khazanah informatika

Jurnal Ilmu Komputer dan Informatika

# Temporal Video Analysis for Identifying Traditional Malay Buildings Using Residual Network and Vision Transformer

**Sunardi[1], Sri Winiarti[2*], Abdul Fadlil[3]**

[*]correspondence: sri.winiarti@tif.uad.ac.id
[1,3]Electrical Engineering Department
Universitas Ahmad Dahlan
Yogyakarta, Indonesia, 55191
[2]Informatics Department
Universitas Ahmad Dahlan
Yogyakarta, Indonesia, 55191

**Abstract** The lack of digital documentation in preserving traditional Malay architecture faces serious challenges, especially with the modernization that slowly obscures the shape and authenticity of the building. Essential elements such as roof shapes, stage structures, and typical ornamental carvings are difficult to identify manually without special skills and considerable time. Malay architecture is an integral part of Indonesia's cultural heritage that needs to be documented systematically and digitally. Along with advances in Artificial Intelligence (AI) technology, traditional buildings' intense learning, identification, and classification can now be done automatically through video-based visual data processing. This study uses a video-based deep learning approach to develop and evaluate a classification system for traditional Malay buildings. Two types of architecture are used: Residual Network (ResNet) and Vision Transformer (ViT). The dataset in the form of videos of traditional buildings was collected from the Pekanbaru, Riau Province, then processed through frame extraction, spatial-temporal augmentation, and visual annotation, resulting in a total of 1,500 frames as training data. This study also presents a novel aspect by comparing the performance of five deep learning models: ResNet18, ResNet34, ResNet50, ResNet101 (CNN), and ViT based on self-attention. ViT, which is rarely used in traditional video-based architecture, shows competitive accuracy and proves its effectiveness in understanding global visual relationships. The training method is carried out using supervised learning and evaluated based on classification accuracy. The test results show that all models can accurately identify visual features of Malay architecture. ResNet50 recorded the highest accuracy (100%), followed by ResNet18 (96.0%), ResNet101 (94.9%), ResNet34 (93.9%), and ViT (93.9%). These findings strengthen the potential for utilizing deep learning in cultural preservation through a video-based automatic documentation system.

**Keywords:** *Traditional Malay Architecture, Video Classification, Residual Network, Vision Transformer, Deep Learning*

***Article info:*** *submitted April 24, 2025, revised February 19, 2026, accepted March 3, 2026*

## 1 Introduction

The application of Artificial Intelligence (AI) technology in the recognition of traditional buildings, especially Riau Malay architecture, has excellent potential in efforts to preserve culture and develop tourism. Traditional buildings can be identified and documented automatically and efficiently using AI-based video classification techniques.

Indonesia is rich in cultural heritage, including traditional buildings that reflect local wisdom and high his-

torical values. One example is traditional Malay architecture, which has spread across Sumatra and its surroundings. These buildings function as residences or centers of social activities and represent a community's cultural identity. However, along with the increasing flow of modernization and the lack of systematic documentation, many traditional buildings have been damaged or lost without being optimally identified and preserved [1], [2].

Identification of traditional Riau Malay buildings in the context of architecture faces various challenges that hinder preservation and documentation efforts. The main problems include the lack of standard documentation and digitalization, where many buildings have not been documented in digital format, both in terms of structure and aesthetic elements [3], [4]. In addition, modernization and changes in materials have caused the loss of authenticity of traditional architectural forms, so an automatic identification system is needed that can recognize the original character of the building [5], [6]. The complexity of Malay architectural elements such as pyramid roofs, carvings, and stage structures also complicates the manual classification process, which requires time and high expertise [7], [8]. On the other hand, the lack of specific expertise in traditional architecture among architects and students is an obstacle to research and preservation. For this reason, applying AI technology, primarily through video classification models based on Residual Network (ResNet) and Vision Transformer (ViT), is a potential solution. This technology is not only able to accelerate the process of identification and documentation of traditional buildings, but also supports more accurate digital archiving and reconstruction [9], [10].

The selection of deep learning models in this study is based on the need to identify Malay traditional buildings accurately and efficiently from video data. The ResNet18, ResNet34, ResNet50, and ResNet101 models were chosen because they are reliable Convolutional Neural Networks (CNN) architectures in image classification, with the advantage of residual learning that can overcome vanishing gradients and capture complex visual features in stages [11]. The depth variation in these models provides flexibility between processing speed and accuracy in recognizing architectural details of traditional buildings. Meanwhile, ViT was chosen because it uses a self-attention mechanism that allows for comprehensive and practical visual pattern recognition to handle variations in shape, ornamentation, and multi-label classification of Malay buildings rich in visual elements [9], [12]. The combination of the ResNet and ViT approaches in this study aims to compare the performance of each architecture and optimize an AI-based video identification system that can support the digital preservation of Malay traditional architecture [13].

The application of AI technology, especially in the form of machine learning and deep learning, has grown rapidly and shown great potential in the field of pattern recognition and visual classification, including in the context of architectural form recognition [14]. In previous studies, AI has been successfully used to identify architectural elements such as building facades, geometric structures, and design styles with fairly high accuracy [15]. However, most of these studies are still focused on modern buildings. At the same time, the application of AI in the context of traditional architecture, especially those with high historical and cultural value, is still limited. The main challenge in developing an AI-based traditional building identification system is the lack of representative datasets containing images, videos, and metadata of traditional buildings from various regions. In addition, not many AI classification models currently available consider the spatial and temporal aspects inherent in the transformation of traditional architecture [16], [17]. The integration of spatial-temporal information (spatiotemporal augmentation) is very important for understanding the dynamics of changes in traditional buildings from time to time, both in terms of structure, materials, and social function [14].

