JSIKTI (Jurnal Sistem Informasi dan Komputer Terapan Indonesia)
Vol.8, No.4, June 2026, pp. 91~104
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.33173/jsikti.314

Classifying Indonesian Batik Motifs by Region Using Swin Small Transformer Architecture

Ida Bagus Ketut Sukanegara*[1], Aniek Suryanti Kusuma[2], Putu Sugiartawan[3]
[1,2,3] Magister Program of Informatics, Institut Business and Technology, Indonesia
e-mail: ***[1] sukanegara@student.instiki.ac.id**, [2]anieksuryanti@instiki.ac.id,
[3]putu.sugiartawan@instiki.ac.id

Abstrak

Batik merupakan salah satu warisan budaya Indonesia yang memiliki peran penting dalam merepresentasikan identitas daerah, nilai filosofis, serta konteks sosial dan sejarah yang terkandung pada setiap motif. Selain memiliki nilai budaya yang tinggi, batik juga berkontribusi terhadap perkembangan industri kreatif dan sektor pariwisata. Klasifikasi otomatis batik berdasarkan daerah asal dapat mendukung proses dokumentasi, edukasi, dan autentikasi, namun masih menjadi tantangan karena adanya kemiripan visual antar motif, variasi yang tinggi dalam satu kelas, serta perbedaan antar daerah yang bersifat halus. Seiring dengan perkembangan Vision Transformer, penelitian ini mengkaji penerapan arsitektur Swin Small Transformer untuk mengklasifikasikan motif batik Indonesia ke dalam lima kategori wilayah, yaitu Jawa Barat, Jawa Tengah, Jawa Timur, Madura, dan Yogyakarta. Model yang diusulkan menggunakan arsitektur swin_small_patch4_window7_224 yang diinisialisasi dengan bobot pralatih ImageNet dan selanjutnya dilakukan proses fine-tuning menggunakan dataset batik yang telah dikurasi. Mekanisme hierarchical shifted-window attention pada Swin Transformer dimanfaatkan untuk menangkap karakteristik visual lokal berupa pola berulang sekaligus struktur komposisi global yang menjadi ciri khas masing-masing daerah. Hasil pengujian pada data uji yang terdiri atas 80 citra menunjukkan performa yang sangat baik. Model memperoleh akurasi, macro-precision, macro-recall, dan macro F1-score masing-masing sebesar 100%, tanpa ditemukan kesalahan klasifikasi pada seluruh kategori batik. Hasil tersebut menunjukkan bahwa arsitektur yang diusulkan mampu mempelajari representasi fitur yang diskriminatif untuk membedakan motif batik berdasarkan wilayah asalnya. Temuan penelitian ini menunjukkan bahwa hierarchical Vision Transformer merupakan pendekatan yang efektif dalam memodelkan karakteristik visual yang kompleks pada motif batik serta berpotensi menjadi alternatif yang unggul dibandingkan pendekatan berbasis Convolutional Neural Network (CNN). Selain untuk klasifikasi batik, kerangka kerja yang diusulkan juga berpotensi diterapkan pada berbagai aplikasi tekstil warisan budaya lainnya guna mendukung pelestarian digital, kegiatan edukasi, serta dokumentasi aset budaya secara lebih luas.

Kata kunci*— batik Indonesia; Swin Transformer; Vision Transformer; citra warisan budaya; pengenalan pola tekstil.*

Abstract

Batik plays a crucial role in Indonesian cultural heritage, with regional motifs encoding local philosophies, identities, and socio-historical contexts while also sustaining creative industries and tourism. Automated classification of batik by region can support documentation, education, and authentication, yet remains challenging due to visually overlapping patterns, high intra-class variability, and subtle inter-regional differences. Building on recent advances in Vision Transformers, this study investigates the Swin Small Transformer architecture for classifying Indonesian batik motifs into five regional categories: Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta. The proposed framework employs the

Received May 1st,2026; Revised May 25th, 2026; Accepted June 10th, 2026

swin_small_patch4_window7_224} model initialized with ImageNet-pretrained weights and fine-tuned on a curated regional batik dataset. The hierarchical shifted-window attention mechanism of Swin is leveraged to capture both local repetitive elements and broader compositional structures that characterize regional styles. Experimental evaluation on a held-out test set consisting of 80 images demonstrates outstanding performance. The model achieves perfect classification results with overall accuracy, macro-averaged precision, recall, and F1-score all reaching 1.0000. No misclassifications are observed across any regional category, indicating that the proposed architecture effectively learns discriminative representations of regional batik motifs. These findings suggest that hierarchical Vision Transformers can robustly model the nuanced visual cues underpinning regional identity in batik patterns and provide a strong alternative to conventional convolutional neural network approaches. Beyond batik classification, the proposed framework may be extended to other cultural-heritage textile applications, supporting digital preservation, educational initiatives, and large-scale documentation of traditional artistic assets.

Keywords— *Indonesian batik; Swin Transformer; vision transformers; cultural heritage imaging; textile pattern recognition*

1. INTRODUCTION

Indonesian batik represents one of the most significant forms of textile-based cultural heritage, reflecting centuries of artistic development, cultural identity, and regional knowledge transmission. Batik motifs embody symbolic meanings associated with local philosophies, historical narratives, social structures, and environmental influences. Different regions across Indonesia have developed distinctive batik traditions characterized by unique motif compositions, color palettes, ornamentation styles, and production techniques. Consequently, batik serves not only as a decorative textile but also as a cultural artifact that preserves collective memory and regional identity [1].

