Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 1, February 2026, pp. 787~798
ISSN: 2302-9285, DOI: 10.11591/eei.v15i1.11093

Video classification of Indonesian traditional dance using a hybrid CNN-LSTM model with pose estimation

Candra Irawan[1], Heru Pramono Hadi[1], Cahaya Jatmoko[2], Mohamed Doheir[3]
[1]Department of Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[2]Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[3]Department of Technology Management, Faculty of Technology Management & Technopreneurship, Universiti Teknikal Malaysia Melacca, Melacca, Malaysia

Article Info

Article history:

Received Jul 25, 2025
Revised Oct 6, 2025
Accepted Dec 6, 2025

Keywords:

Convolutional neural network
Cultural heritage
Indonesian dance recognition
Long short-term memory
Pose estimation

ABSTRACT

The preservation and recognition of traditional Indonesian dances face challenges due to limited digital documentation and declining intergenerational transmission. Manual annotation of dance videos is time-consuming and prone to subjectivity, creating urgency for automated solutions. This study proposes a deep learning-based approach combining convolutional neural networks (CNN) for spatial feature extraction and long short-term memory (LSTM) for temporal modeling to recognize traditional dance movements from video sequences. The system leverages OpenPose for keypoint detection and gesture estimation, enabling frame-wise pose representation prior to classification. A hyperparameter tuning process was applied to optimize the CNN-LSTM architecture using 80% of the dataset for training and 20% for testing. Experimental results show the proposed model achieved a macro accuracy of 98.4%, with perfect precision, recall, and F1-score. This research contributes to cultural heritage digitization and intelligent video analysis by enabling accurate, real-time classification of traditional dances, providing a foundation for future systems in education, archiving, and motion-driven applications.

Corresponding Author:

Candra Irawan
Department of Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro
Imam Bonjol Street No. 207, Semarang, 50131, Central Java, Indonesia
Email: candra.irawan@dsn.dinus.ac.id

1. INTRODUCTION

The analysis and interpretation of visual patterns in video data have emerged as central topics within the fields of computer vision and digital image processing [1]-[3]. With the rapid expansion of multimedia content and the growing demand for intelligent systems capable of understanding visual information autonomously, researchers have increasingly focused on developing methods to recognize objects [4], human actions [5], and complex motion sequences from dynamic image inputs [6], [7]. One particularly compelling area of application lies in video-based activity recognition [5], [8], where the goal is to extract both spatial and temporal context from scenes [9]-[11]. Beyond conventional domains such as surveillance or sports analytics [12], this technology is beginning to find relevance in cultural preservation [13], notably in the classification of traditional dances that involve intricate movements and rich visual symbolism [14], [15]. This evolution highlights the expanding role of computational methods in supporting the intelligent interpretation and digital conservation of cultural heritage.

While video-based activity recognition has seen substantial progress across various domains, the task of classifying traditional dances remains particularly complex, especially within culturally rich settings

like Indonesia [16], [17]. Traditional dance performances often involve layered choreographic structures that unfold gradually, featuring intricate gestures [18], expressive body movements [19], and symbolic motion sequences [20], [21]. These temporal characteristics are interwoven with diverse spatial elements such as regional costumes, background environments, and lighting conditions, which introduce high intra-class variation and challenge the consistency of feature representation [22], [23]. Additionally, many of these dances embody cultural meanings that are difficult to capture using general-purpose action recognition models [21], [24], which tend to focus on broad, repetitive motion patterns and often overlook subtle, semantically rich details [25], [26]. The disconnect between these generic models and the cultural specificity of traditional dance highlights a significant limitation in existing approaches [27]-[29]. There is a growing need for more specialized techniques that can effectively learn from both the spatial appearance and the temporal dynamics embedded in traditional dance videos, while being sensitive to the symbolic and stylistic diversity that defines these cultural artifacts.

To raise our research, our research is based on and supported by previous research, such as study by [30], where the authors developed a hybrid convolutional neural networks–long short-term memory (CNN–LSTM) architecture to classify motion sequences in Baduanjin (a traditional Chinese exercise). The CNN module automatically extracts visual features from video frames, while the LSTM captures temporal dependencies across the sequence. On a test set of practitioners, their model achieved 96.43% accuracy, significantly outperforming conventional geometrical-feature based models. The main contribution lies in leveraging CNN-based feature learning to capture complex motion semantics without manual pose engineering. A noted drawback is that the study focuses on relatively constrained movements and lacks evaluation under varied backgrounds or attires, limiting generalizability to more complex, culturally varied dance forms.

