CENTIVE
Conference on Electrical Engineering, Informatics
, Industrial Technology, and Creative Media
ISBN : 978-602-53004-0-0

Comparative Study of CNN, Vision Transformer, and Hybrid CNN–ViT Models for Indonesian Batik Pattern Classification

Naufal El Kamil Aditya Pratama Rahman [1*], Akmelia Zahara[2], Bintang Yudhistira[3]
[123]Telkom University Purwokerto
[123] JL. DI Panjaitan No.128, Banyumas, Central Java
*naufalelkamiladityap@student.telkomuniversity.ac.id
akmeliazahara@student.telkomuniversity.ac.id
bintangyudhistira@student.telkomuniversity.ac.id

Abstract — Batik is an Indonesian cultural heritage with unique visual characteristics and deep philosophical value. The complexity of motifs, color variations, and geometric details make batik classification an interesting challenge in the field of computer vision. This study conducted a comparative study between three deep learning approaches for classifying Indonesian batik motifs using Convolutional Neural Network (CNN), Vision Transformer (ViT), and a hybrid CNN–ViT model. The dataset used includes more than 3,000 batik images from various regions in Indonesia, with a variety of motifs such as Yogyakarta Kawung, Aceh, Ceplok, and Megamendung.Each model was trained with uniform parameters and augmentations to ensure fair evaluation, resulting in CNN accuracy of 94.43% F1-macro 93.45%, ViT accuracy of 91.55% F1-macro 89.78%, and Hybrid CNN-ViT accuracy of 94.04% F1-macro 92.91%. This is reinforced by the combination of modules (EfficientNet-B2 + CBAM + ArcFace) that can improve model performance furthermore, This study contributes to the development of an automated batik classification system and supports cultural preservation through artificial intelligence- based digitization.

Keywords – Batik Indonesian, CNN, Vision Transformer, Hybrid Model, Image Classification

I. INTRODUCTION

Batik was once a cultural debate between Indonesia and Malaysia in the early 2000s. This occurred because Malaysia often claimed batik as part of its cultural heritage, especially after they popularized Malaysian Batik internationally. Although both countries do have traditions of patterned textiles, Indonesian batik has a much deeper and more complex character, technique, and philosophy, particularly in the process of canting writing and natural dyeing, which is rich in symbolic meaning. This prompted the Indonesian government to fight for official recognition, resulting in batik being recognized by UNESCO as Intangible Cultural Heritage of Humanity in 2009, where UNESCO em- phasized its importance as part of Indonesia's cultural heritage. Batik is not just a patterned fabric, but an ex- pression of identity, philosophy, and local wisdom that has been passed down from generation to generation. Each region in Indonesia has its own unique visual style and symbolic meaning, such as Parang, which symbolizes strength and courage; Kawung, which re- flects purity and balance in life; and Megamendung, which depicts tranquility and wisdom. The diversity of batik creates challenges in automatic recognition and classification. Batik motifs have highly complex shapes, precise yet varied repeating patterns between regions, and combinations of colors and textures. In addition, batik is used in various products such as clothing, sarongs, and many other items. This requires computer systems to have models capable of under- standing visual representations at a deep level, so that they can recognize not only basic shapes but also more abstract patterns. In recent years, advances in deep learning technology, particularly Convolutional Neural Networks (CNN), have shown significant results in the field of image recognition. CNN is capable of extracting important features hierarchically from edges and textures to complex patterns making it effective for identifying and classifying batik motifs with varying degrees of complexity. Through this approach, the

system will learn directly from visual data without the need for manual feature engineering, thereby improving the accuracy, efficiency, and generalization of the model [1], [2].

On the other hand, the Vision Transformer (ViT) method has emerged as a modern architecture that changes the way models understand visual representations. Unlike CNNs, which focus on spatial locality through convolution, ViT uses a self-attention mechanism to capture long-range dependencies between parts of an image, enabling it to learn overall patterns. This approach is particularly useful for solving problems with images that have fine details and complex motif variations, such as batik. Various studies show that ViT can outperform CNN in fine-grained classification tasks when supported by a large amount of data and the right augmentation strategy. ViT offers a new per- spective in visual analysis that was previously difficult to achieve with convolutional approaches [3].