The growing importance of digital cultural heritage preservation has motivated the development of intelligent systems capable of documenting, organizing, and analyzing traditional artifacts. Advances in artificial intelligence, computer vision, and machine learning have enabled the automated classification of cultural objects, supporting museums, archives, educational institutions, and cultural organizations in managing large collections efficiently [2] [3]. Image-based recognition technologies are increasingly used to identify artifacts, analyze visual styles, and facilitate digital access to cultural resources.

Despite these advances, automated batik recognition remains a challenging computer vision problem. Batik motifs often exhibit substantial intra-class variation due to differences in craftsmanship, coloring processes, textile materials, and artistic interpretation. At the same time, visually similar motifs may appear across different regions, creating subtle inter-class distinctions that are difficult to recognize automatically. These characteristics make regional batik classification a fine-grained visual recognition task requiring highly discriminative feature representations [4].

Traditional image classification methods relied primarily on handcrafted feature extraction techniques such as texture descriptors, color histograms, and shape-based representations. Methods based on Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG) have been applied to textile recognition with varying degrees of success [5]. However, handcrafted features often struggle to capture the complex hierarchical structures and semantic relationships embedded within batik motifs. Furthermore, these approaches typically require extensive feature engineering and may exhibit limited generalization across diverse datasets.

The emergence of deep learning has significantly improved image classification performance by enabling automatic representation learning from raw image data. Convolutional

Neural Networks (CNNs) such as AlexNet, VGGNet, ResNet, DenseNet, and EfficientNet have achieved remarkable success across numerous visual recognition tasks and have been applied to various cultural heritage applications [6] [7]. CNN-based approaches learn hierarchical feature representations that progressively encode textures, shapes, and semantic concepts, providing substantial advantages over traditional handcrafted methods.

More recently, Vision Transformers (ViTs) have emerged as a powerful alternative to convolutional architectures. Inspired by transformer models originally developed for natural language processing, Vision Transformers treat images as sequences of visual patches and employ self-attention mechanisms to model relationships among image regions [8]. Unlike CNNs, which primarily focus on local receptive fields, transformer architectures can capture long-range dependencies and global contextual information more effectively. These characteristics make Vision Transformers particularly attractive for fine-grained recognition tasks where subtle visual cues may be distributed across multiple image regions.

Among recent transformer-based models, the Swin Transformer has attracted considerable attention due to its hierarchical architecture and computational efficiency. Swin Transformer introduces shifted-window self-attention mechanisms that restrict attention computation to local windows while enabling information exchange across neighboring regions through window shifting operations [9]. This design significantly reduces computational complexity while preserving the ability to model both local and global visual relationships. As a result, Swin Transformer has achieved state-of-the-art performance across a wide range of computer vision tasks, including image classification, object detection, semantic segmentation, and medical image analysis [10].

The Swin Small Transformer variant provides an attractive balance between model capacity and computational efficiency. By employing hierarchical feature extraction and multi-stage representation learning, Swin Small can effectively capture fine-grained visual details while maintaining relatively moderate computational requirements. These characteristics are particularly relevant for cultural heritage applications, where datasets are often smaller than large-scale industrial benchmarks and computational resources may be constrained.

Recent studies have demonstrated the effectiveness of transformer-based approaches for recognizing traditional foods, textiles, heritage artifacts, and other culturally significant objects [11] [12]. Nevertheless, research specifically focused on Indonesian batik classification using hierarchical Vision Transformers remains relatively limited. Most existing studies continue to rely on CNN-based architectures or conventional machine learning techniques, leaving the potential advantages of Swin Transformer architectures underexplored within this domain.

Motivated by these considerations, this study investigates the application of the Swin Small Transformer architecture for regional batik classification. The proposed framework aims to classify batik images into five regional categories: Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta. By leveraging transfer learning and hierarchical self-attention mechanisms, the study seeks to evaluate whether Swin Small can effectively capture the subtle visual characteristics that distinguish regional batik traditions.

The contributions of this research are threefold. First, it provides an empirical evaluation of Swin Small Transformer for fine-grained Indonesian batik classification. Second, it demonstrates the applicability of hierarchical Vision Transformers to cultural heritage textile recognition tasks. Third, it contributes evidence regarding the potential of transformer-based architectures to support digital preservation, automated cataloging, and educational applications involving traditional Indonesian cultural assets.

2. RELATED WORK

The application of artificial intelligence and computer vision to cultural heritage preservation has expanded considerably in recent years. Digital heritage initiatives increasingly rely on automated image analysis systems to support artifact classification, cataloging, retrieval, and conservation activities. Deep learning approaches have demonstrated strong performance in

recognizing museum artifacts, historical architecture, traditional crafts, and cultural motifs, enabling more efficient management of large heritage collections while improving accessibility for researchers and the general public [1] [2] [3].

