Vol. 8, No. 3, June 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Handcrafted Feature Ablation for Batik Nitik Classification Under Provenance-Aware Evaluation

Aji Priyambodo[1,4*], Rizal Isnanto[2], Ridwan Sanjaya[3]

[1,2] Doctoral Program of Information Systems, Postgraduate School, Diponegoro University, Semarang, Indonesia

[3] Department of Information Systems, Faculty of Computer Science, Soegijapranata Catholic University, Semarang, Indonesia

[4] Information Systems and Technology Study Program, Semarang Institute of Technology and Business, Semarang, Indonesia

Received:
October 19, 2025
Revised:
May 31 2026
Accepted:
June 24, 2026
Published:
June 27, 2026

Corresponding Author:
Author Name*:
Aji Priyambodo
Email*:
ajippro@students.undip.ac.id

DOI:
10.63158/journalisi.v8i3.1679

Abstract. This study re-examines Batik Nitik classification using a leakage-safe provenance-aware evaluation protocol to determine which handcrafted descriptors make a substantive contribution to performance and whether saturated results persist after provenance-based partitioning. Batik Nitik 960 was represented using Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Gray-Level Co-occurrence Matrix (GLCM) descriptors, and grayscale intensity moments. Descriptor ablation, classifier benchmarking, cosine-similarity baselines, four-setting leave-one-provenance-group-out sensitivity analysis, and a supplementary image-level split comparison were evaluated using in-pipeline preprocessing. All HOG-containing feature sets achieved 0.9833 cross-validation accuracy and 1.0000 hold-out accuracy. On fused features, SVM, KNN (Euclidean), and KNN (cosine) achieved 1.0000 hold-out accuracy, while Random Forest reached 0.9958. Raw-pixel, HOG-only, and fused-feature cosine baselines also reached 1.0000 hold-out accuracy. A supplementary image-level HOG-SVM split also produced 1.0000 accuracy. This study contributes a provenance-aware benchmark diagnosis for Batik Nitik classification by identifying HOG as the strongest standalone handcrafted descriptor and by cautioning against deployment-ready interpretation of saturated closed-set accuracy.

Keywords: Batik Nitik classification, Handcrafted features, HOG, Provenance-aware evaluation, Benchmark saturation.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Automatic batik motif recognition remains a significant computer-vision problem because it connects cultural-heritage preservation with practical requirements in digital cataloguing, image retrieval, education, museum documentation, and creative-industry support [1–7]. Batik Nitik is particularly relevant for this purpose because its repeated geometric micro-patterns generate subtle within-family variation: motifs may appear globally similar while differing in local structure, edge organisation, and fine texture. These characteristics make Batik Nitik an appropriate case for evaluating how alternative visual representations behave under fine-grained classification conditions.

Prior studies on batik recognition have examined convolutional neural networks, transfer learning, deep autoencoders, ROI-sensitive preprocessing, augmentation-driven expansion, and colour-texture descriptor combinations [1], [4–10], [11–19]. The Batik Nitik 960 dataset has also been introduced as a benchmark for classification, retrieval, and generative exploration [20], and subsequent Batik Nitik studies have reported model-oriented evaluations on the same collection [3], [16], [21]. Collectively, these studies demonstrate that batik recognition has become an active area of cultural-computing research. At the same time, they reveal substantial variation in dataset construction, preprocessing practice, augmentation strategy, and evaluation strictness, which complicates methodological comparison. Related textile-vision studies likewise emphasise the importance of texture representation, colour-texture analysis, and robust inspection pipelines for patterned-fabric imagery [22–26].

Handcrafted descriptors remain methodologically relevant within this literature for three principal reasons. First, they provide a transparent relationship between visual cues and model behaviour. HOG captures oriented edge structure [27], LBP captures local threshold-pattern distributions [28], GLCM summarises second-order co-occurrence statistics [29], and intensity moments provide a compact description of grayscale intensity distribution [10]. Second, handcrafted pipelines remain practical in data-constrained settings, where deep models may be unnecessarily complex or sensitive to limited training diversity. Third, handcrafted descriptors are well suited to diagnostic

ablation because performance differences can be interpreted in relation to specific representational cues.

Despite this progress, two methodological issues remain insufficiently addressed in the batik-classification literature. First, many studies emphasise the best-performing final configuration without determining whether each descriptor in a combined representation contributes materially to the observed result. As a consequence, it is often unclear whether multi-feature pipelines are genuinely complementary or whether some components are redundant. Second, evaluation protocols are not always described in sufficient detail to exclude optimistic estimates caused by preprocessing leakage, augmentation overlap, or hidden relatedness between training and test samples [30].

A related issue in the broader computer-vision literature concerns the treatment of augmented data during evaluation. Although augmentation can improve learning, it can also compromise experimental validity when transformed variants of the same underlying source image appear in both training and test partitions [30, 31]. This issue is especially important for datasets constructed from rotation, flip, or crop variants. In such cases, provenance-aware grouping is not merely a technical preference but a requirement for defensible internal validation [30, 31]. This concern is directly relevant to Batik Nitik 960 because the dataset includes rotation-derived variants of the same underlying source samples [20].

Preliminary computational results indicated that HOG-based representations were highly effective for Batik Nitik classification. A subsequent audit of the dataset structure and computational workflow, however, showed that the problem required a stricter experimental design. Very high accuracy in this setting may reflect benchmark saturation under a permissive split rather than robust descriptor behaviour across independent sources. The central research question is therefore as follows: which handcrafted descriptors contribute most under a leakage-safe provenance-aware evaluation protocol?

Accordingly, this study revisits Batik Nitik classification using a leakage-safe provenance-aware evaluation protocol and uses that setting to examine descriptor behaviour more rigorously. The contribution is threefold. First, the study implements provenance-aware

evaluation to reduce optimistic bias caused by overlap among related image variants. Second, it conducts descriptor ablation to clarify the relative contribution of HOG, LBP, GLCM, and intensity moments. Third, it interprets saturated performance cautiously by showing that several simple cosine-based baselines also reach ceiling-level hold-out performance on the current closed-set dataset.

2. METHODS

2.1 Dataset

This study uses Batik Nitik 960, a balanced image collection comprising 960 grayscale-convertible batik images distributed across 60 motif classes, with 16 images per class [20]. According to the accompanying dataset publication, each class originates from four source samples, and each source sample is expanded through rotations of 90°, 180°, and 270°. This construction is analytically important because a random image-level split may distribute augmentation siblings derived from the same source sample across training and test partitions, thereby producing overly optimistic estimates of generalisation.