The application of AI in architectural heritage recognition has gained significant attention in recent years. Traditional Riau Malay buildings have distinctive architectural features that reflect cultural heritage and historical value. Automatic recognition using AI-based video classification techniques, especially with deep learning models such as ResNet and ViT, offers a promising solution for digital documentation, preservation, and classification of these structures [18]-[20]. This literature review examines previous studies on AI-based architectural recognition, the effectiveness of deep learning models, and the significance of multi-label classification in preserving architectural heritage. Computer vision techniques have been widely used to analyze and classify architectural structures. Traditional methods rely on manually crafted feature extraction, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) [21]. However, these techniques often struggle with large-scale datasets and variations in architectural designs. The rise of deep learning, especially CNN, has improved the accuracy in recognizing and categorizing architectural elements [22].

## 2 Methods

This study aims to classify traditional Malay architecture using video data through a deep learning approach. The process begins with collecting videos of traditional buildings in Pekanbaru, Riau Province, followed

by frame extraction, preprocessing (resizing, normalization, and augmentation), and manual and automatic labeling. The extracted images are divided into three main classes: Riau Malay, Kalimantan Malay, and Non-Malay. The research method's several main stages, as shown in Figure 1.

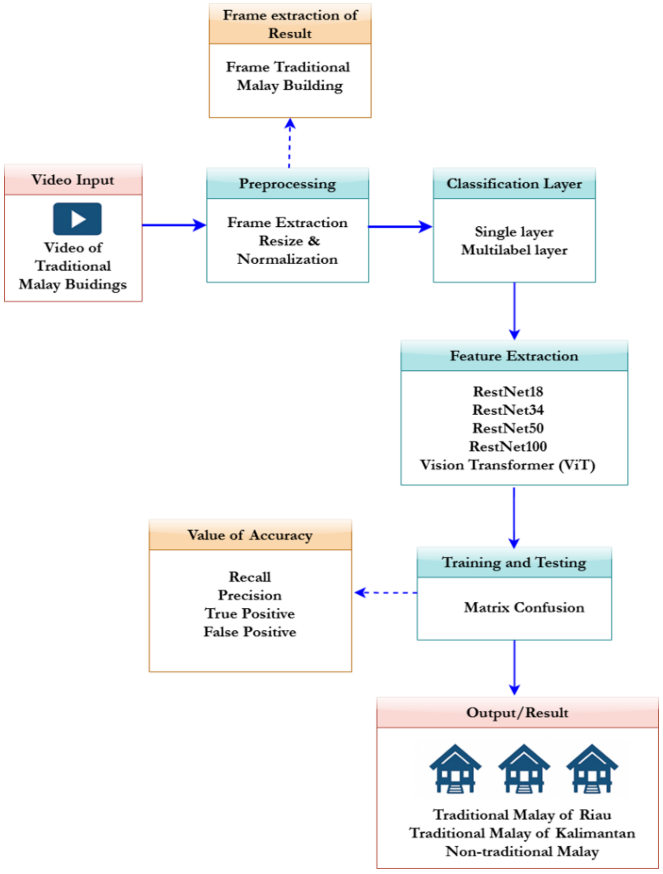


**Figure 1. Research method in identifying traditional Malay buildings based on video data**

The research method shown in the diagram starts from data collection through videos of traditional Malay buildings. This video data is then processed through a preprocessing stage, namely by performing frame extraction to take images per frame from the video, followed by resizing and normalization so that the image has a uniform pixel size and distribution to enter the next stage. After that, the processed image will go through a feature extraction stage, where deep learning models such as ResNet18, ResNet34, ResNet50, and ViT extract critical visual features from architectural elements of the building. These features then enter the classification layer for single-label classification (one type of label per image) and multi-label (more than one label if the building has mixed characteristics), all of which are implemented with the help of the ViT architecture. The final stage of this process is evaluation and output, where the model provides results in the form of a classification of traditional Malay buildings based on their shape, structure, and visual ornaments. These results support digital documentation and cultural preservation and open up opportunities for the development of automated systems to classify and educate traditional architecture more broadly. The evaluation method in this study was carried out to measure the performance of the classification model in recognizing images of traditional Malay buildings that had been extracted from videos. The evaluation used several standard metrics in machine learning: accuracy, loss, and correlation analysis between training parameters. First, validation accuracy measures how accurately the model classifies building images into the correct category based on ground truth labels. A high accuracy value indicates that the model can reliably identify building characteristics. Second, validation loss and training loss are monitored to evaluate the learning process of the model. Low loss indicates that the model does not experience overfitting and can generalize well. Furthermore, training parameter experiments (number of epochs, learning rate, dropout rate) were also carried out in stages to find the optimal combination. Visualizations such as accuracy and loss graphs per experiment and correlation matrices between variables are used to analyze the relationship between parameters and model performance. For example, it was found that the dropout rate has a strong negative correlation with validation accuracy (-0.72), and a positive correlation with loss, indicating that a dropout rate that is too high can reduce model performance. This comprehensive evaluation approach selects the best model based on the combination of parameters that produces the highest validation accuracy and the lowest loss.