This study aims to compare the performance of three modern deep learning architectures, namely CNN with CBAM and ArcFace modules, Vision Transformer (ViT) with ArcFace, and a hybrid CNN–ViT model that also uses ArcFace in the task of classifying In- donesian batik motifs. This study also aims to analyze the relationship between accuracy and computational efficiency in order to identify the most suitable model for real-world applications such as cultural heritage digitization systems, virtual museums based on batik images, and batik product recommendations on e-commerce platforms. With this approach, the study is expected to contribute to the development of an accurate, efficient, and applicable batik motif recognition system in the era of digital transformation, especially in Indonesia [1], [3], [4].

II. RESEARCH METHOD

This research was conducted in systematic stages, beginning with the collection and processing of batik datasets, followed by image preprocessing before being entered into the model for further analysis. [10]. The main process includes training, validation, and testing on three main architectures: EfficientNet-B2 with CBAM and ArcFace modules, Vision Transformer (ViT-B/16) with ArcFace, and a hybrid CNN–ViT model that combines both through the Gated Fusion module. Each model was evaluated using accuracy and macro F1-score metrics on stratified split test data. The overall flow of this research is illustrated in Figure 1, which shows the stages from data preparation and training pipeline to model evaluation.

A. Related Research

Previous research on the application of CNNs in batik motif classification shows that this model is capable of achieving high accuracy on small to medium-sized datasets, especially when utilizing transfer learning techniques from pre-trained models such as VGG,

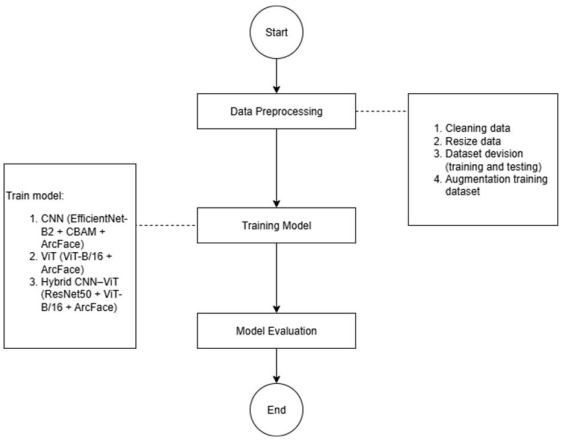


Fig. 1. Workflow of the research

ResNet, MobileNet, or EfficientNet. In addition, the use of fine-tuning on the final layer often improves the model's ability to recognize the fine details and complex textures of each batik motif. These results demonstrate the great potential of CNN as the basis for architecture in batik classification systems [1]–[3], [5].

Meanwhile, Transformer/ViT is effective on images due to global self-attention, and a comparative study on the traditional Indonesian food classification do- main (cakes) reported that ViT was slightly superior to EfficientNet when the differences between classes were subtle [3]. Hybrid CNN–Transformer models are reported to be able to balance accuracy and latency, for example, linear-attention/hybrid designs for facial recognition on mobile devices [6].

B. Batik

Batik is part of Indonesia's cultural heritage, with a wide variety of motifs and philosophies. Each pattern, such as *Solo Parang*, *Pekalongan*, *Aceh*, and *Megamendung*, has its own symbolic meaning, representing the moral and social values of society [3], [5]. The dataset used consists of several classes of popular motifs, taken from online and local collections. All images were resized to 224×224 pixels and normalized with parameters *ImageNet* [2], [5].

C. Dataset

The dataset used in this study was obtained from open sources. All images were converted to .jpg format with a uniform resolution of 224×224 pixels, following the ImageNet model input standard [10]. The main reference includes the Batik Motif Dataset from the Hugging Face Datasets platform, named "Indonesian Batik Motif Classification Dataset", which contains a collection of batik images from various regions in Indonesia. This dataset was curated from several public sources and previous academic projects and has 38 classes, covering various types of motifs such as *Parang*, *Kawung*, *Ceplok*, *Truntum*, dan *Megamendung*.

Overall, the dataset consists of approximately 3,000–4,200 images with resolutions varying between 256×256 and 1024×1024 pixels. All images were then resized to 224×224 pixels to fit ImageNet-based architectures such as EfficientNet and ViT. The complete dataset reference is available at: https://huggingface.co/datasets/muhammadsalmanalfaridzi/Batik-Indonesia.