All images were processed as single-channel grayscale inputs and resized to 64 × 64 pixels. The grayscale formulation was retained to preserve comparability with the computational workflow and to focus the analysis on structural pattern information rather than chromatic variability. Consequently, the descriptors analysed in this study should be interpreted primarily as structural and intensity-based representations rather than colour-dependent representations.

2.2 Image preprocessing and handcrafted descriptors

Four descriptor families were extracted. HOG was used to represent oriented gradient structure, which is particularly relevant for repeated geometric motifs [27]. LBP was used to represent local threshold-pattern distributions [28]. GLCM descriptors were used to summarise second-order texture statistics [29]. Grayscale intensity moments provided a compact description of first-, second-, and third-order distributional shape under the grayscale formulation. No image augmentation was applied during model training so that descriptor behaviour could be interpreted without being confounded by augmentation-driven expansion.

Descriptor settings were fixed throughout the study. HOG used 9 orientations, 8 × 8 pixels per cell, 2 × 2 cells per block, and L2-Hys block normalisation, yielding 1764 features per image [27]. LBP used P = 8, R = 1, and uniform mapping, followed by a 10-bin normalised histogram [28]. GLCM descriptors were computed from 32-level quantised images across four directions; six summary statistics were extracted per direction (contrast, dissimilarity, homogeneity, energy, correlation, and ASM), yielding 24 features [29]. Grayscale intensity moments consisted of mean, standard deviation, and skewness, yielding three features. The resulting dimensionalities were 1774 for HOG + LBP, 1798 for HOG + LBP + GLCM, and 1801 for the fused handcrafted representation.

2.3 Provenance-aware evaluation design

The central methodological component is provenance-aware grouping. A provenance group was defined as one underlying source sample and all of its rotation-derived variants. Group membership was reconstructed directly from the released filenames by removing the optional rotation suffix and preserving the class-specific source index. Table 1 provides an illustrative example of this reconstruction logic. For hold-out testing, one group per class was reserved, producing 720 development images and 240 hold-out images while ensuring zero source-group overlap between partitions. The hold-out group for each class was selected using a fixed random seed of 42 to preserve reproducibility.

Internal validation on the development portion used StratifiedGroupKFold with three folds. Stratification preserved class balance, while grouping preserved provenance integrity. Three folds were selected because each class provides only three development groups after hold-out reservation. All feature-scaling steps were executed within the relevant scikit-learn pipeline so that normalisation parameters were estimated only from the corresponding training fold or training split. PCA was fitted on the development feature matrix and used only for post hoc visualisation; it was not used for training, hyperparameter selection, or model comparison. Figure 1 summarises the end-to-end experimental pipeline. Because provenance groups are reconstructed deterministically from the released filenames, the corresponding group manifest and fixed seed-based hold-out indices can be exported from the accompanying computational notebook as supplementary material.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

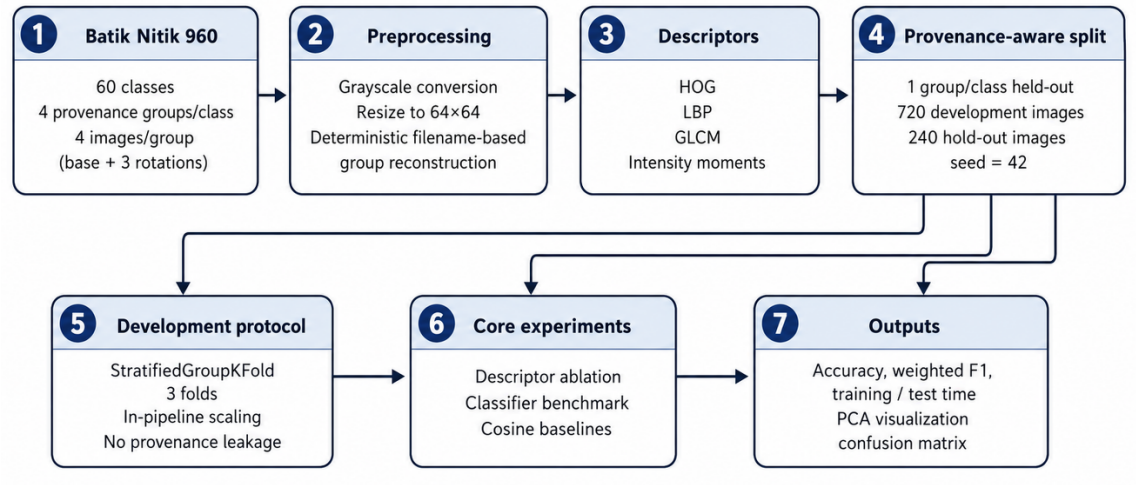

Figure 1. Overview of the provenance-aware experimental pipeline used in this study

Table 1. Illustrative reconstruction of four provenance groups for one Batik Nitik class and their rotation-derived variants

Class label	Source-group ID	Base image	Rotation-derived variants	Images per group
Sekar Kemuning	G1	1 Sekar Kemuning 1.jpg	base + rotate_90 + rotate_180 + rotate_270	4
Sekar Kemuning	G2	1 Sekar Kemuning 2.jpg	base + rotate_90 + rotate_180 + rotate_270	4
Sekar Kemuning	G3	1 Sekar Kemuning 3.jpg	base + rotate_90 + rotate_180 + rotate_270	4
Sekar Kemuning	G4	1 Sekar Kemuning 4.jpg	base + rotate_90 + rotate_180 + rotate_270	4

2.4 Classifier benchmarking and cosine baselines

Descriptor ablation was performed with an RBF-kernel support vector machine because it provides a stable and widely used nonlinear baseline for medium-dimensional handcrafted features [32]. The ablation model used C = 10 and gamma = 'scale'. To determine whether the fused representation materially changed the overall pattern, additional benchmarks were conducted with Random Forest [33], KNN (Euclidean) [34], and KNN (cosine) [34]. The Random Forest model used 300 trees with default Gini splitting and a fixed random_state of 42, whereas KNN used k = 3 in both Euclidean and cosine settings. No data augmentation was applied during model development or evaluation.