## 3 Result

### *3.1 Data Collection*

This study uses a multi-label classification approach to classify traditional Malay building images using various deep learning methods, namely ResNet18, ResNet34, ResNet50, ResNet101, and ViT. The video data obtained is processed through frame extraction, resizing, and normalization to produce an image measuring 640 x 640 pixels before entering the model. One of method used when performing data augmentation is shown in Figure 2.

### *3.2 Pre-processing of Malay Traditional Building Data*

The stages in data processing are as follows:

1. Changing Video data to Frame by frame. Video data is changed frame by frame. The maximum

**Figure 2. Data augmentation using grayscale techniques**

duration of the video used is 2 minutes. After the data is changed to a frame, annotation is carried out to carry out the data labeling process, which can be done manually or automatically.

2. Labeling. The data labeling process can be done manually and automatically. This study uses 3 classes, namely Riau Malay Traditional Buildings, Kalimantan Malay Traditional Buildings, and Non-Malay Traditional Buildings, with labels: ornaments, roof shapes, and shapes of the entire building.

3. Split Dataset. After obtaining a dataset of 162, the data will be split into 70% training data, 20% validation data, and 10% testing data. In this case study of Malay traditional building identification, the training data.

4. Augmentation. The augmentation process is carried out to obtain feature extraction that matches the building object to be identified. In this case study, the Augmentation methods used are: Flip, Rotation, Cropping, Sher, Grayscale, Bluer, and Saturation.

5. Create a matrix data transformation. By selecting the size of the data to be used. In this case study, a size of 640 x 640 is used to create a snapshot of the data at a time with the transformation applied. Furthermore, the data will be ready to make an architectural model.

6. Training the architecture model. The next step is to create an Architecture Model from the labeled dataset using the Restnet10, Restnet25, Restnet50, Restnet101, and Vit Multiclass methods by selecting the custom train menu.

### *3.3 Create Model Architecture*

This study uses the ResNet and ViT architecture models. The reason for choosing the ResNet variant can be adjusted based on specific needs regarding performance, training efficiency, and the complexity of the Malay traditional building images being analyzed. ResNet50 or ResNet101 provides more accurate results for large and varied datasets, while ResNet18 and ResNet34 are still suitable for initial development or light deployment on edge devices. The reason for using the ViT model in this study is that ViT is very suitable for identifying traditional buildings because it can capture global spatial relationships between image elements (such as roof structures and distinctive ornaments), without being limited to local areas like traditional CNNs. This allows the model to distinguish more complex and comprehensive cultural patterns.

#### *3.3.1 ResNet Model*

ResNet is a CNN architecture designed to build deep networks without experiencing performance degradation due to vanishing gradients. In the Malay traditional building identification project, architecture variants such as ResNet18, ResNet34, ResNet50, or ResNet101 can be used, depending on the level of model complexity and the availability of computational resources. The model-building process begins by loading a ResNet model that has been pre-trained on ImageNet as transfer learning (optional), then modifying the output layer (fully connected layer) to match the number of Malay building classes, for example, with the syntax:

```
model.fc = nn.Linear(in_features=2048,
↪  out_features=num_classes).
```

Next, the model is trained using a prepared dataset with an optimization algorithm such as Adam or SGD. After training, the model is evaluated using accuracy, precision, and recall metrics to assess the classification performance against the test data.

In the context of Malay traditional building identification, the selection of ResNet18, ResNet34, ResNet50, and ResNet101 architectures is very relevant because each offers different levels of network depth and complexity, which can be adjusted to the visual characteristics of Malay buildings and the availability of computational resources. Malay traditional buildings have rich visual characteristics such as wood carving ornaments, typical roof shapes such as pyramids or gonjong, and unique pillar structures, all of which require the model's ability to recognize local features and complex patterns. The ResNet architectural model for identifying Malay traditional buildings is shown in Figure 3.

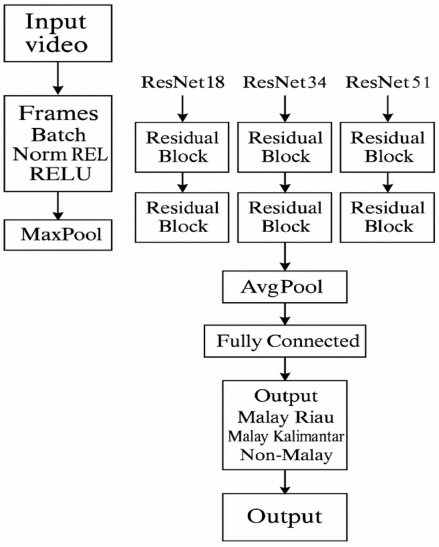


**Figure 3. ResNet architectural model in identifying traditional Malay buildings**

The diagram shown in Figure 3 illustrates the workflow of the modified ResNet architecture to classify Malay traditional buildings using video data as input.