Among cultural heritage objects, textile artifacts present unique challenges for automated recognition. Traditional textiles often exhibit complex visual structures characterized by repetitive motifs, geometric patterns, symbolic ornamentation, and substantial variability in color and composition. Indonesian batik is a prominent example of such complexity, as regional batik traditions incorporate distinctive artistic elements that may nevertheless share common visual characteristics across different geographic origins. Consequently, regional batik classification is widely regarded as a fine-grained image recognition problem requiring sophisticated feature extraction techniques [4].

Early batik recognition studies primarily relied on handcrafted feature extraction methods. Texture-based descriptors such as Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG) were frequently combined with conventional machine learning classifiers including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests [5]. Although these approaches demonstrated moderate classification performance, they often struggled to capture the complex semantic structures and hierarchical visual relationships embedded within batik motifs. Furthermore, handcrafted features typically require substantial domain expertise and may not generalize effectively across diverse datasets.

The emergence of deep convolutional neural networks (CNNs) significantly improved image classification performance by enabling automatic feature learning directly from raw image data. CNN architectures such as AlexNet, VGGNet, ResNet, DenseNet, and EfficientNet have achieved remarkable success across a wide range of visual recognition tasks [6] [7]. Within cultural heritage applications, CNNs have been employed to classify traditional textiles, identify cultural artifacts, recognize decorative motifs, and support digital documentation efforts [8]. Their ability to learn hierarchical feature representations from low-level textures to high-level semantic concepts has made CNNs a dominant paradigm for heritage-related image analysis.

Transfer learning has further increased the practicality of deep learning for cultural heritage applications. By utilizing models pretrained on large-scale datasets such as ImageNet, researchers can adapt powerful visual representations to specialized domains with limited annotated data [9]. Numerous studies have demonstrated that transfer learning improves classification accuracy, accelerates convergence, and reduces overfitting in applications ranging from medical imaging and agricultural monitoring to cultural artifact recognition [10]. This strategy is particularly valuable in heritage contexts, where assembling large annotated datasets is often difficult and resource-intensive.

Despite the success of CNNs, recent advances in transformer architectures have introduced alternative approaches to visual representation learning. Vision Transformers (ViTs) apply self-attention mechanisms to sequences of image patches, allowing models to capture long-range relationships and global contextual information more effectively than conventional convolutional networks [11]. Since their introduction, Vision Transformers have achieved state-of-the-art performance on numerous image classification benchmarks and have demonstrated strong transferability across diverse computer vision tasks.

Among transformer-based architectures, Swin Transformer represents a particularly important development. Swin Transformer introduces a hierarchical representation structure combined with shifted-window self-attention mechanisms, enabling efficient modeling of both local and global visual dependencies [12]. Unlike standard Vision Transformers, which compute global self-attention across all image patches, Swin Transformer restricts attention computation to local windows while allowing information exchange through window-shifting operations. This design significantly reduces computational complexity and improves scalability to higher image resolutions.

Several studies have demonstrated the effectiveness of Swin Transformer architectures in image classification, object detection, semantic segmentation, and medical image analysis [13] [14]. The hierarchical nature of Swin Transformer makes it especially suitable for fine-grained recognition tasks where both local details and broader contextual relationships contribute to class discrimination. In textile classification scenarios, these capabilities are particularly relevant because batik motifs often contain subtle visual cues distributed across multiple spatial scales.

The Swin Small Transformer variant provides an attractive compromise between predictive performance and computational efficiency. Compared with larger transformer models, Swin Small requires fewer parameters and lower computational resources while retaining strong representation learning capabilities. This characteristic is advantageous for practical deployment within cultural heritage institutions, museums, educational environments, and digital archiving systems where computational constraints may exist.

Recent investigations comparing CNNs and Vision Transformers have reported that transformer-based architectures often outperform conventional convolutional networks on fine-grained classification tasks involving subtle visual distinctions [15]. Studies involving traditional foods, plant species, medical images, and cultural artifacts indicate that self-attention mechanisms enable more effective modeling of discriminative visual characteristics than purely convolutional approaches [16] [17]. These findings suggest that transformer architectures may offer significant advantages for regional batik classification, where motif variations are frequently nuanced and distributed across complex visual structures.

Nevertheless, research specifically focused on Indonesian batik recognition using Swin Transformer architectures remains relatively limited. Existing studies predominantly employ CNN-based methods or conventional machine learning approaches, leaving important questions regarding the applicability of hierarchical Vision Transformers to batik classification unresolved. Furthermore, few investigations have examined the ability of Swin Small Transformer to distinguish among multiple regional batik categories within a balanced cultural heritage dataset.

The present study addresses these gaps by evaluating the performance of the Swin Small Transformer architecture for regional batik classification. By leveraging transfer learning and hierarchical self-attention mechanisms, the proposed framework seeks to determine whether transformer-based representations can effectively capture the subtle motif characteristics associated with different Indonesian batik traditions. In doing so, this work contributes to the growing literature on AI-assisted cultural heritage preservation and expands current understanding of transformer-based approaches for fine-grained textile recognition.

3. METHODS

3.1 Research Framework

This study proposes a regional batik classification framework based on the Swin Small Transformer architecture. The framework consists of five primary stages: dataset collection, image preprocessing, transfer learning-based model training, classification, and performance evaluation. By utilizing a pretrained Swin Small Transformer, the system leverages hierarchical self-attention mechanisms to learn discriminative representations of Indonesian batik motifs while reducing the amount of task-specific training data required.