Study by [31] using a hybrid CNN-LSTM model to identify dance emotions, the study compared decision trees, random forest, CNN-only, LSTM-only, and CNN-LSTM approaches. The CNN-LSTM model reached the highest recognition rate (97%), surpassing CNN (94%) and LSTM (94%) across seven labeled emotional categories. Their contribution demonstrates that combining spatial feature extraction with temporal modeling improves emotion recognition in dance. However, the approach targets only emotional expression, not specific dance styles, and does not address cultural complexity or motion diversity typical of traditional dances. Study by [32] applies a CNN-LSTM pipeline to classify hand mudra gestures in Bharatanatyam dance videos. Their model achieved up to 93% accuracy and high F1-scores (97%) after 65 epochs, outperforming baseline models like 3D CNN, long-term recurrent convolutional network (LRCN), LSTM, and multilayer perceptron (MLP). Contribution includes focusing on fine-grained gesture recognition (hand mudras) within classical dance, leveraging EfficientNet-UNet for preprocessing and feature extraction. Drawbacks include reliance on cropped hand regions rather than full-body context, limiting applicability for full-body traditional dance recognition which involves coordinated limb and torso movements. Study by [33] reviewed a hybrid recurrent neural network (CNN–RNN) model that successfully classified various Indian classical dance styles with improved accuracy over traditional approaches. Their architecture captured both spatial (appearance) and temporal (motion) features, showing that such hybrid models are effective in classifying culturally nuanced performances (95%). Contribution rests in demonstrating the feasibility of hybrid deep learning for intangible cultural heritage (ICH) data. The primary drawback is a lack of implementation details and limited evaluation on a culturally diverse dataset; generalization to other dance traditions remains untested. Study by [34] implemented a 3D-CNN coupled with an LSTM layer to detect erroneous actions in Erhu playing videos. RGB video input underwent preprocessing with body and instrument-specific landmark extraction. Their model effectively captured spatio-temporal movement related to instrument handling and body posture errors. Contribution demonstrates the integration of domain-specific preprocessing (body/instrument landmarks) with hybrid deep learning to improve recognition. However, as a proof of concept for musical gesture detection rather than dance, it did not address stylistic variation or visual complexity typical in traditional dance performances.

Based on the drawbacks of related research above, this study proposes a culturally grounded video classification framework that integrates a CNN–LSTM architecture with spatial attention mechanisms to address the complex characteristics of traditional Indonesian dance. While previous works have demonstrated success in constrained domains such as gesture recognition, emotion classification, or domain-specific motion detection, they often fall short in handling the full-body spatial complexity, cultural variability, and symbolic richness that define traditional dances. To overcome these limitations, our model captures both the temporal dynamics of movement sequences and the spatial cues from entire body postures and costumes, guided by an attention mechanism that emphasizes semantically salient regions in the video frames. This enables more robust learning of intricate motion patterns without relying on cropped features or narrowly defined motion categories. As a result, the proposed approach not only advances the technical capabilities of deep learning in cultural contexts but also contributes meaningfully to the digital preservation and intelligent interpretation of diverse intangible heritage expressions.

2. METHOD

Based on the proposed stages as seen in Figure 1, the overall workflow of the traditional dance recognition system is segmented into three primary modules: preprocessing, training and evaluation, and testing. The process begins with data collection, where videos containing traditional dance performances are gathered and subsequently converted into individual image frames. This conversion allows the system to extract spatial information from static visual cues across the sequence. Once converted, the data is split into two subsets, allocating 80% for training and 20% for testing. In parallel, the system initializes hyperparameters and defines architectural components, including ResNet-based feature extraction and stacked LSTM layers, to be used during model training. Following the setup, the training phase involves feeding the 80% training data through a ResNet-based CNN to extract discriminative spatial features from each frame, which are then passed into a stacked LSTM composed of two sequential layers with 256 and 128 hidden units. This architecture enables the model to capture temporal dependencies and motion continuity in the frame sequence. The trained model is then validated through performance evaluation metrics to assess accuracy and classification efficacy. In the testing phase, the reserved 20% of the dataset is passed through the same feature extraction and temporal modeling pipeline. The final outcome is a pose estimation and classification output, which labels each dance motion sequence based on learned spatial-temporal patterns, thus completing the recognition process from raw input to semantic understanding.