The process starts from video input, which is broken down into individual image frames. The input video $V$ consists of several frames calculated by equation (1):

$$V = \{x_1, x_2, ..., x_T\}, \quad x_t \in \mathbb{R}^{H \times W \times C} \tag{1}$$

Each $x_t$ the frame is processed through: Resize to $640 \times 640$ with Normalization:

$$x'_t = \frac{x_t - \mu}{\partial} \tag{2}$$

Each frame undergoes preprocessing processes, such as resizing (e.g., to $640 \times 640$ pixels), normalization, and data augmentation to improve the model's generalization ability. After that, the processed frames are sent to the ResNet backbone (such as ResNet18, ResNet34, ResNet50, or ResNet101), which functions to extract deep visual features through convolutional layers and residual blocks. The obtained features can then be averaged or processed temporally, then fed into a fully connected layer (FC) that has been adjusted to produce three output classes: Riau Malay, Kalimantan Malay, and Non-Malay. The processed frame $x'_t$ is fed into the initial convolutional layer:

$$Z_0 = \text{ReLU}(\text{BN}(W_0 * x'_t + b_0)) \tag{3}$$

where ReLU is searched using the equation [23]:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \ll 0 \\ x & \text{if } x > 0 \end{cases} \tag{4}$$

Meanwhile, max pooling is calculated using the equation [24]:

$$z_1 = \text{MaxPool}(z_0) \tag{5}$$

By using average pooling:

$$z_{\text{gap}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} z_L(i,j) \tag{6}$$

For a fully connected layer, you can use the equation:

$$z_{\text{fc}} = W_{\text{fc}} \cdot Z_{\text{gap}} + b_{\text{fc}}, \quad z_{\text{fc}} \in \mathbb{R}^3 \tag{7}$$

In the final stage, the Softmax activation function is used to generate the probability of each class, both for each frame and the entire video segment. The logit values from the FC Layer are converted into probabilities through the softmax function [25], [26]:

$$P(y = k \mid x) = \frac{e^{z^k_{\text{fc}}}}{\sum_{j=1}^{3} e^{z^k_{\text{fc}}}}, \quad k \in \{1, 2, 3\} \tag{8}$$

The RestNet stages in a complete mathematical formulation can be seen in Figure 4.

This pipeline enables the system to accurately identify traditional architectural styles from video inputs by leveraging the representational power of the ResNet network. ResNet18 and ResNet34 tend to be lighter and faster to train, suitable for identifying simple patterns such as roof shapes or basic building structures. These models are efficient when used on datasets that are not too large or when trained on devices with GPU limitations.

ResNet50 and ResNet101, with their deeper layers, can capture more complex and detailed visual features such as ornamental details or building material textures. This makes them ideal for recognizing subtle differences between types of Malay traditional buildings that may appear similar globally but have significant local differences.

Thus, the selection of ResNet variants can be adjusted based on specific needs regarding performance,

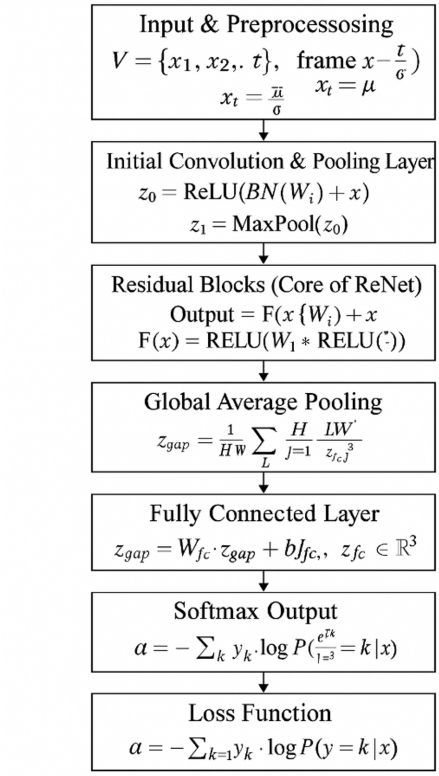


**Figure 4. ResNet stages with mathematical formulation**

training efficiency, and the complexity of the Malay traditional building images being analyzed. ResNet50 or ResNet101 provides more accurate results for large and varied datasets, while ResNet18 and ResNet34 are still suitable for initial development or light deployment on an edge device.

### *3.3.2 ViT Model*

Another architectural model used to identify traditional Malay buildings is the ViT. The reason for using the ViT architectural model method is shown in Figure 5.

Figure 5 illustrates the ViT model workflow to classify building images into three cultural categories: Riau Malay, Kalimantan Malay, and Non-Malay. The following are the stages in the architecture:

**Input Image** A building image (size 640×640 pixels) is entered as input. Using equation 1, the input image $x \in \mathbb{R}^{H\times W\times C}$ is divided into $N$ patches of size $P \times P$, so that [9]:

$$N = \frac{H \times W}{P^2} \tag{9}$$

Then, each patch is flattened and projected onto a fixed-

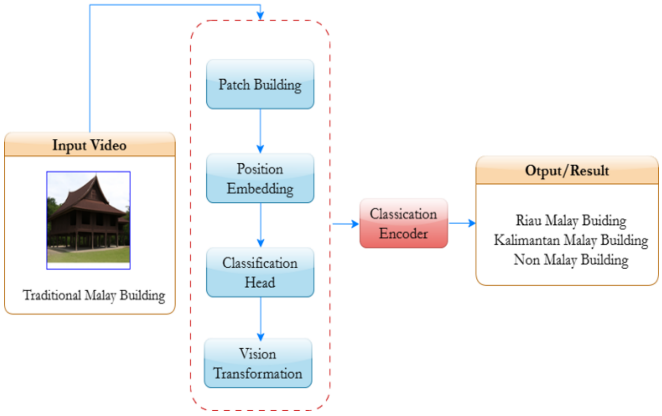


**Figure 5. Architectural Model of Traditional Malay Buildings using ViT**

dimension vector $D$ using linear projection $E$:

$$Z_0^i = E \cdot x^i + E_{pos}^i, \quad \text{for } i = 1, \ldots, N \tag{10}$$

Where $E_{pos}$ is the positional embedding added to make the model pay attention to spatial order.