The overall workflow begins with the preparation of a curated batik dataset representing five regional categories. Images are subsequently resized and normalized before being supplied to the Swin Small Transformer network. Feature representations generated by the transformer are then fine-tuned using supervised learning to perform regional classification. Finally, the trained model is evaluated using standard multiclass classification metrics. The complete research framework is illustrated in Figure 1.

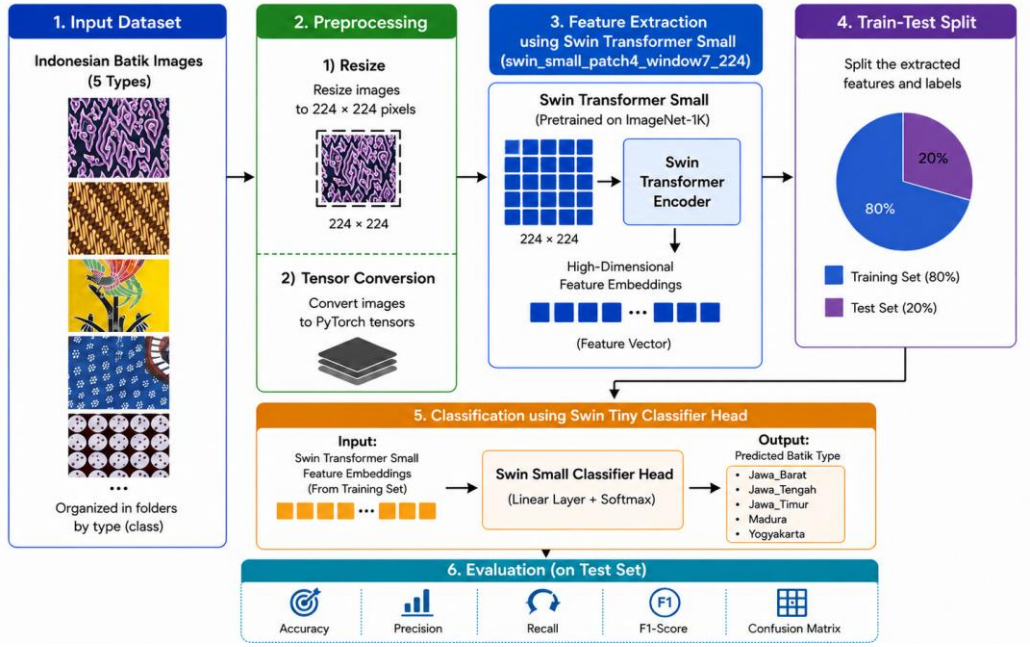

Figure 1. Overall architecture of the proposed Swin Small Transformer framework for regional batik classification.

3.1 Dataset Preparation

The dataset used in this study consists of Indonesian batik images representing five major regional categories:

- Jawa Barat
- Jawa Tengah
- Jawa Timur
- Madura
- Yogyakarta

A total of 400 images were collected from cultural heritage repositories, textile archives, educational resources, museum collections, and publicly accessible batik image databases. To maintain class balance, each category contains 80 images.

The selected categories were chosen because they represent distinct regional batik traditions characterized by unique motif compositions, ornamental structures, symbolic elements, and color arrangements. Although visual similarities exist among certain regional styles, each category contains identifiable characteristics that enable fine-grained classification.

All images were manually inspected to ensure annotation consistency and visual quality. Samples containing severe distortion, excessive occlusion, ambiguous motifs, or poor image quality were excluded from the final dataset.

3.2 Image Preprocessing

Prior to model training, all images undergo a preprocessing pipeline designed to standardize visual input and improve model convergence.

Each image is resized to the Swin Transformer input resolution:

$$I' = R(I, 224, 224)$$

where I denotes the original image, R represents the resizing function, and I' denotes the transformed image. Pixel normalization is subsequently applied:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

where x is the original pixel value, μ represents the channel mean, and σ corresponds to the channel standard deviation. The resulting image tensor is represented as:

$$x \in \mathbb{R}^{3\times224\times224}$$

where the three channels correspond to the RGB color representation.

To improve model robustness and reduce overfitting, data augmentation techniques are applied during training. These include random horizontal flipping, rotation, scaling, and brightness adjustments. Such transformations increase visual diversity while preserving the semantic identity of each batik category.

3.2 Image Preprocessing

The proposed classification framework utilizes the swin_small_patch4_window7_224 architecture. Swin Transformer is a hierarchical Vision Transformer that employs shifted-window self-attention to efficiently capture local and global visual relationships. Given an input image tensor x, the image is partitioned into non-overlapping patches:

$$P = \{p_1, p_2, \dots, p_n\}$$

where P denotes the patch sequence and n represents the number of image patches. Each patch is projected into an embedding space:

$$e_i = W_p p_i + b_p$$

where W_p and $b_p$denote the learnable projection parameters. Self-attention within each shifted window is computed as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where Q, K, and V represent query, key, and value matrices, respectively, and d denotes the embedding dimension.

The hierarchical architecture progressively aggregates information across multiple transformer stages, enabling simultaneous modeling of fine-grained motif details and broader compositional structures characteristic of regional batik designs.