Here are examples of several datasets based on their classes in the figure 2

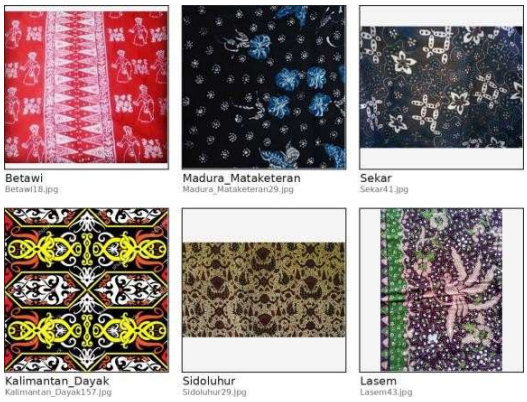


Fig. 2. Examples of Batik Dataset

D. ArcFace

ArcFace is an Additive Angular Margin Loss used to optimize model performance with the constraint of angles and arcs from both ViT and CNN. In this case, ArcFace can improve resistance to image noise. The use of an ArcFace margin of 0.50 with a scale of 30 adds a penalty and increases the similarity value before the loss function is executed [12].

E. CNN with CBAM and ArcFace

The CNN architecture uses EfficientNet-B2 for the main backbone due to its balance between parameter efficiency and accuracy performance [5]. CBAM (Convolutional Block Attention Module), this model is equipped with a CBAM (Convolutional Block Attention Module), an efficient attention module that serves to adaptively refine the feature map extracted by CNN. This module works by sequentially determining "what" is important (through Channel Attention) and "where" the important information is located (through Spatial Attention), so that the network can focus on informative features while suppressing less relevant areas [7]. This stage consists of two main stages: (1) Channel Attention, which studies the importance weight of each feature channel through average pooling and max pooling operations followed by a multi-layer perceptron (MLP); and (2) Spatial Attention, which pays attention to important areas in the image through 7×7 convolution on the channel aggregate map [3], [8].

After the attention stage, global features are averaged using global average pooling and then projected onto a 512-dimensional embedding space. To improve inter-class separability in the features, an ArcFace Loss classification layer with angle margin is used $m = 0.5$ and scale $s = 30$ [3]. The integration of CBAM and ArcFace improves the model's ability to distinguish similar motifs, especially those with complex color and texture variations such as *Nitik* and *Sekar Jagad*. [3], [8]. The following is an overview of the CNN model architecture:

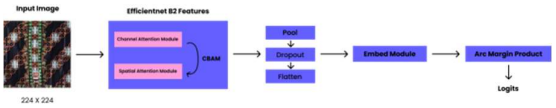


Fig. 3. CNN with CBAM and ArcFace Model

F. Vision Transformer (ViT) with ArcFace

The ViT-B/16 model projects images into small 16×16 pixel patches using an initial convolutional layer. Each patch is represented as an embedding vector that is added with positional embedding and the [CLS] classification token [6]. Next, all tokens are processed by several Transformer encoder blocks with a multi-head self-attention mechanism to capture global spatial relationships between image parts and strengthen their features [6], [7]. The [CLS] vector from the encoder is then normalized through Layer- Norm and dropped using dropout, before being pro- jected onto a 512-dimensional embedding and classi- fied with ArcFace Loss similar to CNN. The combi- nation of ViT and ArcFace provides better inter-class margins and strengthens the model's generalization of batik motifs that have repeating global patterns, such as *Yogyakarta Kawung* and *Garutan* [6], [7]. The following is an illustration of the model:

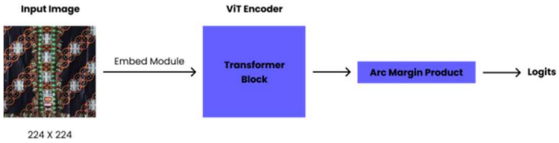


Fig. 4. Vision Transformer (ViT) with ArcFace

G. CNN–ViT Hybrid with ArcFace

The Hybrid CNN–ViT model is designed to combine the local features of CNN with the global context of ViT in a single classification framework [1], [9]. This architecture uses ResNet50 as a feature extractor CNN (f_{cnn}) and ViT-B/16 as a Transformer extractor (f_{vit}). The two vectors are combined through the Gated Fusion module, which uses a double linear layer and sigmoid function to generate gate weights g_{cnn} and g_{vit}. The fusion vector is calculated using the equation:

$$f_{\text{fused}} = [\, g_{cnn} \odot f_{cnn},\ g_{vit} \odot f_{vit},\ f_{cnn} \,\|\, f_{vit} \,], \quad (1)$$

It is then projected into a 1024-dimensional space, dropped out, and mapped to a 512-dimensional embedding space. The final layer uses ArcFace to generate a

fixed angular margin, improving inter-class separation. This hybrid model balances the strengths of CNN in capturing micro textures and ViT in understanding macro patterns of batik fabric, while maintaining high inference efficiency. [1], [9]. The following is an illustration of the model:

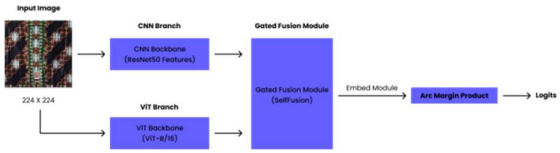
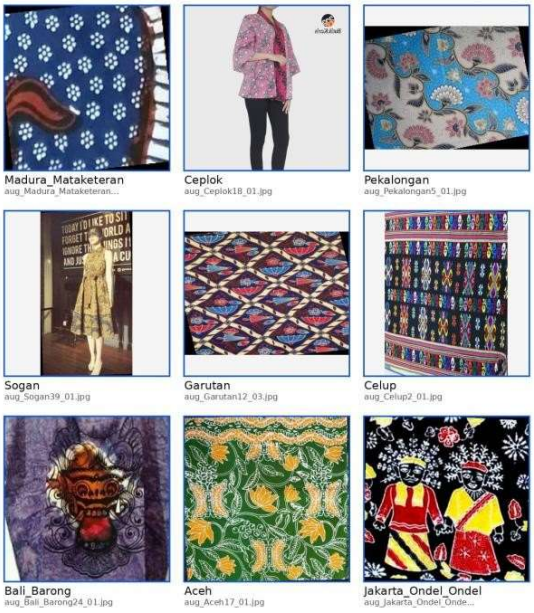


Fig. 5. CNN–ViT Hybrid with ArcFace

H. Preprocessing Data

At this stage, the pixel values of the data are normalized based on ImageNet standards and the image size is changed to match the model created. An augmentation process is also carried out for the training data with a distribution of 2079 to 3608 for the training data and 512 for the test data. The data is divided into *stratified* with a ratio of 80% training data and 20% testing data. This division process is carried out randomly while maintaining a balanced class distribution. The augmentation process includes Flip, Rotate, Brightness, Contrast, Color, Perspective transform, and Gaussian noise [11]. The following is an example of augmented data



I. Model Evaluation

The evaluation was conducted on a test set to assess the model's ability to consistently distinguish various batik motifs. Class predictions were obtained by calculating the cosine score between the trained embedding vector and the normalized class weight vector, then selecting the class with the highest score. In addition, performance was summarized through a confusion matrix and macro accuracy and F1-score metrics to ensure fair assessment of unbalanced class distributions.

Cosine-based prediction. Given an embedding $e \in \mathbb{R}^d$ and a class weight $w_k \in \mathbb{R}^d$, both are normalized such that $\tilde{e} = e/\|e\|$ and $\tilde{w}_k = w_k/\|w_k\|$. The score for class-$k$ is defined as

$$\text{logit}_k = \langle \tilde{e}, \tilde{w}_k \rangle, \qquad \hat{y} = \underset{k}{\text{argmax}} \text{logit}_k. \tag{2}$$

Confusion matrix and derived metrics. The confusion matrix calculates four basic components per class: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From these components, metrics are evaluated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

F1-score makro. For potentially imbalanced multi-class data, global performance is summarized using the unweighted inter-class F1 average:

$$\text{F1}_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i, \tag{7}$$

with N is the number of classes. In addition to these aggregate values, the general visual evaluation includes a row-normalized confusion matrix and an F1 bar chart per class to review error patterns and relatively difficult classes.

III. RESULT

A. Experimental Setup

All experiments were conducted using a batik dataset obtained from Hugging Face. The dataset con- sists of a number of batik images from various regions in Indonesia and is divided into 80% training data and 20% test data. The images were converted to RGB format with a size of 224×224 pixels. The training process was carried out for 100 *epoch* use *batch size* 32, *learning rate* 1×10^{-4}, and optimizer AdamW with *scheduler* Cosine Annealing. Model training was performed on a GPU to accelerate and simplify the convergence process. The loss function used was CrossEntropyLoss with ArcFace's angular margin to strengthen separability between classes.