**Patch Embedding** The image is broken down into several small patches, for example, 16 × 16 pixels per patch. Each patch is represented as a vector and then flattened into a linear sequence.

**Position Embedding** Since the transformer does not take position order into account naturally, spatial position information is added to each patch to keep the image structure in mind.

**Transformer Encoder Layers** Each encoder layer has two main components: Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN). These patches are then processed by several encoder layers consisting of:

*MHSA:* to capture the relationship between patches of the entire image [27]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{11}$$

Where:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{12}$$

$d_k$ is the key dimension that MHSA combines several attention heads [9], [28]:

$$\text{MHSA}(X) = [\text{head}_1; \ldots; \text{head}_h]W^0 \tag{13}$$

*Feed Forward Neural Networks (FFN):* to strengthen feature representation [29]:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{14}$$

*Normalization and residual connection:* are used to maintain training stability. Each sub-block is equipped with a residual connection:

$$X' = \mathrm{LN}(X + \mathrm{MHSA}(X)) \tag{15}$$

$$X'' = \mathrm{LN}(X' + \mathrm{FFN}(X')) \tag{16}$$

*Class Token*: a special token is added at the beginning of the patch sequence and represents the entire image. This token will be developed during the encoder process and used as the final input for classification. A class token $Z_0^{(0)} \in \mathbb{R}^D$ is added to the beginning of the patch sequence, thus becoming [28]:

$$Z_0 = [Z_0^{(0)} + Z_0^{(1)} + Z_0^{(2)} + \cdots + Z_0^{(N)} + E_{pos}] \tag{17}$$

*MLP Head*: the final class token is passed to the Multi-Layer Perceptron (MLP) [18], [30], which produces 3 class outputs: Riau Malay, Kalimantan Malay, and Non-Malay.

After the $L$ transformer layer, the class token $Z_L^{(0)}$ is used as the final representation and enters the classifier:

$$\hat{y} = \mathrm{softmax}(W_c Z_L^{(0)} + b_c) \tag{18}$$

Figure 6 shows the stages of the ViT and the mathematical formulation used to identify traditional Malay buildings.

### *3.4 Experiment and Model Evaluation*

The test results show that the ResNet50 method produces the highest accuracy, 99.2%, outperforming other tested methods. This shows that the ResNet50 architecture can effectively capture visual patterns and characteristics of traditional Malay building images. Furthermore, ResNet18 recorded an accuracy of 98.0%, followed by ResNet101 with 99.1%, and ResNet34 and ViT each with an accuracy of 99.4%.

Although ViT is known to excel in complex visual classification, its performance on this dataset has not surpassed ResNet50, which has proven more efficient in extracting local features typical of Malay architecture.

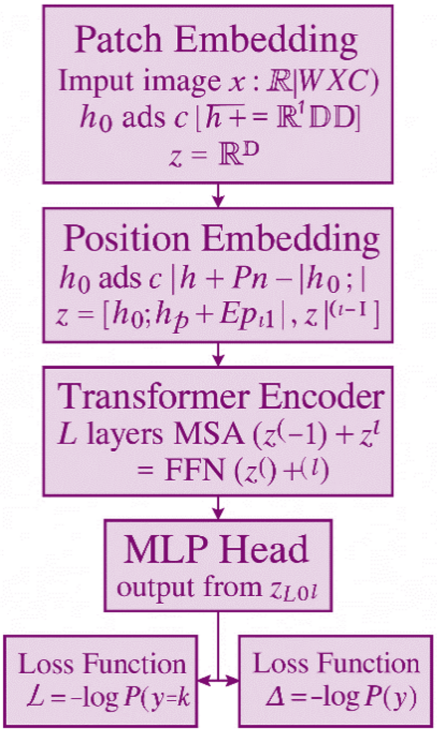


**Figure 6. shows the stages of the ViT in the mathematical model**

These results show that CNN models such as ResNet are still very relevant and competitive in classifying traditional building images. The stable but not outstanding performance of ViT indicates that this architecture still needs further optimization for medium-sized datasets.

Although ResNet101 is a deeper model than ResNet50, its accuracy results are not superior, indicating that increasing network depth is not always directly proportional to performance if not accompanied by increased efficiency in feature representation. In addition, ViT, although known for its ability to handle spatial and global context of images, in this study showed performance equivalent to ResNet34, but has not exceeded the performance of ResNet50. These results show that classic CNN architectures, such as ResNet, are still very competitive and superior in specific image classification tasks compared to traditional architectures. This study proves that choosing the right model significantly affects classification accuracy, and ResNet50 is the main recommendation in the context of the dataset and scenario used. Model Training Results. After the testing process is carried out, the test results will be displayed. In this case study, 5 architectural models were used to test the Riau Malay traditional building model. All experiments were conducted with 640 × 640 pixel inputs to ensure a fair comparison

between architectures. Despite being computationally light, ResNet18 and ResNet34 showed excellent accuracy (0.98 and 0.985, respectively). Meanwhile, ResNet50 and ResNet101 provided improved accuracy (0.992 and 0.991) due to their greater network depth, which allows for more complex feature extraction. ViT recorded the highest accuracy of 0.994, benefiting from its self-attention mechanism that can highlight the most informative parts of the image without the local biases common in convolutions. Overall, the results show that the deeper or more complex the architecture used, the higher the accuracy achieved, with ViT leading due to its global attention capability.