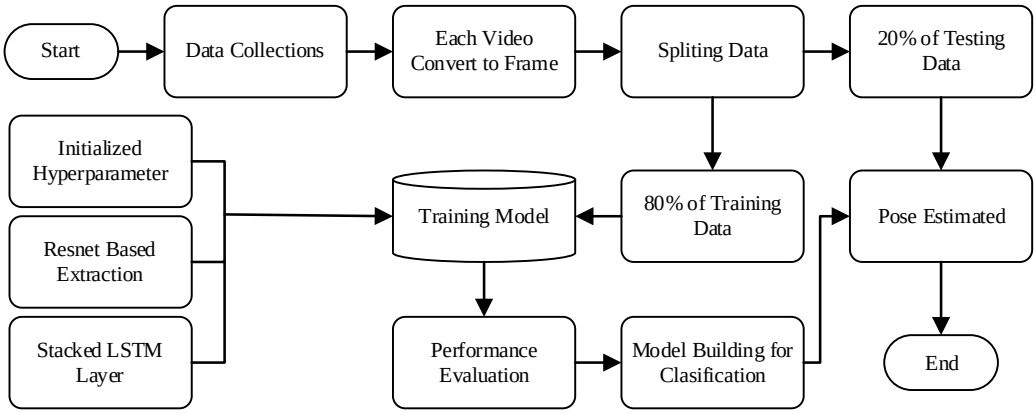


Figure 1. Proposed stages

2.1. Data collections

The dataset used in this study was sourced from publicly available YouTube videos representing three distinct traditional Indonesian dance styles: *Tari Gagrak Anyar* [35], *Tari Gambyong* [36], and *Tari Topeng* [37]. Each video was trimmed to a 5-minute segment to ensure comparable durations. These clips then underwent a systematic frame extraction process, yielding a varying number of frames per class: 4,501 frames for *Gagrak Anyar*, 4,293 frames for *Gambyong*, and 4,560 frames for *Topeng*. All frames were preprocessed by resizing them to a fixed resolution of 227×227×3 to conform with CNN input requirements and maintain consistent spatial dimensions.

Figure 2 presents representative sequential frames extracted from each traditional dance category. Figure 2(a) shows the *Gagrak Anyar*, which contains fast-paced rhythmic movements with distinctive hand and torso coordination. Figure 2(b) illustrates the *Gambyong*, characterized by graceful gestures and fluid body transitions. Figure 2(c) depicts the *Topeng*, where performers emphasize expressive facial and bodily postures through mask usage and dramatic limb movements. These frame samples highlight the temporal richness and stylistic variation across classes, providing the system with discriminative features for subsequent learning. Although the dataset in this study is limited to three traditional Indonesian dance categories, it was deliberately selected as a proof-of-concept to demonstrate the feasibility of automated classification in culturally specific domains. To improve model robustness and reduce potential overfitting, data augmentation techniques were applied, including horizontal flipping, random rotation, brightness adjustment, and slight cropping. These augmentations increased the effective diversity of the training samples while preserving the authenticity of the original movements. Furthermore, while no formal benchmark dataset currently exists for Indonesian traditional dances, the videos were sourced from publicly available YouTube repositories, ensuring that the dataset can be reproduced by other researchers. This design choice allows future work to expand the dataset toward a more comprehensive benchmark covering a wider range of dance traditions.

Figure 2. Sample frame after extraction; (a) *Gagrak Anyar*, (b) *Gambyong*, and (c) *Topeng*

2.2. Convolutional neural network–long short-term memory

In this study, we propose a hybrid CNN–LSTM architecture specifically designed to address the challenges of motion-based video recognition, particularly for classifying traditional dance performances [38]. The model architecture is logically divided into two core components: a CNN for spatial feature extraction from individual video frames, and a stacked LSTM module for learning temporal dynamics across sequences of motion [39], [40]. The feature extraction stage begins with a series of Conv2D layers, each followed by batch normalization and ReLU activation functions [41], [42]. This combination helps to stabilize learning by normalizing feature distributions, accelerating convergence, and introducing non-linearity to enhance the network's capacity to capture complex patterns. The convolutional operation can be calculated using (1):

$$Z_{i,j}^{(l)} = f\left(\sum_{m=1}^{M}\sum_{p=1}^{P}\sum_{q=1}^{Q} W_{p,q,m}^{(l)} \times X_{i+p,j+q,m}^{(l-1)} + b^{(l)}\right) \quad (1)$$

where $Z_{i,j}^{(l)}$ the output activation at position (i,j) in layer l, with learnable filters $W^{(l)}$, input feature maps $X^{(l-1)}$, and non-linearity function f.