B. Scenario 1 (CNN with CBAM and ArcFace)

The CNN model used is *EfficientNet-B2* with the addition of a CBAM (Convolutional Block Attention Module) on each main convolution block. CBAM consists of two stages: (1) *Channel Attention* to determine the most relevant feature channels, and (2) *Spatial Attention* to enhance important areas in the image. The final layer uses ArcFace to generate a fixed corner margin, which reinforces separability between classes. The proposed CNN architecture utilizes the pretrained EfficientNet-B2 backbone as an initial feature extractor. The resulting 1408-channel feature map is then processed by the CBAM module, which consists of Channel Attention and Spatial Attention. This module improves the quality of representation by emphasizing important features and reducing the influence of noise or irrelevant backgrounds. Without CBAM, the pooling process would treat all channels and spatial locations equally, putting the model at risk of losing the distinguishing details in batik motifs. The feature map strengthened by CBAM is then aggregated with Global Average Pooling and projected onto a 512-dimensional embedding through a fully connected layer and Batch Normalization. This embedding is further processed by ArcFace, which applies angular margin with parameters $m = 0.5$ and $s = 30$ to increase the distance between classes and tighten intra-class representations. The combination of feature filtering by CBAM and feature separation by ArcFace improves the model's ability to distinguish batik motifs that have high visual similarity.

Figure 7 shows the confusion matrix results of the model. This CNN model tends to converge stably and produces high performance on strongly textured patterns and has significant differences.

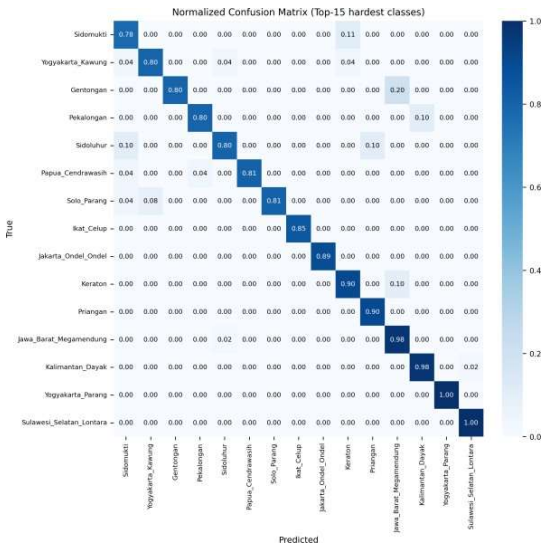


Fig. 7. Confusion Matrix Top 15 CNN with CBAM and ArcFace.

In the confusion matrix normalized for the 15 most difficult batik classes, the EfficientNet-B2 model with CBAM and ArcFace (ArcMarginProd- uct) showed fairly good performance with diag- onal accuracy values ranging from 0.78 to 1.00. Classes such as *Sidomukti*, *Yogyakarta Kawung*, *Gen- tongan*, *Pekalongan*, and *Sido Luhur* still shows a fairly high level of confusion with an accu- racy of around 0.78–0.81, while classes such as *Jawa Barat Megamendung*, *Kalimantan Dayak*, *Yo- gyakarta Parang*, and *Sulawesi Selatan Lontara* is al- ready very stable with an accuracy close to 1.00.

Classification errors mostly occur between motifs that have similar patterns and visual structures, such as between variants *Parang*, *Kawung*, and *Sidomukti*, which have similar geometric shapes and patterns. This shows that the model still has difficulty distinguishing fine details in motifs with complex repeating patterns or varying color contrasts.

Overall, the combination of CBAM helps the model focus on important feature areas and channels, while ArcFace strengthens the separation between classes in the embedding space, allowing classes with strong visual characteristics to be separated well. However, the main challenge for this model is distinguishing between classes that are structurally very similar and handling variations in color and size of motif elements.

In general, this model has shown strong results, and with minor adjustments to the augmentation strategy and loss function, its performance has the potential to improve, especially for batik classes that have a high degree of similarity in form.

C. Scenario 2 (Vision Transformer (ViT) with ArcFace)

The ViT-B/16 model is a model-based approach *self-attention* to capture global relationships between patches in an image. Each image is divided into 16×16 pixel patches, converted into linear embeddings, and assigned *positional encoding* so that the spatial order is maintained. The classification layer also uses ArcFace for the same angular margin as CNN. Each token is then processed through a transformer block consisting of multi-head self-attention and a feed-forward net- work. This mechanism allows the model to learn batik motif patterns comprehensively, not only from local patches but also from inter-patch relationships. The representation of class tokens is normalized and projected into the embedding space before being fed into ArcFace, so that the distance between classes becomes more angularly separated. This strategy improves the model's ability to distinguish between motifs that have high visual similarity.

Figure 8 shows the confusion matrix. The ViT model demonstrates better ability in recognizing global patterns, but requires longer training time.