Table 1. Accuracy Test Results with 5 Models

| Method | Data Matrix Size | Accuracy |
|---|---|---|
| Restnet18 | 640 x 640 | 0.980 |
| Restnet34 | 640 x 640 | 0.985 |
| Restnet50 | 640 x 640 | 0.992 |
| Restnet101 | 640 x 640 | 0.991 |
| ViT | 640 x 640 | 0.994 |

The visualization of testing for the RestNet and ViT models is shown in Figure 7. The comparison of the classification accuracy of five deep learning models in identifying traditional Malay buildings, namely ResNet18, ResNet34, ResNet50, ResNet101, and ViT, with a multi-label approach. The graph shows that the ResNet50 model achieves the highest performance with an accuracy of 99.2%, followed by ResNet18 with an accuracy of 98.0% and ResNet101 with 99.1%. Meanwhile, ResNet34 and ViT Multi-label have the same accuracy, namely 99.4%. These results indicate that the deeper the ResNet architecture (up to a certain point), the more the performance tends to increase, but not always linearly, as seen from the decrease in accuracy in ResNet34 and ResNet101 compared to ResNet50. The ViT model, although modern and powerful for many visual tasks, in the context of Malay traditional building identification shows slightly lower results than ResNet50, possibly due to the need for more extensive training data or the specific visual characteristics of Malay traditional buildings that are better suited to be identified by convolutional models. This finding suggests that CNN architectures such as ResNet50 are still very effective in image-based classification tasks for specific datasets like this.

### *3.5 Discussion*

This study confirms that model selection significantly affects classification accuracy. The superior performance of ResNet50 highlights the importance of selecting a model appropriate for the dataset's visual complexity and scale. Although ViT has not surpassed ResNet50, its ability to

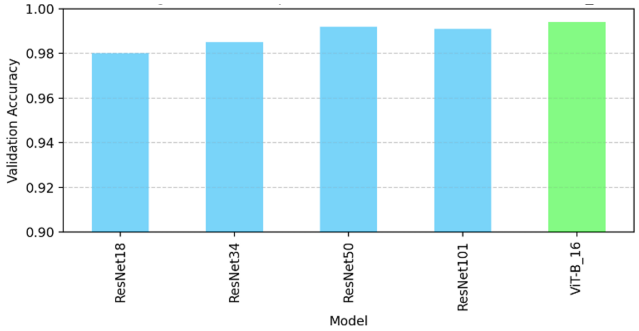


**Figure 7. Visualization results of the accuracy test of traditional Malay buildings with five architectural models**

model global patterns remains valuable, especially when combined with sufficient training data and the right fine-tuning strategy. ResNet and ViT methods in identifying Malay buildings are based on their respective architectural strengths in recognizing complex visual patterns.

ResNet is a convolutional model that effectively extracts spatial features from images, such as carving patterns, roof shapes, and typical ornaments on Malay buildings. With residual connections, ResNet can build a very deep network without experiencing accuracy degradation problems, making it suitable for detecting subtle and distinctive architectural details. Meanwhile, ViT works by dividing the image into small patches and analyzing them as sequences, like in Natural Language Processing (NLP), allowing this model to capture global relationships between parts of the image. This is very useful in identifying the overall features of Malay building structures, which often have symmetrical patterns and visual relationships between building elements. The combination of the two provides a complementary approach. ResNet excels in local details, while ViT is strong in understanding global context, making both very relevant for classifying traditional architectural images such as Malay buildings.

The results of the training and testing carried out in the performance test of the ResNet and ViT models can be seen in Table 2.

Based on the accuracy data shown in the figure, it can be seen that the five deep learning models used, ResNet18, ResNet34, ResNet50, ResNet101, and ViT, show very good performance in classifying Malay traditional buildings. Here is a more detailed explanation:

1. ViT has the highest accuracy, 99.4%, showing its superiority in capturing spatial relationships and visual context between image parts through self-attention. This indicates that ViT is very effective in traditional architectural image-based classification tasks.

Table 2. Matrix Test Results

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Riau Malay | 0,986301 | 0,986301 | 0,986301 | 438 |
| Kalimantan Malay | 0,991197 | 0,984266 | 0,987719 | 572 |
| Non Malay | 0,991903 | 1 | 0,995935 | 490 |