The cosine experiments were designed as supplementary interpretive baselines rather than substitutes for the primary ablation analysis. Three cosine settings were evaluated: raw pixels with KNN (cosine), HOG-only with KNN (cosine), and fused handcrafted features with KNN (cosine). These baselines were included to assess whether cosine similarity itself could account for the ceiling-level hold-out results and whether descriptor-level interpretation remained informative under the stricter protocol. In all cosine settings, input vectors were standardised within the relevant training fold or split, class labels were encoded once and reused consistently across experiments, and no augmentation or external calibration step was introduced.

2.5 Implementation details and metrics

The experiments were executed through the computational notebook using Python 3.12.13 with OpenCV, scikit-image, scikit-learn, NumPy, pandas, Matplotlib, and statsmodels in a Linux/Google-Colab-style CPU environment. The recorded runtime used an Intel(R) Xeon(R) CPU @ 2.20 GHz, one physical core, two logical CPUs, and 12.67 GB RAM. The global random seed was fixed at 42 for deterministic provenance-aware splits and model initialisation. Performance was summarised using cross-validation accuracy, cross-validation standard deviation, hold-out accuracy, weighted F1-score, training time, and test time. Class-level accuracy and class-level F1-score were also computed on the provenance-aware hold-out set for a representative best-performing model to verify whether saturated aggregate performance concealed class-specific degradation. For perfect hold-out accuracy, an exact 95% Clopper-Pearson confidence interval was reported to avoid treating an observed value of 1.0000 as complete statistical certainty. Train and test times were measured once for the classifier fit and predict stages only;

image loading, grayscale conversion, resizing, handcrafted feature extraction, PCA visualisation, and non-classifier preprocessing stages were excluded from these elapsed-time measurements.

The workflow was intentionally constrained. No deep-network fine-tuning, denoising stage, augmentation regime, or data-driven feature-selection procedure was introduced beyond the predefined handcrafted configurations. This design isolates the effects of descriptor choice and evaluation protocol, which are the central analytical concerns of the study. Table 2 summarises feature dimensionality, baseline hyperparameter settings, hardware details, and reproducibility information so that the descriptor-ablation and classifier-benchmark results can be interpreted in relation to representation size and implementation context.

Table 2. Descriptor dimensionality, baseline settings, and reproducibility details.

Component	Setting	Dimensionality / value	Remark
HOG	9 orientations; 8x8 cell; 2x2 block; L2-Hys	1764	Primary standalone descriptor
LBP	P=8; R=1; uniform histogram	10	Local micro-texture
GLCM	32 gray levels; 4 directions; 6 statistics	24	Second-order texture
Intensity moments	Mean; standard deviation; skewness	3	Grayscale moments
Fused features	HOG + LBP + GLCM + moments	1801	Used for classifier benchmark
SVM / RF / KNN	C=10, gamma=scale / 300 trees / k=3	Seed=42	Same labels across experiments
Software / hardware environment	Python 3.12.13; OpenCV, scikit-image, scikit-learn, NumPy, pandas,	CPU: Intel(R) Xeon(R) @ 2.20GHz; 1 physical / 2 logical cores	RAM: 12.67 GB total; 11.08 GB available; Linux Colab runtime

Component	Setting	Dimensionality / value	Remark
	Matplotlib, statsmodels		
Grouped hold-out split	1 provenance group per class reserved	720/240 images (75:25)	Hold-out groups selected with seed = 42
Timing protocol	Single run per final model	Classifier fit / predict stages only	Image loading, preprocessing, feature extraction, PCA, and plotting excluded
PCA usage	PCA fitted on development matrix; hold-out projected using fitted transformation	Descriptive only	Not used for model selection

3. RESULTS AND DISCUSSION

3.1 Feature-ablation results

Table 3 reports the feature-ablation benchmark, and Figure 2 visualises the same pattern. All feature configurations containing HOG achieved 1.0000 hold-out accuracy, whereas the corresponding grouped cross-validation accuracy was 0.9833 with a standard deviation of 0.0289. In contrast, the non-HOG descriptors were substantially weaker when used in isolation. LBP reached 0.6250 cross-validation accuracy and 0.7458 hold-out accuracy; GLCM reached 0.7000 cross-validation accuracy and 0.8542 hold-out accuracy; and intensity moments reached 0.4764 cross-validation accuracy and 0.5833 hold-out accuracy. These results indicate that gradient-oriented structure is the dominant discriminative cue among the tested handcrafted descriptor families.

Table 3. Feature-ablation benchmark under provenance-aware evaluation.

Feature	CV Acc.	CV Std.	Test Acc.	CV F1	Test F1	Train (s)	Test (s)
HOG only	0.9833	0.0289	1.0000	0.9791	1.0000	0.6391	0.2969
LBP only	0.6250	0.0333	0.7458	0.6288	0.7248	0.0547	0.0225
GLCM only	0.7000	0.0331	0.8542	0.6898	0.8320	0.0420	0.0226
Intensity moments only	0.4764	0.0638	0.5833	0.4608	0.5599	0.0295	0.0194
HOG + LBP	0.9833	0.0289	1.0000	0.9791	1.0000	0.5620	0.2928
HOG + LBP + GLCM	0.9833	0.0289	1.0000	0.9786	1.0000	0.5618	0.3352
All fused features	0.9833	0.0289	1.0000	0.9786	1.0000	0.8501	0.4617

As shown in Figure 2, the performance gap between HOG-containing and non-HOG configurations remains substantial. The figure indicates that HOG-containing representations produce saturated hold-out performance, whereas grouped cross-validation provides a more conservative indication of internal variability.