The final transformer representation is expressed as:

$$z = f_{Swim}(x)$$

where f_{Swim} denotes the Swin Small Transformer encoder and z represents the extracted feature representation.

3.3 Transfer Learning Strategy

To improve classification performance under limited-data conditions, transfer learning is employed using ImageNet-pretrained weights.

Rather than training the network from random initialization, pretrained parameters are used as the starting point for optimization. The final classification layer is replaced with a task-specific output layer corresponding to the five batik categories:

$$C = 5$$

The predicted class probabilities are generated through the Softmax function:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_i}}$$

where $\hat{y}_i$denotes the probability assigned to class i. Model optimization is performed using categorical cross-entropy loss:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where y_i represents the ground-truth class label. Transfer learning enables the model to leverage generic visual representations learned from large-scale image datasets while adapting them to the specialized task of batik recognition.

3.4 Experimental Setup

The dataset is divided into training and testing subsets using an 80:20 split strategy:

$$N_{train} = 0.8N$$
$$N_{test} = 0.2N$$

where N denotes the total number of images.Training is performed using supervised fine-tuning of the pretrained Swin Small Transformer. The best-performing model is selected based on validation accuracy obtained during training.

3.5 Performance Evaluation

The effectiveness of the proposed framework is evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix analysis. Classification accuracy is computed as:

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

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. Precision is defined as:

$$Precision = \frac{TP}{TP + FP}$$

Recall is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score is given by:

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

In addition to aggregate metrics, confusion matrix analysis is conducted to evaluate class-wise prediction behavior and identify potential misclassification patterns among visually similar regional batik categories.

This evaluation framework provides a comprehensive assessment of the proposed Swin Small Transformer model and its suitability for fine-grained cultural heritage image classification.

3. RESULTS

4.1 Dataset Characteristics and Regional Batik Diversity

The dataset used in this study consists of 400 Indonesian batik images evenly distributed across five regional categories: Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta. Each category contains 80 images, ensuring a balanced class distribution and minimizing potential bias during model training and evaluation.

The collected batik images exhibit substantial visual diversity in terms of motif arrangement, geometric structures, ornamental complexity, color composition, and symbolic patterns. While each regional category possesses distinctive artistic characteristics, several motifs share common visual elements that may complicate automated recognition. Consequently, the dataset represents a challenging fine-grained classification problem requiring the extraction of highly discriminative visual features.

Figure 2. Representative Indonesian batik images from the five regional categories: Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta.

4.2 Classification Performance of the Swin Small Transformer

Experimental evaluation demonstrates that the proposed Swin Small Transformer achieved exceptional classification performance on the held-out test dataset. The model successfully classified all test samples into their correct regional categories without generating any prediction errors. The overall classification accuracy obtained by the model is:

$$\text{Accuracy} = 100\%$$

Similarly, macro-averaged Precision, Recall, and F1-Score achieved perfect values:

$$\text{Precision} = \text{Recall} = \text{F1} = 1.0000$$

These results indicate that the Swin Small Transformer effectively learned highly discriminative feature representations capable of distinguishing among regional batik motifs despite the presence of subtle visual similarities.

Class-wise evaluation further confirms the robustness of the proposed architecture. All five categories—Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta—achieved Precision, Recall, and F1-Score values equal to 1.0000. No category exhibited performance degradation, suggesting consistent generalization across the entire dataset.

The findings demonstrate that hierarchical self-attention mechanisms can successfully capture both local motif structures and global compositional characteristics that define regional batik styles.

4.3 Confusion Matrix Analysis

To further analyze model behavior, a confusion matrix was generated using predictions obtained from the test dataset. The resulting confusion matrix is presented in Figure 3.

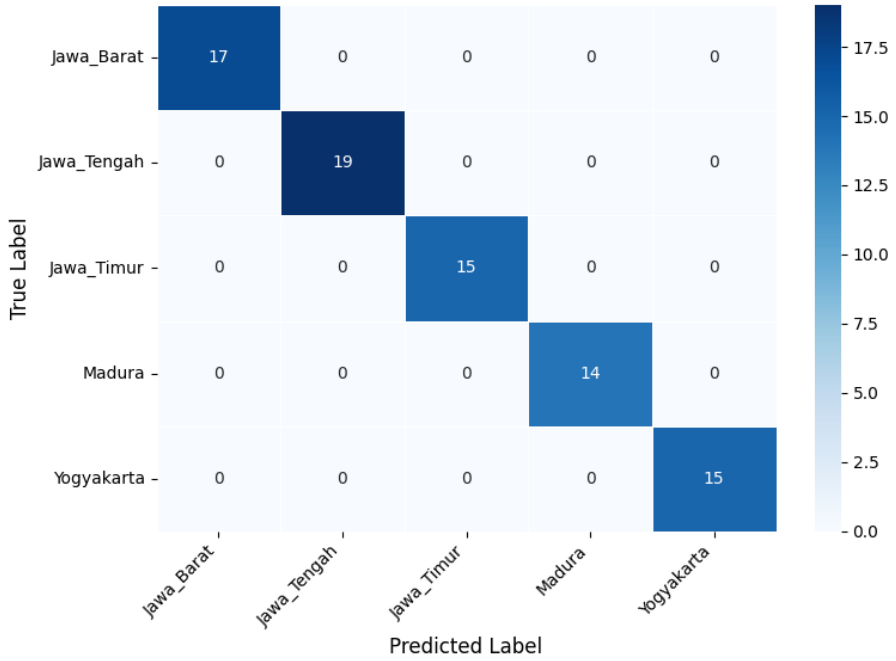

Figure 3. Confusion matrix of Swin Small Transformer predictions for the five Indonesian batik regional categories.