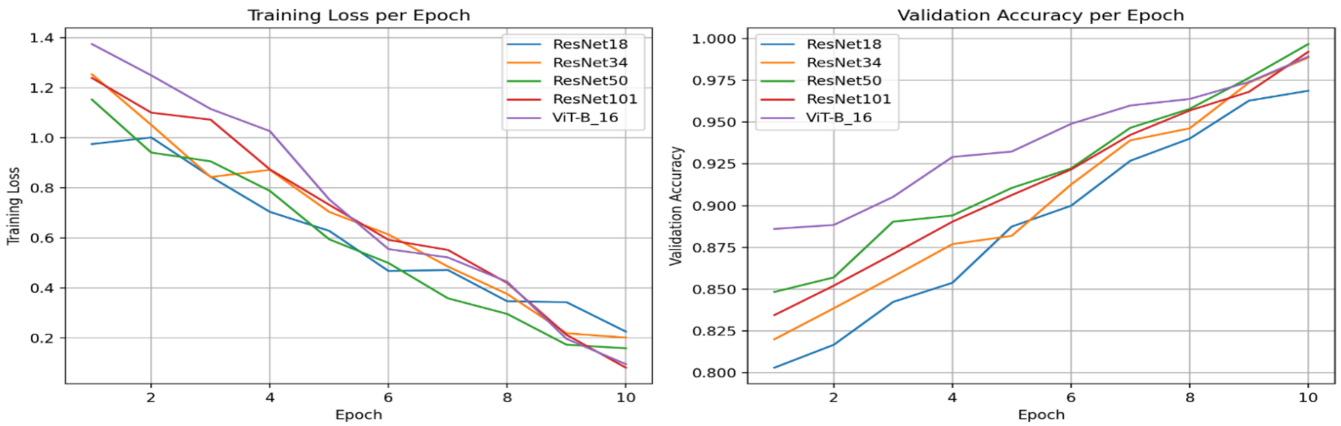


**Figure 8. Visualization of training and validation accuracy using ResNet and ViT Models**

2. ResNet50 and ResNet101 also show high performance, with accuracies of 99.2% and 99.1%, respectively. This reflects that models with medium to high depth can effectively extract features from Malay building images.

3. ResNet34 has an accuracy of 98.5%, slightly higher than ResNet18 (98.0%), indicating that increasing network depth provides performance improvements, although not too significant.

Figure 8 shows a visualization of each architecture's convergence speed and performance stability throughout training in the Training Loss process, where each line shows the change in loss during training. It can be seen that all models experience a decrease in loss as the number of epochs increases, indicating that the model learns to adjust its weights. ViT-B_16 (the top line at the beginning) finally reaches the lowest loss, followed by ResNet101, ResNet50, ResNet34, and ResNet18. The Validation Accuracy results explain that each line maps the accuracy on the validation data per epoch. All models show an increasing trend in accuracy. ViT reaches the fastest accuracy ( $\sim 0.994$), while ResNet18 increases the slowest. ResNet50 and ResNet101 also approach their peaks earlier than other ResNet variants.

These results indicate that all tested models are well suited for the Malay traditional building classification task, with ViT and ResNet50 being the strongest candidates. In the future, combining CNN and Transformer or exploring ensemble techniques can also be a direction for further research to improve overall accuracy. This study proves that deep learning approaches, especially with ResNet50 and ViT, can be used effectively in video-based traditional architecture classification. The success of ResNet50 in classifying traditional Malay buildings shows the importance of selecting a model appropriate to the data's type and structure. Meanwhile, ViT has great potential for further development, especially when combined with more complex spatial-temporal augmentation techniques. The combination of these two models in the future can produce a classification system that is accurate and adaptive to variations in local cultural elements.

Although ResNet50 recorded the best performance in this dataset, ViT remains a strong candidate for classification that requires an understanding of global visual context. Expanding the dataset, testing the model in real scenarios, and exploring fine-tuning techniques on ViT to improve multi-label classification performance is recommended for further research. Future improvements may include a hybrid approach combining CNN and Transformer architectures to leverage the advantages of local and global feature extraction. Expanding the dataset and integrating real-world conditions may improve the models generalization capabilities.

## 4 Conclusion

Based on the results of the accuracy test of five classification models with a data matrix size of 640 × 640 , ResNet50 showed the best performance with an accuracy of 99.2%, followed by ResNet18 (98.0%), and ResNet34 (98.5%), ResNet101 (99.1%), and ViT Multi-label Classification which each obtained an accuracy of 99.4%. These results indicate that models with deeper architectures, such as ResNet50, can extract more complex features effectively, although lighter models, such as ResNet18, still provide competitive results with better computational efficiency. This study shows that the ResNet and ViT models can automatically classify Malay traditional buildings based on video data. ResNet50 is the most effective model, but ViT remains a promising alternative for tasks that require an understanding of the global visual context. For further research, it is recommended that further testing of these models using more extensive and more varied datasets be conducted to evaluate their resilience to overfitting and test their performance under real conditions. In addition, exploring the ViT model with deeper data augmentation or fine-tuning techniques can be an interesting alternative to improve accuracy on multi-label classification tasks.

## Acknowledgement

Thank you to the Research and Community Service Institute of Universitas Ahmad Dahlan, which has supported the implementation of this research in moral and material support. Especially to Mr. Heri Pramano, S.T., M.Ars., who is an expert in the field of architecture for his scientific support, which provides knowledge transfer related to digital architecture and traditional buildings to facilitate the acquisition of data for this research. Thanks to digital architecture expert Prof. Ir. Prasasto Satwiko, M. Build.Sc., Ph.D., IAI., from the Doctoral Program of Atmajaya University for his cooperation in this research, which provides knowledge transfer for the field of digital architecture.

## References

[1] Y. Pisolkar, "Cultural heritage management and sustainable development: Major themes and research trajectories," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 2417–2431, 2024.