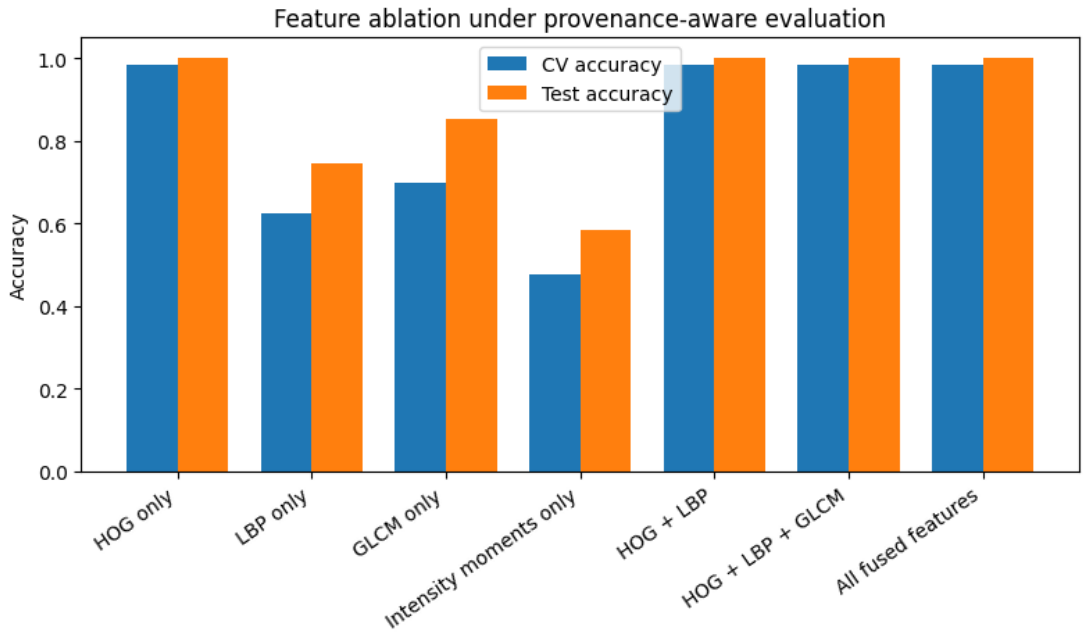

Figure 2. Feature-ablation accuracy under the provenance-aware evaluation protocol.

Configurations containing HOG achieved 1.0000 hold-out accuracy, whereas the LBP-only, GLCM-only, and intensity-moments-only settings remained materially less effective, particularly in grouped cross-validation.

3.2 Classifier benchmarking on fused features

Table 4 and Figure 3 compare classifier behaviour on the fused handcrafted representation. SVM (RBF) [32], KNN (Euclidean) [34], and KNN (cosine) [34] achieved 1.0000 hold-out accuracy, whereas Random Forest [33] reached 0.9958 hold-out accuracy. Cross-validation accuracy ranged from 0.9736 for Random Forest to 0.9833 for the remaining classifiers. These results suggest that, once the fused representation is used, classifier choice has limited influence on hold-out accuracy in the current closed-set benchmark. The comparison remains informative, however, when the computational profile is considered.

Table 4. Classifier benchmark on the fused handcrafted representation.

Classifier	CV Acc.	CV Std.	Test Acc.	CV F1	Test F1	Train (s)	Test (s)
SVM (RBF)	0.9833	0.0289	1.0000	0.9786	1.0000	0.9149	0.3661
Random Forest	0.9736	0.0296	0.9958	0.9702	0.9958	22.6659	0.1097
KNN (Euclidean)	0.9833	0.0289	1.0000	0.9800	1.0000	0.0180	0.0301
KNN (Cosine)	0.9833	0.0289	1.0000	0.9795	1.0000	0.0242	0.0352

Figure 3 illustrates the computational trade-off. In this benchmark, runtime rather than hold-out accuracy becomes the more visible discriminator, with Random Forest requiring the largest training time in the recorded CPU run.

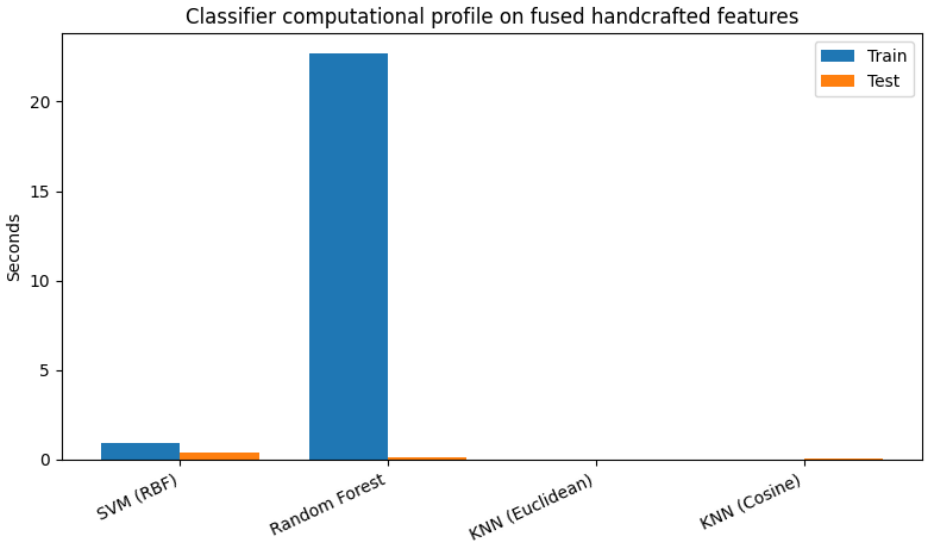

Figure 3. Computational profile of the classifier benchmark on fused handcrafted features.

Random Forest required the largest training cost in the recorded run, while KNN required minimal fitting time and low prediction time in this small closed-set benchmark. These timing values should be interpreted as relative indicators within the recorded CPU environment rather than as portable hardware-independent benchmarks.

3.3 Cosine-similarity baselines

Because cosine similarity is an important comparative baseline in this study, additional cosine-based experiments were evaluated and are summarised in Table 5 and Figure 4. Raw pixels + KNN (cosine), HOG-only + KNN (cosine), and fused handcrafted features + KNN (cosine) each achieved 0.9833 cross-validation accuracy and 1.0000 hold-out accuracy. These findings do not invalidate the descriptor-ablation analysis; rather, they refine its interpretation. HOG remains the strongest individual handcrafted descriptor among the tested feature families, but the supplementary baselines show that the current dataset is sufficiently separable for several simple representations to achieve saturated hold-out performance.

Table 5. Cosine-similarity baselines under provenance-aware evaluation.