From the normalized confusion matrix for the 15 most difficult batik classes, the Vision Transformer (ViT) model with ArcFace showed good performance

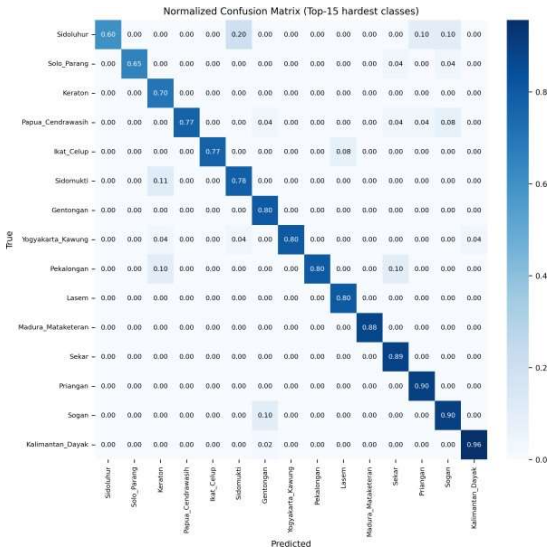


Fig. 8. Confusion Matrix Model Vision Transformer with ArcFace

in recognizing most batik patterns, with diagonal accuracy ranging from 0.60 to 0.96. Several classes such as *Sidoluhur*, *Solo_Parang*, and *Keraton* still experiences a relatively high prediction error rate, with an accuracy below 0.75. This indicates that the model still has difficulty distinguishing between motifs that have similar geometric structures and repeating patterns. Meanwhile, classes such as *Sogan*, *Priangan*, and *Kalimantan_Dayak* has demonstrated excellent performance with an accuracy above 0.90.

In general, the use of ArcFace successfully helped separate classes in the embedding space so that motifs with strong visual characteristics could be clearly separated. However, because the ViT architecture relies on global representations of image patches, its performance declined when local shape information of motifs was not captured well due to variations in texture, color, or image cropping. Overall, the ViT model with ArcFace is already capable of recognizing most motifs well, but improvements are still needed for classes with similar visual patterns or complex fine details.

D. Scenario 3 (Hybrid CNN–ViT with ArcFace)

Hybrid CNN–ViT model combines *ResNet50* and ViT-B/16 to balance the local strength of CNN and the global context of ViT. Features from both backbones are combined using a mechanism *Gated Fusion* which generates gate weight g_{cnn} and g_{vit} to balance the contributions of each representation:

$$f_{\text{fused}} = [\, g_{cnn} \odot f_{cnn},\ g_{vit} \odot f_{vit},\ f_{cnn} \,\|\, f_{vit} \,]. \quad (1)$$

The fusion results were then projected into 512 dimensions and classified using ArcFace. Figure 9 shows the hybrid model Confusion Matrix, which demonstrates the highest stability and performance compared to the other two models.

The Hybrid Model Architecture is designed to leverage the complementary strengths of CNN local feature extraction and ViT global context understanding. The data processing runs in parallel: the input image is passed through the ResNet50 backbone without the final fully connected layer to generate a local feature vector (f_{cnn}), while the same image is processed by the ViT-B/16 backbone, where the [CLS] token output after the Transformer Encoder is used as the global feature vector (f_{vit}). These two representations then enter the Gated Fusion module, which is a gate network consisting of Linear, ReLU, and Sigmoid layers to generate attention weights g_{cnn} and g_{vit}. These weights allow the model to adaptively balance the contributions between local and global features, so that the model can focus on texture when necessary or on spatial context when more relevant.

In accordance with the implementation and Equation (1), the weighted vectors are recombined through a skip-connection mechanism before being projected into a 1024-dimensional fusion space. This final vector is then projected onto a 512-dimensional embedding through a fully connected layer and Batch Normalization, before being passed to ArcFace with parameters $m = 0.5$ and $s = 30$ for the classification stage. ArcFace enhances separability between classes, so that the combination of dual feature extraction and the Gated Fusion mechanism produces more stable and accurate performance in distinguishing batik motifs that have high visual similarity.