[2] C. Ge, "The review of AI and cultural heritage protection: Taking the whole process of cultural heritage protection as an example," in *Applied and Computational Engineering*, 2024, pp. 137–143.

[3] D. Buragohain, Y. Meng, C. Deng, Q. Li, and S. Chaudhary, "Digitalizing cultural heritage through metaverse applications: challenges, opportunities, and strategies," *Herit. Sci.*, vol. 12, no. 1, 2024.

[4] B. Aithal and P. P. S., *Building Feature Extraction with Machine Learning*. Park Square, Milton Park, Abingdon, Oxon: CRC Press, 2022.

[5] F. Fontanella, F. Colace, M. Molinara, A. S. D. Freca, and F. Stanco, "Pattern recognition and artificial intelligence techniques for cultural heritage," *Pattern Recognit. Lett.*, vol. 138, pp. 23–29, 2020.

[6] M. L. Giudice, F. Mariani, G. Caliano, and A. Salvini, "Deep learning for the detection and classification of adhesion defects in antique plaster layers," *J. Cult. Herit.*, vol. 69, pp. 78–85, 2024.

[7] Z. Zain, N. A. Uray, and S. Christabella, "Identifikasi arsitektur melayu: Rumah tinggal tradisional dan masjid di semenanjung malaysia," *EMARA Indones. J. Archit.*, vol. 7, no. 1, pp. 42–59, 2021.

[8] O. D. Bakare-Fatungase, F. E. Adejuwon, and T. O. Idowu-Davies, "Integrating artificial intelligence in education for sustainable development," *Using Tradit. Des. Methods to Enhance. AI-Driven Decis. Mak.*, vol. 7, no. 2, pp. 231–245, 2024.

[9] 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," in *International Conference on Learning Representations (ICLR)*, 2021.

[10] J. Llamas, P. M. Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo, "Classification of architectural heritage images using deep learning techniques," *Appl. Sci.*, vol. 7, no. 10, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning, PMLR*, 2021, pp. 10 347–10 357.

[13] M. D. Pranatha, M. A. Maricar, and G. H. Setiawan, "Implementasi arsitektural resnet-34 dalam klasifikasi gambar penyakit pada daun kentang," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 6, no. 3, pp. 575–580, 2024.

[14] T. D., E. Vissers-Similon, and J. D. Walsche, "Classification of artificial intelligence techniques for early architectural design stages," *Int. J. Archit. Comput.*, vol. 0, no. 0, pp. 1–18, 2024.

[15] N. Matter and N. Gado, "Artificial intelligence in architecture: Integration into architectural design process," *Al/Engineering Res. J.*, vol. 181, pp. 1–16, 2024. [Online]. Available: https://www.datacamp.com/blog/how-to-learn-ai

[16] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–41, 2022.

[17] K. Ma, B. Wang, Y. Li, and J. Zhang, "Image retrieval for local architectural heritage recommendation based on deep hashing," *Buildings*, vol. 12, no. 6, pp. 1–16,

2022.
[18] J. Liu, A. T. Becerra, J. F. Bienvenido-Barcena, X. Yang, Z. Zhao, and C. Zhou, "CFFI-Vit: Enhanced vision transformer for the accurate classification of fish feeding intensity in aquaculture," *J. Mar. Sci. Eng.*, vol. 12, no. 7, 2024.
[19] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, pp. 1–32, 2025.
[20] E. Ergün, "High precision banana variety identification using vision transformer-based feature extraction and support vector machine," *Sci. Rep.*, vol. 15, no. 1, p. 10366, 2025.
[21] R. K. D. D. Perhavec, "The use of artificial intelligence in building engineering for historic buildings built in the austro-hungarian monarchy," *ACM J. Comput. Cult. Herit.*, vol. 18, no. 1, pp. 10:1–10:23, 2025.
[22] H. Li, X. Yue, Z. Wang, W. Wang, H. Tomiyama, and L. Men, "A survey of convolutional neural networks — from software to hardware and the applications in measurement," *Meas. Sensors*, vol. 18, p. 1000080, 2021.
[23] Y. Bai, "RELU-function and derived function review," in *SHS Web of Conferences*, 2022, p. 02006.
[24] D. Todoli-Ferrandis, J. Silvestre-Blanes, V. Sempere-Payá, and S. Santonja-Climent, "Polling mechanisms for industrial IoT applications in long-range wide-area networks," *Futur. Internet*, vol. 16, no. 4, pp. 1–18, 2024.
[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://deeplearning.net/
[26] V. Shatravin, D. Shashev, and S. Shidlovskiy, "Implementation of the SoftMax activation for reconfigurable neural network hardware accelerators," *Appl. Sci.*, vol. 13, no. 23, p. 12784, 2023.
[27] Y. Zhou, P. Liu, Y. Cui, C. Liu, and W. Duan, "Integration of multi-head self-attention and convolution for person re-identification," *Sensors*, vol. 22, no. 16, p. 6293, 2022.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5999–6009.
[29] J. Sreeranga, K. T. Vishal, N. Sudarshan, and B. Nethravathi, "Leveraging feedforward neural networks for image processing: an overview and analysis," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 6, no. 1, pp. 1961–1967, 2024.
[30] G. Taiwo, S. Vadera, and A. Alameer, "Vision transformers for automated detection of pig interactions in groups," *Smart Agric. Technol.*, vol. 10, p. 100774, 2025.