Baseline	CV Acc.	CV Std.	Test Acc.	CV F1	Test F1	Train (s)	Test (s)
Raw pixels + KNN (cosine)	0.9833	0.0289	1.0000	0.9787	1.0000	0.0618	0.0927
HOG only + KNN (cosine)	0.9833	0.0289	1.0000	0.9800	1.0000	0.0220	0.0264
Fused features + KNN (cosine)	0.9833	0.0289	1.0000	0.9795	1.0000	0.0298	0.0265

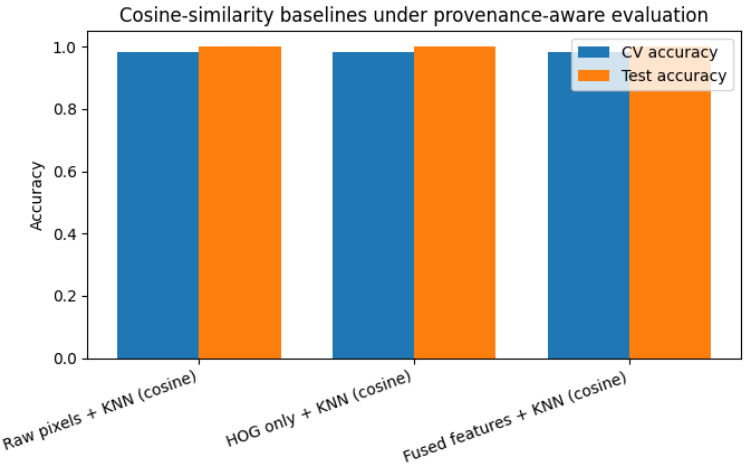

Figure 4. Cosine-similarity baselines under provenance-aware evaluation.

The saturated hold-out performance observed across raw-pixel and handcrafted variants suggests that the present closed-set dataset is highly separable under multiple representational configurations, although grouped cross-validation remains below an absolute ceiling.

3.4 Grouped-split sensitivity analysis

As a robustness check, this study further conducted a four-setting leave-one-provenance-group-out sensitivity analysis to examine whether the saturated performance remained stable across different reserved provenance groups. Each run held out the same source-group position across all 60 classes (G1, G2, G3, or G4), producing 720 training images, 240 hold-out images, 180 training groups, and 60 hold-out groups per setting. As shown in Table 6, both HOG-SVM and fused-SVM retained 1.0000 hold-out accuracy in all four settings, while grouped cross-validation accuracy remained 0.9833. A supplementary 10-repetition grouped hold-out check using seeds 42-51 also produced 1.0000 hold-out accuracy in every repetition, with mean cross-validation accuracy of 0.9600 and a range of 0.9167-0.9833. These sensitivity analyses indicate that the principal benchmark pattern is not dependent on a single arbitrary grouped split.

Table 6. Leave-one-provenance-group-out sensitivity analysis under provenance-aware evaluation.

Held-out group	Pipeline	Train images	Hold-out images	Train groups	Hold-out groups	CV Acc.	CV Std.	Test Acc.	Test F1
G1	HOG-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G1	Fused-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G2	HOG-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G2	Fused-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G3	HOG-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G3	Fused-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000

Held-out group	Pipeline	Train images	Hold-out images	Train groups	Hold-out groups	CV Acc.	CV Std.	Test Acc.	Test F1
G4	HOG-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000
G4	Fused-SVM	720	240	180	60	0.9833	0.0289	1.0000	1.0000

3.5 Visual analysis of class separability

As an additional diagnostic, a supplementary image-level random split using the same HOG-SVM configuration produced 1.0000 accuracy and 1.0000 weighted F1-score. This comparison used a stratified 75:25 image-level split with random seed 42, the same 64 × 64 grayscale input representation, the same HOG settings, and the same in-pipeline standardisation strategy as the main experiment. The key difference was that provenance groups were not enforced; consequently, rotation-derived variants could still be distributed across training and test partitions. The identical hold-out score therefore reinforces the emphasis of this study on evaluation validity rather than on maximising a single aggregate score. To contextualise the saturated scores, Figure 5 presents a sample visual grid of several Batik Nitik classes with closely related geometric structure and subtle local differences.

Figure 5. Sample visual grid of Batik Nitik classes used to illustrate the fine-grained texture variation present in the dataset.

Beyond the visual sample grid, Figures 6 and 7 provide qualitative support for the quantitative results. Figure 6 presents a post hoc two-dimensional PCA projection of the handcrafted feature space. PCA was fitted on the development feature matrix and used solely for visual interpretation; it was not used in training, hyperparameter selection, or model comparison. Figure 7 presents a normalised confusion matrix for the representative best-performing SVM (RBF) model computed on the provenance-aware

hold-out set. On this hold-out set, no motif class showed lower performance than the others; all 60 motif classes achieved per-class accuracy and per-class F1-score of 1.0000. For the saturated SVM hold-out result (240/240 correct predictions), the exact 95% Clopper-Pearson confidence interval for accuracy was 0.9847-1.0000, indicating that the observed ceiling-level result should still be interpreted within the statistical limits of the current closed-set benchmark. Taken together, these results indicate that ceiling-level performance is not driven by a small subset of easier classes.