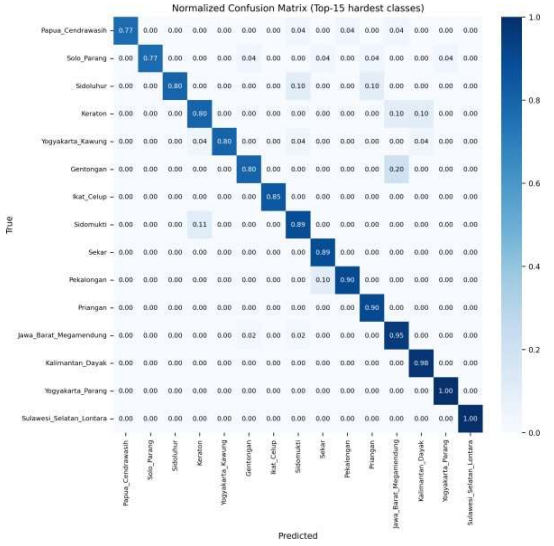


Fig. 9. Confusion Matrix Model Hybrid CNN–ViT with ArcFace

Based on the confusion matrix for the 15 most difficult batik classes, the Hybrid CNN-ViT model with ArcFace showed quite good results with diagonal accuracy ranging from 0.77 to 1.00. Several classes such as *Papua_Cendrawasih* and *Solo_Parang* still has a fairly high error rate, whereas classes such as *Yogyakarta_Parang* and *Sulawesi_Selatan_Lontara* has achieved perfect accuracy.

This hybrid model is capable of utilizing the ad-

vantages of two architectures, namely CNN, which is strong in capturing local texture patterns, and ViT, which is good at understanding the global shape of motifs. The combination of the two makes the model more capable of recognizing batik motifs with complex and diverse patterns. However, errors still often occur in motifs that have similar shapes and structures, such as between *Parang*, *Kawung*, and *Sidomukti*.

Overall, this hybrid model has demonstrated good performance in batik motif classification and has the potential to be improved with augmentation and training strategies.

E. Performance Comparison

The quantitative evaluation results are shown in Table 1. Accuracy and F1-macro values are measured based on prediction results in the test dataset to assess performance consistency between models.

Table 1. Comparison of performance results between models

Model	Accuracy	F1-macro
CNN (EfficientNet-B2 + CBAM + ArcFace)	94.43%	93.45%
ViT (ViT-B/16 + ArcFace)	91.55%	89.78%
Hybrid CNN–ViT (ResNet50 + ViT-B/16 + ArcFace)	94.04%	92.91%

In general, all three models performed well in batik motif classification tasks. The CNN model produced the highest accuracy and F1-macro scores, demonstrat- ing its ability to recognize local texture details in batik images. The ViT model produced slightly lower results, as it focused more on global relationships between image areas. Meanwhile, the Hybrid CNN-ViT model has balanced and stable performance, as it is able to combine the advantages of CNN in recognizing local patterns and the ability of ViT in understanding global context. Thus, all models are able to classify batik motifs quite well, with no significant differences in performance.

IV. DISCUSSION

The results of the experiment show that each architecture has its own characteristics and advantages in recognizing complex batik motifs. The CNN model (EfficientNet-B2 + CBAM + ArcFace) achieved the highest performance with an accuracy of 94.43% and an F1-macro of 93.45%. This improvement was mainly due to the combination of the CBAM module, which strengthened the model's attention to important features in batik textures, and the use of ArcFace, which expanded the distance between classes in the embedding space, resulting in more discriminative feature representations. In addition, the dataset used was more varied and larger than in previous studies. In the study [9], apart from the difference in the number of datasets, the accuracy of the CNN model produced showed quite good results with a value of 89%. However, the CNN model with our module combination experienced a significant improvement. This certainly reinforces that the combination of modules (EfficientNet-B2 + CBAM + ArcFace) in the CNN model is capable of improving its performance.

Furthermore, the ViT model (ViT-B/16 + ArcFace) showed slightly lower performance than our experimental results, with an accuracy of 91.55% and an F1-macro score of 89.78%. This can be explained by the basic nature of Vision Transformer, which is more effective in understanding global relationships between patches, but less optimal in capturing local details when training data is limited or the texture of the motif is very dense. Nevertheless, ViT still shows good ability in recognizing patterns with large structures and clear contrasts. This is also clarified in the study [3] that ViT works well in generalizing traditional cake classification images. The experimental results show that the best ViT accuracy is obtained with a value of 96.25% with self-attention in ViT.

Then, the Hybrid CNN–ViT approach (ResNet50 + ViT-B/16 + ArcFace) provided balanced results between the two, with an accuracy of 94.04% and an F1-macro of 92.91%. This architecture successfully leverages the advantages of CNN in extracting local details and the ability of ViT in understanding global context, resulting in a model that is more robust to variations in shape and color in batik motifs. In general, these results show that the combination of these two approaches can be an effective solution for classification problems with complex characteristics and similar visuals between classes.