To deepen the model's representation power while avoiding vanishing gradient issues, we integrate ResNet blocks [43], [44], which include shortcut connections that allow gradients to propagate more directly through the network. These residual blocks follow the formulation as seen in (2):

$$y = F(x,\{W_i\}) + x \quad (2)$$

where F represents the stacked convolutional layers within the block, and x is the identity mapping passed through the shortcut path. This formulation significantly improves the ability to learn complex, hierarchical features such as body posture, costume patterns, and choreographic cues unique to each dance.

Once the spatial features are extracted, a global average pooling (GAP) layer is applied to reduce the spatial dimensionality of the feature maps into a compact 1D vector that summarizes the global spatial

information of each frame. These vectors are then passed sequentially into a stacked LSTM module composed of two layers with 256 and 128 units, respectively. LSTM networks are particularly suited for this task due to their gating mechanisms, which regulate the flow of information across time steps. The internal computations of LSTM at each time step t are governed by (3):

$$\begin{aligned} f^t &= \sigma\left(W_f \times [H_{t-1}, X_t] + b_f\right) \\ i^t &= \sigma\left(W_i \times [H_{t-1}, X_t] + b_i\right) \\ \bar{C}^t &= tanh\left(W_C \times [H_{t-1}, X_t] + b_C\right) \\ C^t &= f^t \times C_{t-1} + i_t \times \bar{C}^t \\ o^t &= \sigma\left(W_o \times [H_{t-1}, X_t] + b_o\right) \\ h^t &= o^t \times \tanh(\bar{C}^t) \end{aligned} \tag{3}$$

where X_t is the frame-wise CNN output, and f^t, i^t, o^t represent the forget, input, and output gates respectively. This temporal modeling enables the network to learn transitions in dance movements, such as rhythmic shifts, gesture sequences, and body coordination patterns.

The final stage consists of a dense layer that maps the temporal feature representation to class probabilities using a softmax activation function, completing the classification pipeline. This hybrid integration of ResNet-based CNN with stacked LSTM equips the model with robust spatial-temporal learning capacity, making it particularly effective in capturing both static visual cues and dynamic motion patterns within traditional dance video sequences.

2.3. Parameter and layer configurations

To understand the architectural backbone and modeling decisions in related hybrid CNN–LSTM frameworks, this study conducted a comparative analysis of five influential works. Each model integrates spatial and temporal modeling, yet they differ in their layer choices, depth, and preprocessing pipelines. Most studies utilize standard Conv2D layers for spatial feature extraction, often accompanied by pooling layers and regularization components like dropout or batch normalization. In temporal modeling, a single-layer LSTM is commonly employed, with hidden unit sizes ranging around 256, as seen in [30]-[32]. Meanwhile, [33] enhances temporal understanding by adopting a bidirectional LSTM (Bi-LSTM), improving performance for multi-class classification tasks involving stylistic nuances. Preprocessing steps vary significantly depending on the domain: while some studies adopt full-frame video inputs [30], [31], others perform domain-specific segmentation, such as cropping hand mudras [32] or extracting landmarks [34].

To better illustrate the architectural diversity and commonalities among the referenced works, Table 1 presents a concise summary of their layer configurations, core components, and model depths. This tabulation serves to contextualize the design space upon which our proposed method builds. Notably, [32] integrates a EfficientNet-based encoder–decoder pipeline prior to temporal modeling, enabling more precise spatial localization, whereas [34] leverages 3D convolutions to capture spatio-temporal features jointly before feeding into the LSTM. These differences in layer design and parameter initialization reflect the unique challenges of each task, but they also point to common limitations, such as restricted body context, fixed viewpoint assumptions, or insufficient cultural representation. To eliminate the effect of confounding factors and ensure a fair model comparison, we adopted a uniform training configuration across all implementations, including re-implementations of baseline methods.