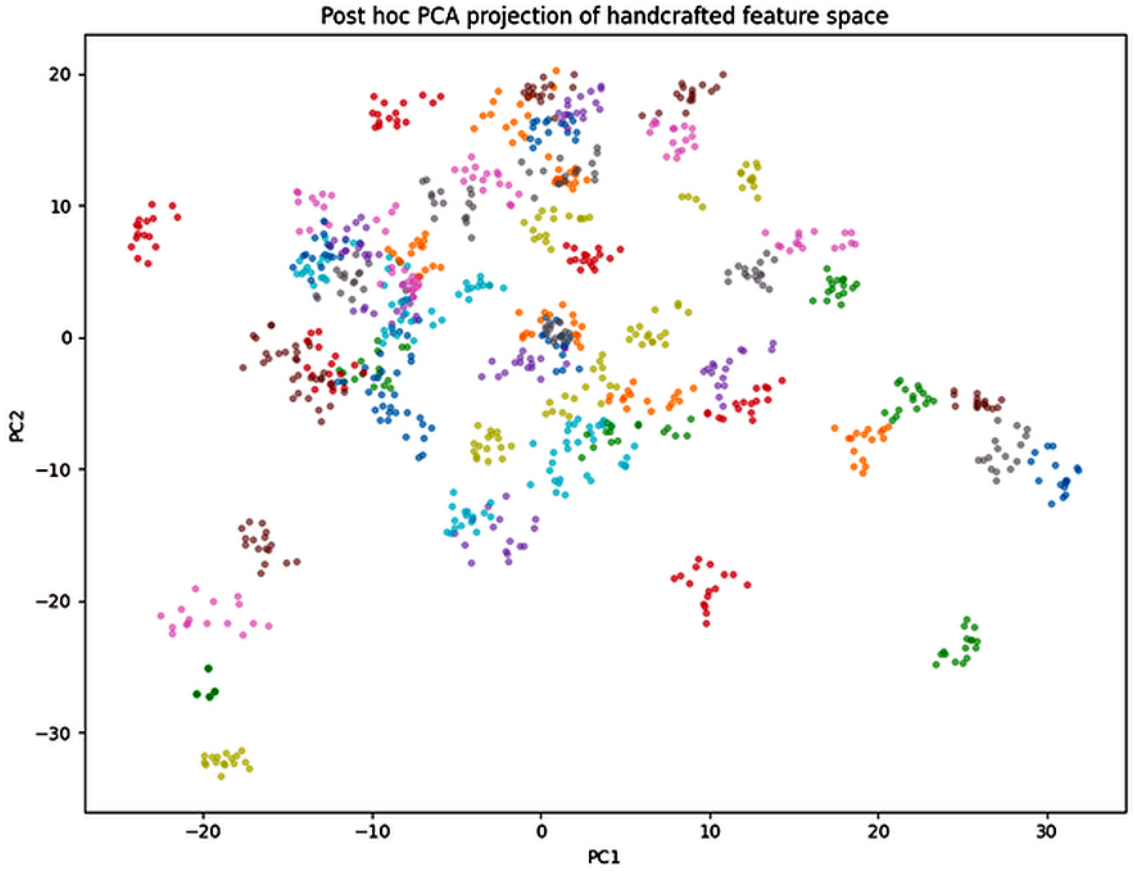

Figure 6. PCA projection of the handcrafted feature space.

The observed clustering structure is consistent with the strong class separability indicated by the quantitative results.

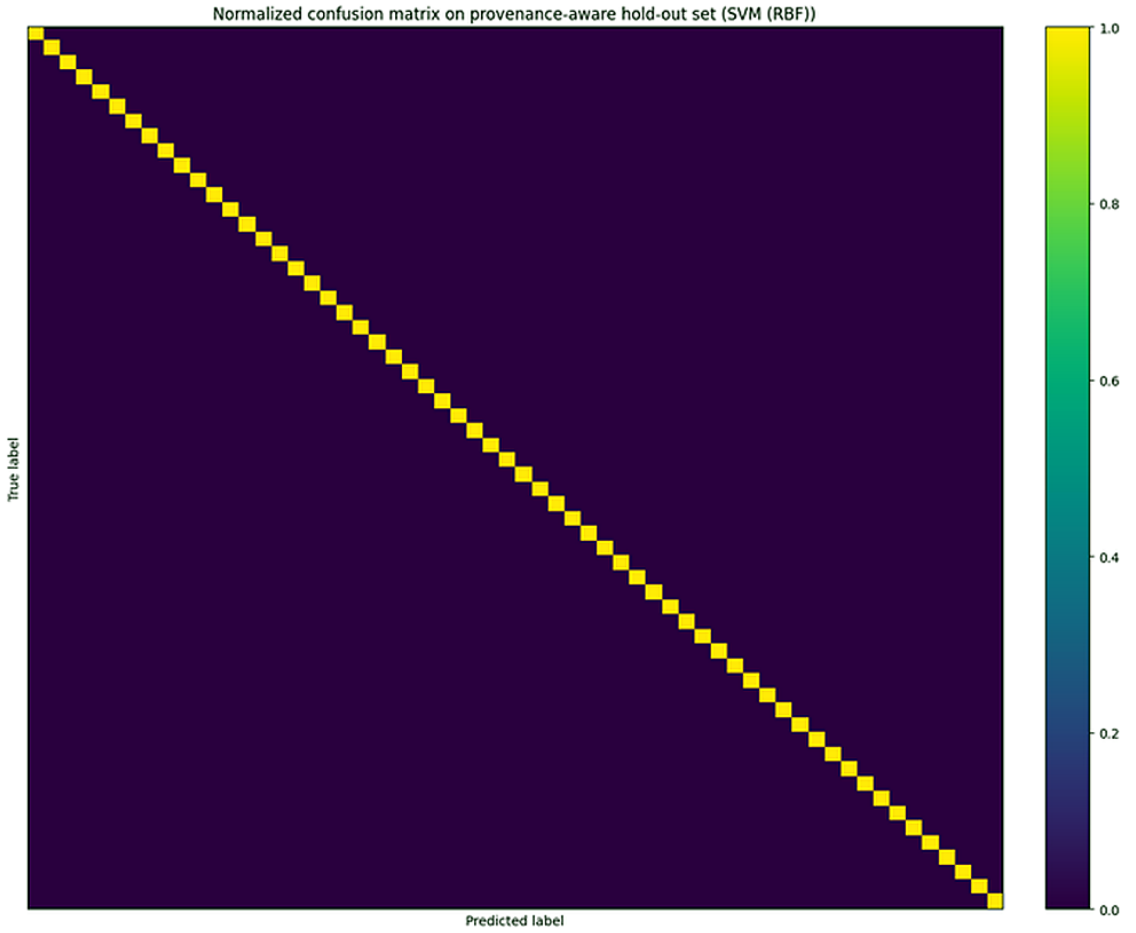

Figure 7. Normalized confusion matrix for a representative best-performing model computed on the hold-out set.

The near-diagonal structure suggests that strong class discrimination is maintained across the full set of motif labels and is consistent with uniformly perfect per-class hold-out results. To consolidate the benchmark logic across descriptor ablation, classifier comparison, cosine-based baselines, grouped-split sensitivity analysis, and qualitative visual analysis, Table 7 summarises the principal evidence and its interpretive implications.

Table 7. Compact summary of experimental conclusions.

Experiment	Evidence	Main conclusion	Interpretive implication
Descriptor ablation	Table 3	HOG is strongest among tested handcrafted descriptors.	Standalone descriptor ranking is meaningful within the tested handcrafted setting.

Experiment	Evidence	Main conclusion	Interpretive implication
Classifier benchmark	Table 4	SVM and KNN variants reach saturated hold-out accuracy; Random Forest is slightly lower.	Representation and dataset separability dominate classifier choice under the current dataset.
Cosine baseline	Table 5	Raw pixels and handcrafted vectors reach saturated hold-out performance.	Closed-set separability extends beyond one descriptor family.
Grouped-split sensitivity	Table 6	All four leave-one-provenance-group-out settings retain 1.0000 hold-out accuracy.	The main benchmark pattern is not dependent on a single held-out source-group position.
PCA visualisation	Figure 6	Low-dimensional projection is descriptive only.	PCA does not affect model selection or reported metrics.
Confusion matrix	Figure 7	Representative hold-out confusion matrix is near-diagonal.	No single class drives the saturated result.

