative Analysis of CPU and GPU Profiling for Deep Learning Models," 2023, [Online]. Available: http://arxiv.org/abs/2309.02521

NOMENCLATURE

The symbols used in the following equations summarize the key components in evaluating the performance of the face recognition model.

TP_i, TN_i : True positive in class i, True negative in class i

FP_i, FN_i : False positive in class i, False negative in class i

K : Total number of classes

N : Total number of data points across all classes

n_i : The number of data samples in class i