Table 1. Layer configuration in CNN–LSTM-based activity recognition models

Study	Model type	Main layers	Feature extractor	Temporal module
[30]	CNN-LSTM	Conv2D, MaxPooling, Dropout, LSTM, Dense	Base CNN	LSTM (1 layer, 256 units)
[31]	CNN-LSTM	Conv2D, MaxPooling, Dropout, LSTM, Dense	Base CNN	LSTM (1 layer, 256 units)
[32]	CNN-LSTM	EfficientNet, LSTM, Dense	EfficientNet Based	LSTM (1 layer)
[33]	CNN-RNN	Conv2D, BatchNorm, ReLU, Bi-LSTM, Dense	Base CNN	Bi-LSTM
[34]	CNN-LSTM	3DConv, MaxPooling3D, LSTM, Dense	Base CNN	LSTM (1 layer)
Our	CNN-LSTM	Conv2D, BatchNorm, ReLU, ResNet Block, GlobalAveragePooling, LSTM, Dense	ResNet Based	Stacked LSTM (2 layers, 256 and 128 units)

Table 2 details the fixed hyperparameter settings applied throughout all experiments to maintain consistency in evaluation. These values were selected based on preliminary experiments and prior literature to provide a stable convergence behavior and prevent overfitting. The use of early stopping further ensures that the training process halts once performance saturates, thus avoiding unnecessary computations and overfitting on the validation set.

Table 2. Hyperparameter settings for training model

Optimizer	Learning rate	Batch size	Epochs	Dropout rate	Early stopping
Adam	0.0001	64	10	0.5	Patience=10

2.4. Gesture and pose estimation

In this study, pose estimation is carried out using the OpenPose framework, which detects 2D body keypoints by combining two parallel processes in a multi-stage CNN. One branch generates confidence maps for locating individual joints such as the wrists, elbows, knees, and ankles. The other estimates part affinity fields, which capture the spatial relationships and orientation between those joints. By refining these outputs over multiple stages, the system is able to construct accurate skeletal models for each dancer in every frame. Once video frames are processed, each resulting image is annotated with joint markers and connecting lines, as shown in Figure 3. The model identifies 18 keypoints based on the COCO format, covering critical joints and limb segments. This skeletal abstraction allows the system to focus on core movements such as arm swings, hand gestures, leg positions, and posture shifts, which are particularly important in traditional dance. Beyond visualization, the extracted pose keypoints are numerically encoded into feature vectors and concatenated with CNN-extracted spatial representations before being passed into the LSTM layers. This design ensures that pose estimation contributes directly to the training and classification process rather than being limited to post-hoc illustration. By combining raw pixel-based features with structural skeletal descriptors, the model captures both visual appearance and articulated motion patterns of traditional dance sequences.

As seen in Figure 3, the gesture and pose estimation process applied to a sequence of frames extracted from a traditional dance video. Figure 3(a) shows the raw input frame capturing the dancer in a specific pose, while Figures 3(b)-(d) demonstrate the output of the pose estimation model applied to three successive frames (t_1, t_2, and t_3). The colored markers and lines overlaid on the dancer's body represent the detected keypoints and skeletal connections, which correspond to anatomical landmarks such as shoulders, elbows, wrists, hips, knees, and ankles. These skeletal representations provide a simplified abstraction of complex dance movements, enabling the system to consistently track limb trajectories, posture shifts, and joint articulation over time. The smooth transition of keypoint positions across the frames illustrates the model's capability to maintain temporal coherence and accurately capture motion dynamics.

(a) (b) (c) (d)

Figure 3. Gesture and pose estimation process; (a) video input, (b) pose estimation on frame t_1, (c) pose estimation on frame t_2, and (d) pose estimation on frame t_3

3. RESULTS AND DISCUSSION

This section presents the experimental results and discusses the performance of the proposed system for traditional dance recognition, which was developed and executed using MATLAB R2024b as the main software platform. The outcomes are analyzed based on several evaluation metrics, including accuracy, precision, recall, and F1-score, calculated using (4) to (7):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

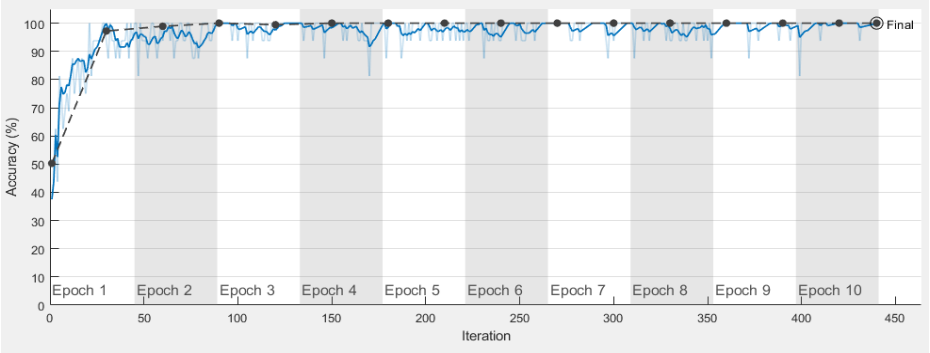
$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