3.6 Discussion

The principal contribution of this study is methodological rather than merely numerical. Although strong results were already observed in preliminary computational workflows, the present analysis demonstrates that the same substantive pattern remains after provenance-aware grouping, in-pipeline preprocessing, four-setting leave-one-provenance-group-out testing, and repeated grouped hold-out checks are imposed. This finding redirects the interpretation from a sole focus on aggregate accuracy toward benchmark validity and the limits of inference from saturated results.

The cosine baselines nevertheless require careful qualification of the study's claims. Because raw-pixel and cosine-based variants also achieve saturated hold-out performance, the current dataset is more appropriately described as highly separable under closed-set conditions than as evidence that a single handcrafted descriptor family

is universally sufficient for batik recognition. The raw-pixel result is particularly informative because it suggests that controlled acquisition, cropping, alignment, or background regularity may contribute materially to the observed separability, in addition to genuine motif structure. A more defensible interpretation is that Batik Nitik 960 is highly separable in its current closed-set form and that the present study clarifies how this separability appears under different representational choices [22], [24], [31].

The ablation analysis remains informative after the cosine baselines are considered. HOG produces the strongest standalone descriptor performance among the tested families, achieving 0.9833 grouped cross-validation accuracy and 1.0000 hold-out accuracy, whereas LBP, GLCM, and intensity moments remain materially lower when used alone. This result indicates that gradient-oriented structure carries the main explanatory burden in the present handcrafted comparison, without implying that HOG is universally superior for batik recognition.

The findings also reposition the present work relative to previous batik studies. Whereas some prior studies emphasise architectural complexity, hybrid preprocessing, or extensive augmentation [1], [4-10], [11-17], the present study shows that a carefully controlled handcrafted pipeline can already produce saturated hold-out performance on the current dataset. The value of this paper therefore lies less in outperforming prior studies than in clarifying what can and cannot be concluded from saturated internal benchmarks.

A further practical implication concerns future dataset construction. If several simple baselines achieve perfect hold-out performance under a provenance-aware protocol, subsequent collections should place greater emphasis on acquisition diversity, cross-device variation, lighting change, scale variation, partial occlusion, wear, background clutter, and harder between-class similarity. For future Batik Nitik datasets, provenance metadata, capture-device metadata, and cross-collection test splits should be treated as core benchmark components rather than optional documentation [22-26].

3.7 Threats to Validity

Internal validity is strengthened by the provenance-aware split and in-pipeline standardisation, but residual regularities may still remain within the dataset beyond those captured by the inferred provenance groups. Because the grouping variable was reconstructed from filenames, the approach depends on the correctness and consistency of the released naming convention. Any undocumented regularity not reflected in the filenames could still influence performance.

Construct validity is limited by the outcome measures used in the study. Accuracy and weighted F1-score are appropriate for the balanced closed-set task examined here, but they do not measure robustness to acquisition shift, calibration quality, retrieval utility, or open-set recognition. Consequently, ceiling-level performance on these metrics should not be conflated with practical readiness.

External validity remains the most important limitation. The dataset contains only four provenance groups per class, all produced within a single collection design [20, 21]. Rotation is a controlled and useful transformation, but it does not substitute for broader real-world variation in capture conditions, motif deterioration, background clutter, or class ambiguity. The evaluation is therefore less susceptible to leakage than an image-level split, yet it remains limited in acquisition diversity.

Finally, interpretive overreach remains a risk. Because multiple settings achieve 1.0000 hold-out performance, the paper should not imply that one classifier or one feature design has been proven optimal in a general sense. The more accurate claim is that the study clarifies descriptor behaviour and benchmark saturation under a stricter internal evaluation protocol.

4. CONCLUSION

This study examined Batik Nitik classification under a stricter and more transparent evaluation protocol. The results confirm that HOG is the strongest descriptor within the handcrafted ablation analysis, while also showing that multiple simple baselines can reach saturated hold-out performance on the Batik Nitik 960 dataset when provenance-aware grouping is enforced. The four-setting leave-one-provenance-group-out analysis and

repeated grouped hold-out check further support the robustness of this internal benchmark pattern, although grouped cross-validation remains more conservative than the hold-out score. The principal contribution is therefore methodological: the study provides a more defensible internal benchmark and a more cautious interpretation of saturated results in cultural-image classification. These findings should not be interpreted as evidence of deployment-ready batik recognition because the benchmark remains closed-set and limited to four provenance groups per class. Future work should move beyond saturated internal benchmarks toward external validation, acquisition-shift experiments, cross-collection testing, and dataset designs with broader visual diversity. In this respect, the present study provides a reference point for understanding how handcrafted features, cosine similarity, and grouped evaluation interact within a controlled batik-recognition benchmark.

ACKNOWLEDGMENT

The authors gratefully acknowledge the academic environment that supported the preparation of this manuscript and the curators and contributors who enabled the Batik Nitik materials to be documented and studied. The authors also thank the editor and reviewers for their constructive suggestions, which improved the methodological clarity and reporting quality of this paper.

REFERENCES

[1] A. A. M. Perdana, M. Fajar B, and A. M. Mappalotteng, "Enhancing batik classification leveraging cnn models and transfer learning," *JOIV Int. J. Informatics Vis.*, vol. 9, no. 3, p. 1033, May 2025, doi: 10.62527/joiv.9.3.2535.

[2] S. Ariessaputra, V. H. Vidiasari, S. M. Al Sasongko, B. Darmawan, and S. Nababan, "Classification of lombok songket and sasambo batik motifs using the convolution neural network (cnn) algorithm," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 1, p. 38, Mar. 2024, doi: 10.62527/joiv.8.1.1386.

[3] A. E. Minarno, I. Soesanti, and H. A. Nugroho, "Batik image representation using multi texton co-occurrence histogram," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3–2, p. 1582, Nov. 2024, doi: 10.62527/joiv.8.3-2.3095.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[4] A. E. Minarno, M. Y. Hasanuddin, and Y. Azhar, "Batik images retrieval using pre-trained model and k-nearest neighbor," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 1, p. 115, Feb. 2023, doi: 10.30630/joiv.7.1.1299.