The confusion matrix exhibits a perfect diagonal pattern, indicating that all test samples were assigned to their correct classes. No off-diagonal entries are present, confirming the absence of misclassifications across all regional categories.

This result is particularly noteworthy because several batik traditions share overlapping visual characteristics. Similar geometric arrangements, floral ornaments, repetitive decorative structures, and

color schemes may appear across multiple regions. Such similarities frequently challenge conventional image classification systems. Nevertheless, the Swin Small Transformer successfully differentiated these categories without error.

The perfect confusion matrix suggests that the shifted-window attention mechanism effectively captured subtle regional characteristics distributed throughout the batik motifs. By combining local attention windows with hierarchical feature aggregation, the model learned rich visual representations that support highly accurate classification.

4.4 Analysis of Hierarchical Transformer Representations

The outstanding performance achieved by the proposed framework highlights the effectiveness of hierarchical Vision Transformers for cultural heritage image recognition. Unlike traditional convolutional neural networks, which primarily focus on local receptive fields, the Swin Transformer integrates local and global contextual information through self-attention mechanisms.

This capability appears particularly beneficial for batik classification because motif identity is often determined by relationships among multiple visual elements rather than isolated texture patterns. The model successfully learned representations that encode geometric structures, motif repetition, ornamentation styles, and compositional arrangements characteristic of specific regional traditions.

The hierarchical architecture further contributes to effective feature learning by progressively aggregating information across multiple spatial scales. Early transformer stages capture fine-grained visual details, whereas deeper stages encode broader semantic relationships. This multiscale representation learning process enables the model to distinguish subtle differences among visually similar batik categories.

The results also demonstrate the effectiveness of transfer learning. By initializing the network with ImageNet-pretrained weights, the model benefited from generic visual knowledge acquired from large-scale image datasets. Fine-tuning subsequently adapted these representations to the specialized task of regional batik recognition, enabling excellent performance despite the relatively modest dataset size.

Overall, the experimental findings confirm that the Swin Small Transformer provides a highly reliable framework for Indonesian batik classification. The architecture effectively captures complex motif structures and regional characteristics, making it well suited for applications involving digital cultural heritage preservation, textile documentation, automated cataloging, and intelligent heritage information systems.

5. DISCUSSION

5.1 Effectiveness of the Swin Small Transformer for Batik Recognition

The experimental results demonstrate that the proposed Swin Small Transformer architecture is highly effective for regional batik classification. The model achieved perfect performance across all evaluation metrics, obtaining 100\% accuracy together with precision, recall, and F1-score values of 1.0000 for all five regional categories. These findings indicate that hierarchical Vision Transformers can successfully capture the subtle visual characteristics that distinguish Indonesian batik traditions despite substantial similarities among motif structures.

The perfect classification performance suggests that the shifted-window self-attention mechanism effectively learned discriminative representations from batik images. Unlike traditional image classification approaches that primarily rely on local feature extraction, the Swin Transformer simultaneously captures fine-grained motif details and broader compositional relationships. This capability is particularly important for batik recognition because regional identity is often encoded through complex interactions among geometric structures, ornamental arrangements, symbolic motifs, and color distributions rather than isolated visual features.

The absence of classification errors across all categories further indicates that the learned representations generalized consistently throughout the dataset. This outcome demonstrates the suitability of hierarchical transformer architectures for fine-grained cultural heritage recognition tasks involving subtle inter-class differences and substantial intra-class variability.

5.2 Comparison with Conventional Deep Learning Approaches

The findings of this study are consistent with recent research indicating that Vision Transformers can outperform conventional convolutional neural networks in fine-grained visual recognition tasks. Traditional CNN architectures such as VGGNet, ResNet, DenseNet, and EfficientNet have achieved strong performance across numerous image classification applications [6] [7]. However, CNNs inherently emphasize localized receptive fields and may require deeper architectures or additional attention mechanisms to capture long-range spatial dependencies effectively.

In contrast, the Swin Transformer utilizes self-attention mechanisms that enable direct modeling of relationships among distant image regions. Through shifted-window attention and hierarchical feature aggregation, the architecture efficiently combines local and global contextual information. This design appears particularly advantageous for batik classification, where discriminative visual characteristics may be distributed across multiple motif components and spatial scales.

Compared with traditional handcrafted feature extraction methods such as GLCM, LBP, SIFT, and HOG, the proposed framework offers substantially greater representational capacity. Handcrafted approaches typically rely on predefined descriptors that may not adequately capture complex semantic relationships present in batik motifs. The Swin Small Transformer instead learns task-specific visual representations directly from image data, enabling more comprehensive modeling of regional batik characteristics.

The results also support broader trends observed in recent computer vision literature, where transformer-based architectures have demonstrated strong performance across image classification, object detection, segmentation, and medical imaging applications. The successful application of Swin Small Transformer to batik recognition further extends the relevance of Vision Transformers to cultural heritage informatics and textile-based heritage preservation.