V. CONCLUSION

Based on the results of the research conducted, it can be concluded that the CNN model (EfficientNet-B2 + CBAM + ArcFace) provides the best performance in batik motif classification tasks with an accuracy of 94.43% and an F1-macro of 93.45%. This model demonstrates high adaptive capabilities to variations in patterns and complex textures in batik images.

The ViT model (ViT-B/16 + ArcFace) also provides good results but still has limitations in recognizing subtle local patterns. Meanwhile, the Hybrid CNN–ViT model (ResNet50 + ViT-B/16 + ArcFace) successfully combines the advantages of both and produces competitive performance, showing great potential for use in broader batik image classification applications.

For further research, it is recommended to make improvements through more diverse data augmentation and the application of a gradual fine-tuning strategy. With these developments, it is hoped that the deep learning-based batik motif classification system can achieve higher accuracy and be able to generalize a wider variety of motifs.

REFERENCES

[1] E. Sugiarto, F. Budiman, and A. Fahmi, "Implementation of Deep Learning Based on Convolution Neural Network for

Batik Pattern Recognition," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 10, no. 1, pp. 11–16, Feb. 2025, doi: 10.22219/kinetik.v10i1.2019.

[2] I. P. Sari and L. Elvitaria, "Data-driven Approach for Batik Pattern Classification Using Convolutional Neural Network (CNN)," *Jurnal Mandiri*, vol. 13, no. 3, pp. 323–331, 2025.

[3] D. Trisnawarman, A. A. Supriyanton, V. C. Mawardi, and U. A. Okengwu, "Comparative Study of CNN and Vision Transformers on Indonesian Traditional Cakes Classification," *International Journal of Advances in Artificial Intelligence and Machine Learning*, vol. 2, no. 2, pp. 86–94, July 2025, doi: 10.58723/ijaaiml.v2i2.405.

[4] N. Puspitasari, A. Tejawati, and A. P. A. Masa, "East Kalimantan Batik Image Classification Using CNN Architecture with Advanced Feature Extraction," *International Journal on Electrical Engineering and Informatics*, vol. 17, no. 2, pp. 223–237, June 2025, doi: 10.15676/ijeei.2025.17.2.6.

[5] N. Setyawan, C.-C. Sun, M.-H. Hsu, W.-K. Kuo, and J.-W. Hsieh, "FaceLiVT: Face Recognition Using Linear Vision Transformer with Structural Reparameterization for Mobile Device," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2025, doi: 10.1109/ICIP55913.2025.11084611.

[6] P.-S. Xie, J.-L. Wang, H. Wang, Y.-C. Pan, X.-Y. Li, and T. Feng, "Industrial Internet Vulnerability Detection Method Based on CBAM-CNN-SVM," *International Journal of Network Security*, vol. 25, no. 3, pp. 385–393, May 2023, doi: 10.6633/IJNS.202305.

[7] M. Dahbali, N. Aboutabit, and N. Lamghari, "A Hybrid Model for Arabic Script Recognition Based on CNN-CBAM and BLSTM," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 10, no. 3, pp. 294–302, Sept. 2024.

[8] M. F. Dzulqarnain, A. Fadlil, and I. Riadi, "Performance Comparison of Learned Features from Autoencoder and Shape-Based Hu Moments for Batik Classification," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 4, pp. 1729–1744, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4827.

[9] M. B. Kurniawan and E. Utami, "Comparative Analysis of Contrast Enhancement Methods for Classification of Pekalongan Batik Motifs Using Convolutional Neural Network," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 6, pp. 1779–1787, Dec. 2024, doi: 10.52436/1.jutif.2024.5.6.2621.

[10] D. A. Ramadhan and D. Ramadhani, "Classification of Riau Batik Motifs Using the Convolutional Neural Network (CNN) Algorithm," *International Journal of Electrical, Energy and Power System Engineering*, vol. 7, no. 3, pp. 201–211, Nov. 2024, doi: 10.31258/ijeepse.7.3.201-211.

[11] A. M. Akbar, M. Perdana, M. Fajar, and A. M. Mappalotteng, "Enhancing Batik Classification Leveraging CNN Models and Transfer Learning," *International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 354–361, 2023. [Online]. Available: https://www.joiv.org/index.php/joiv

[12] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4037–4050, 2022, doi: 10.1109/TPAMI.2021.3087709.