[5] M. T. D. Putra *et al.*, "Batiknet: batik classification-based management application for inexperienced user," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 4, p. 2411, Dec. 2024, doi: 10.62527/joiv.8.4.3086.

[6] D. W. Pratama, A. Sudiarso, D. S. E. Atmaja, and M. K. Herliansyah, "Multi-architectural transfer learning cnn for klowong batik fabric defect classification," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2123–2138, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4806.

[7] H. Sastypratiwi, H. Muhardi, and Y. Yulianti, "Batik recognition and classification using transfer learning and mobilenet approach," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 4, p. 2400, Dec. 2024, doi: 10.62527/joiv.8.4.2407.

[8] L. Fitriani, D. Tresnawati, and M. B. Sukriyansah, "Image classification on garutan batik using convolutional neural network with data augmentation," *JUITA J. Inform.*, vol. 11, no. 1, p. 107, May 2023, doi: 10.30595/juita.v11i1.16166.

[9] R. F. Alya, M. Wibowo, and P. Paradise, "Classification of batik motif using transfer learning on convolutional neural network (cnn)," *J. Tek. Inform.*, vol. 4, no. 1, pp. 161–170, Feb. 2023, doi: 10.52436/1.jutif.2023.4.1.564.

[10] A. Akmal, R. Munir, and J. Santoso, "Automatic weight of color, texture, and shape features in content-based image retrieval using artificial neural network," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 3, pp. 665–672, Sep. 2023, doi: 10.30630/joiv.7.3.1184.

[11] M. Latief, S. Syahrul, and A. M. Mappalotteng, "Artificial intelligence-based karawo motif formation using genetic algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 3, p. 1820, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1820-1828.

[12] D. A. Anggoro, A. A. T. Marzuki, and W. Supriyanti, "Classification of solo batik patterns using deep learning convolutional neural networks algorithm," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 22, no. 1, p. 232, Aug. 2023, doi: 10.12928/telkomnika.v22i1.24598.

[13] D. G. T. Meranggi, N. Yudistira, and Y. A. Sari, "Batik classification using convolutional neural network with data improvements," *JOIV Int. J. Informatics Vis.*, vol. 6, no. 1, p. 6, Mar. 2022, doi: 10.30630/joiv.6.1.716.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[14] S. Joseph, I. Hipiny, H. Ujir, S. F. Samson Juan, and J.-L. Minoi, "Performance evaluation of sift against common image deformations on iban plaited mat motif images," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 3, p. 1470, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1470-1477.

[15] M. Mao, A. Lee, and M. Hong, "Efficient fabric classification and object detection using yolov10," *Electronics*, vol. 13, no. 19, pp. 1–23, Sep. 2024, doi: 10.3390/electronics13193840.

[16] S. Suprapto, M. N. Tentua, and A. R. Maulana, "Optimizing nitik batik classification through comparative analysis of image augmentation," *IAES Int. J. Artif. Intell.*, vol. 14, no. 5, p. 3970, Oct. 2025, doi: 10.11591/ijai.v14.i5.pp3970-3981.

[17] D. Sinaga, C. Jatmoko, S. Suprayogi, and N. Hedriyanto, "Multi-layer convolutional neural networks for batik image classification," *Sci. J. Informatics*, vol. 11, no. 2, pp. 477–484, May 2024, doi: 10.15294/sji.v11i2.3309.

[18] E. Sugiarto, F. Budiman, and A. Fahmi, "Implementation of deep learning based on convolution neural network for batik pattern recognition," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 10, no. 1, pp. 11–16, Jan. 2025, doi: 10.22219/kinetik.v10i1.2019.

[19] M. F. Dzulqarnain, A. Fadlil, and I. Riadi, "Performance comparison of learned features from autoencoder and shape-based hu moments for batik classification," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1729–1744, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4827.

[20] A. E. Minarno, I. Soesanti, and H. A. Nugroho, "Batik nitik 960 dataset for classification, retrieval, and generator," *Data*, vol. 8, no. 4, p. 63, Mar. 2023, doi: 10.3390/data8040063.

[21] A. E. Minarno, I. Soesanti, and H. A. Nugroho, "Batik classification using microstructure co-occurrence histogram," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 1, p. 134, Mar. 2024, doi: 10.62527/joiv.8.1.2152.

[22] R. Carrilho, E. Yaghoubi, J. Lindo, K. Hambarde, and H. Proença, "Toward automated fabric defect detection: a survey of recent computer vision approaches," *Electronics*, vol. 13, no. 18, p. 3728, Sep. 2024, doi: 10.3390/electronics13183728.

[23] N. Rout, J. Hu, G. Baciu, P. Pattanaik, K. Nakkeeran, and A. Khandual, "Color and texture analysis of textiles using image acquisition and spectral analysis in calibrated sphere imaging system-ii," *Electronics*, vol. 12, no. 9, p. 2135, May 2023, doi: 10.3390/electronics12092135.

[24] R. Machado, L. A. M. Barros, V. Vieira, F. D. da Silva, H. Costa, and V. Carvalho, "Textile defect detection using artificial intelligence and computer vision—a preliminary deep learning approach," *Electronics*, vol. 14, no. 18, p. 3692, Sep. 2025, doi: 10.3390/electronics14183692.

[25] M. Fan, N. Deng, B. Xin, and R. Zhu, "Recognition and analysis of fabric texture by double-sided fusion of transmission and reflection images under compound light source," *J. Text. Inst.*, vol. 114, no. 11, pp. 1634–1646, Nov. 2023, doi: 10.1080/00405000.2022.2145428.

[26] Y.-F. Tu, M.-Y. Kwan, and K.-L. Yick, "A systematic review of ai-driven prediction of fabric properties and handfeel," *Materials (Basel)*, vol. 17, no. 20, p. 5009, Oct. 2024, doi: 10.3390/ma17205009.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 24, no. 7, pp. 971–987, 2002, doi: 10.1109/TPAMI.2002.1017623.

[29] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[30] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[31] K. John, D. D. Saurette, and B. Heung, "The problematic case of data leakage: a case for leave-profile-out cross-validation in 3-dimensional digital soil mapping," *Geoderma*, vol. 455, no. March, p. 117223, Mar. 2025, doi: 10.1016/j.geoderma.2025.117223.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[33] L. Breiman, "Random forests," *Mach. Learn*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[34] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.