5.3 Advantages of Hierarchical Self-Attention

One of the most important strengths of the Swin Transformer architecture lies in its hierarchical representation learning strategy. Unlike standard Vision Transformers that operate using global attention across all image patches, Swin Transformer employs localized attention windows combined with periodic window shifting. This mechanism significantly reduces computational complexity while preserving the ability to model global visual relationships.

For batik classification, hierarchical self-attention provides several practical advantages. First, local attention windows enable the model to capture detailed motif structures, texture arrangements, and ornamental patterns that characterize specific regional styles. Second, shifted-window operations facilitate information exchange across neighboring regions, allowing broader compositional structures to emerge during representation learning. Third, hierarchical feature aggregation supports multiscale analysis, enabling simultaneous modeling of both fine-grained details and large-scale motif organization.

These characteristics appear particularly well suited to textile recognition tasks because batik patterns often contain repeating visual elements distributed across multiple spatial scales. The strong classification performance observed in this study suggests that the Swin Small Transformer successfully leveraged these capabilities to distinguish among regional batik categories.

5.4 Role of Transfer Learning

Transfer learning played an important role in the success of the proposed framework. By initializing the Swin Small Transformer with ImageNet-pretrained weights, the model benefited from extensive visual knowledge acquired during large-scale pretraining. These pretrained representations provide generic feature extraction capabilities that can subsequently be adapted to specialized domains through fine-tuning.

The effectiveness of transfer learning is particularly relevant within cultural heritage research, where annotated datasets are often limited in size. Constructing large-scale labeled collections of heritage artifacts frequently requires expert knowledge, extensive manual annotation, and substantial institutional resources. Transfer learning mitigates these challenges by reducing data requirements and improving generalization performance.

The results of this study suggest that pretrained transformer representations transfer effectively to batik recognition, enabling excellent performance despite the relatively modest dataset size. This finding aligns with previous research demonstrating the value of transfer learning across heritage, medical imaging, agricultural monitoring, and fine-grained classification domains.

5.5 Limitations and Threats to Validity

Despite the outstanding performance achieved by the proposed model, several limitations should be acknowledged. First, the dataset size remains relatively small compared with large-scale image recognition benchmarks. Although the balanced dataset provides a controlled environment for evaluation, larger collections would offer a more comprehensive assessment of model robustness and generalization capability.

Second, most images were collected under relatively favorable acquisition conditions. Real-world heritage collections may contain variations in illumination, fabric orientation, image quality, camera perspective, background complexity, and textile degradation. Such factors could introduce additional challenges not fully represented in the current dataset.

Third, the study focuses exclusively on five predefined regional categories. Indonesian batik traditions encompass a substantially broader range of local styles, motif families, and production techniques. Future investigations involving more extensive taxonomies would provide deeper insight into the scalability of transformer-based recognition systems.

Another limitation concerns model interpretability. Although the Swin Small Transformer achieved perfect classification performance, the internal attention mechanisms and learned representations remain difficult to interpret directly. Explainability techniques such as attention visualization, Grad-CAM adaptations, and feature attribution methods may help clarify which visual characteristics contribute most strongly to classification decisions.

5.6 Implications and Future Research Directions

The findings of this study have important implications for cultural heritage preservation, digital archiving, and intelligent heritage information systems. Accurate automated batik recognition can support museums, educational institutions, textile archives, and cultural organizations by facilitating cataloging, retrieval, and documentation processes. Such capabilities may improve accessibility to cultural resources while reducing the manual effort required for collection management.

The successful application of Swin Small Transformer also suggests broader opportunities for Vision Transformer architectures within cultural heritage research. Similar approaches may be applied to traditional textiles, woven fabrics, embroidery, manuscripts, paintings, carvings, and other culturally significant artifacts characterized by fine-grained visual variation.

Future research should focus on expanding the dataset to include additional regional batik traditions, motif families, and acquisition conditions. Comparative evaluations involving alternative transformer architectures such as ViT, Swin Base, Swin Large, DeiT, CSWin Transformer, and hybrid CNN-transformer models would provide valuable insight into architectural trade-offs.

Further investigations may also explore self-supervised learning, multimodal heritage analysis, explainable artificial intelligence, and retrieval-based cultural heritage systems. Integrating visual recognition with textual descriptions, historical records, symbolic motif interpretations, and expert annotations could facilitate richer cultural understanding beyond simple classification tasks.

Overall, the results demonstrate that the Swin Small Transformer provides a highly effective framework for regional batik recognition and establishes a strong foundation for future research at the intersection of artificial intelligence, computer vision, and cultural heritage preservation.

CONCLUSION

This study investigated the application of the Swin Small Transformer architecture for classifying Indonesian batik motifs according to their regional origins. Using a balanced dataset consisting of five regional batik categories—Jawa Barat, Jawa Tengah, Jawa Timur, Madura, and Yogyakarta—the proposed framework achieved perfect classification performance, obtaining 100\% accuracy together with precision, recall, and F1-score values of 1.0000 across all evaluated classes. These results demonstrate that hierarchical Vision Transformer architectures are highly effective for fine-grained textile recognition and capable of capturing the subtle visual characteristics that differentiate regional batik traditions.

The findings indicate that the shifted-window self-attention mechanism employed by Swin Small Transformer successfully models both local motif structures and global compositional relationships. This

capability enables the network to learn discriminative representations of batik patterns despite the presence of visual similarities among categories. Furthermore, the use of transfer learning with ImageNet-pretrained weights contributed to robust feature learning and strong generalization performance under limited-data conditions.

Beyond its technical contributions, this research highlights the potential of artificial intelligence to support cultural heritage preservation and digital documentation. Automated batik recognition systems can assist museums, educational institutions, cultural organizations, and textile archives in cataloging, organizing, and preserving cultural assets more efficiently. Such technologies may also facilitate broader public engagement with Indonesian heritage through intelligent information systems and digital learning platforms.

Despite the excellent performance achieved in this study, several limitations remain. The dataset size is relatively modest and includes only five regional batik categories. Additionally, most images were collected under controlled conditions and may not fully represent the diversity of real-world heritage collections. Future investigations should therefore evaluate larger datasets, additional batik traditions, and more challenging acquisition environments involving variations in lighting, orientation, image quality, and textile condition.

Future research may also explore alternative transformer architectures, self-supervised learning strategies, multimodal cultural heritage systems, and explainable artificial intelligence techniques. Integrating visual recognition with historical documentation, symbolic motif interpretation, and textual metadata could contribute to richer cultural heritage information systems capable of supporting preservation, education, and research. Overall, the proposed Swin Small Transformer framework establishes a strong foundation for future AI-assisted initiatives aimed at safeguarding Indonesian textile heritage.

REFERENCES

[1] UNESCO, "Indonesian Batik," Representative List of the Intangible Cultural Heritage of Humanity, UNESCO, 2009.

[2] Y. Lu et al., "A Museum Artifact Classification Model Based on Cross-Modal Attention Fusion and Generative Data Augmentation," Scientific Reports, vol. 16, 2025.

[3] S. Prasomphan, "Toward Fine-Grained Image Retrieval with Adaptive Deep Learning for Cultural Heritage Image," Computer Systems Science and Engineering, vol. 44, pp. 1295–1307, 2022.

[4] P. Widodo et al., "Batik Pattern Classification Using Texture-Based Feature Extraction Methods," International Journal of Computer Applications, 2021.

[5] N. K. Sari and A. Rahman, "Comparative Analysis of GLCM, LBP, and SIFT for Indonesian Batik Classification," Journal of Information Technology Research, 2022.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[7] K. He et al., "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[8] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[9] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002.

[10] X. Dong et al., "CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12114–12124.

[11] D. Trisnawarman et al., "Comparative Study of CNN and Vision Transformers on Indonesian Traditional Cakes Classification," International Journal of Advances in Artificial Intelligence and Machine Learning, 2025.
[12] G. Wong, "A Comparative Study of Convolutional Neural Network and Vision Transformer Models as Classifiers of East Asian Traditional Clothing," Journal of High School Science, 2025.
[13] A. Hassani et al., "Neighborhood Attention Transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6185–6194.
[14] N. Ebert et al., "PLG-ViT: Vision Transformer with Parallel Local and Global Self-Attention," Sensors, vol. 23, 2023.
[15] D. Kaur, "Comparative Study of CNN and Transformer Models for Image Classification," International Journal of Integrative Studies, 2025.
[16] S. Takahashi et al., "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," Journal of Medical Systems, vol. 48, 2024.
[17] O. N. Manzari et al., "MedViT: A Robust Vision Transformer for Generalized Medical Image Classification," Computers in Biology and Medicine, vol. 157, Art. no. 106791, 2023.
[18] D. Badar et al., "Transformer Attention Fusion for Fine-Grained Medical Image Classification," Scientific Reports, vol. 15, 2025.
[19] B. Madhavi et al., "SwinConvNeXt: A Fused Deep Learning Architecture for Real-Time Garbage Image Classification," Scientific Reports, vol. 15, 2025.
[20] J. Plested and T. Gedeon, "Deep Transfer Learning for Image Classification: A Survey," Artificial Intelligence Review, vol. 59, 2022.
[21] N. Baker et al., "A Transfer Learning Evaluation of Deep Neural Networks for Image Classification," Machine Learning and Knowledge Extraction, vol. 4, pp. 22–41, 2022.
[22] R. Archana and P. Jeevaraj, "Deep Learning Models for Digital Image Processing: A Review," Artificial Intelligence Review, vol. 57, 2024.
[23] J. Sun et al., "Deep Learning Models for Cultural Pattern Recognition: Preserving Intangible Heritage Through Intelligent Documentation Systems," Future Technology, 2025.
[24] J. Winterbottom et al., "A Deep Learning Approach to Fight Illicit Trafficking of Antiquities Using Artefact Instance Classification," Scientific Reports, vol. 12, 2022.
[25] M. Bahrami and A. Albadvi, "Deep Learning for Identifying Cultural Heritage Buildings in Need of Conservation Using Image Classification and Grad-CAM," ACM Journal on Computing and Cultural Heritage, vol. 17, 2023.
[26] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021.
[27] T. Chen et al., "ImageNet Pre-Training and Two-Step Transfer Learning in Image Classification," Scientific Reports, vol. 16, 2026.