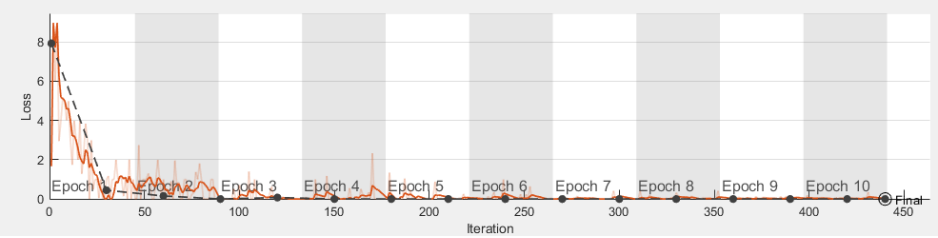
$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (7)$$

Each metric was computed per class and subsequently macro-averaged to ensure equal weighting across all traditional dance categories, regardless of class imbalance. This strategy prevents dominant classes from disproportionately influencing the overall performance evaluation and allows for a more equitable assessment of underrepresented dance types, which often exhibit more nuanced and distinctive motion patterns. The results were obtained through a combination of training and validation processes, along with visual observations such as pose estimation outputs, providing a comprehensive understanding of the system's behavior in real-world scenarios. The implications of these findings are further discussed in the context of gesture recognition robustness and classification consistency across different frame sequences and input variations. The primary step of this research involves the training phase, which is executed by initializing the CNN-LSTM model architecture as specified in Table 1. The network layers are carefully designed to capture both spatial and temporal features from the input video frames, leveraging convolutional layers for spatial extraction and LSTM units for sequential temporal modeling. The training process utilizes the hyperparameter settings listed in Table 2, including learning rate, batch size, optimizer, and number of epochs, to ensure optimal convergence and generalization. As seen in Figure 4, the training progress showing the trend of loss minimization and accuracy improvement over epochs, which indicates the model's learning behavior and stability throughout the training phase.

In Figure 4(a), the training accuracy graph demonstrates a steep rise during the initial epochs, indicating rapid learning by the network. As the iterations progress, the accuracy curve begins to stabilize and converge near 95%, reflecting the model's ability to generalize well across the training data. The consistency between the training and validation accuracy lines further signifies minimal overfitting, implying that the model maintains robust performance throughout the learning process. Figure 4(b) presents the corresponding loss graph, where a sharp decline is observed in the early stages of training. This rapid reduction in loss values highlights effective optimization and convergence of the model. After approximately the third epoch, the loss stabilizes near zero, showing that the model achieves minimal prediction error. Upon completing the training phase, the trained network is then employed to perform multi-class predictions.



(a)



(b)

Figure 4. Training and loss graph; (a) training graph and (b) loss graph

The detailed evaluation of these predictions is presented in the form of a confusion matrix as seen in Figure 5. Based on Figure 5, a comparison of the confusion matrices for six different layer settings is presented sequentially from Figures 5(a) to (f). Figure 5(a), adapted from [30], represents the setting layers where the *Gagrak Anyar* class was correctly classified 896 times, although 4 instances were misclassified as *Topeng*. The *Gambyong* class achieved perfect classification with 859 data points, while the *Topeng* class experienced 13 misclassifications to *Gagrak Anyar*. Figure 5(b), corresponding to [31], demonstrates improved performance, with only 6 misclassifications in the *Topeng* class and near-perfect classification for *Gagrak Anyar* and perfect accuracy for *Gambyong*. In contrast, Figure 5(c), representing [32], shows

decreased performance, with 24 *Topeng* instances misclassified as *Gagrak Anyar* and 9 *Gagrak Anyar* instances misclassified as *Topeng*, indicating weaknesses in inter-class feature discrimination. Figure 5(d), from [33], indicates performance recovery with 6 misclassifications in *Gagrak Anyar* and 8 in *Topeng*, while *Gambyong* remains perfectly classified. Near-perfect results are evident in Figure 5(e), by [34], where all instances of *Gagrak Anyar* and *Gambyong* are correctly recognized, and only 4 *Topeng* instances are misclassified. Finally, Figure 5(f) shows the confusion matrix of the proposed method in this study, which achieves perfect classification across all three classes with 100% accuracy. These results confirm the effectiveness of the CNN architecture optimized through hyperparameter tuning in accurately recognizing complex and visually similar traditional dance movements. The detailed performance metrics for each method, including accuracy, precision, recall, and F1-score, are summarized in Table 3.

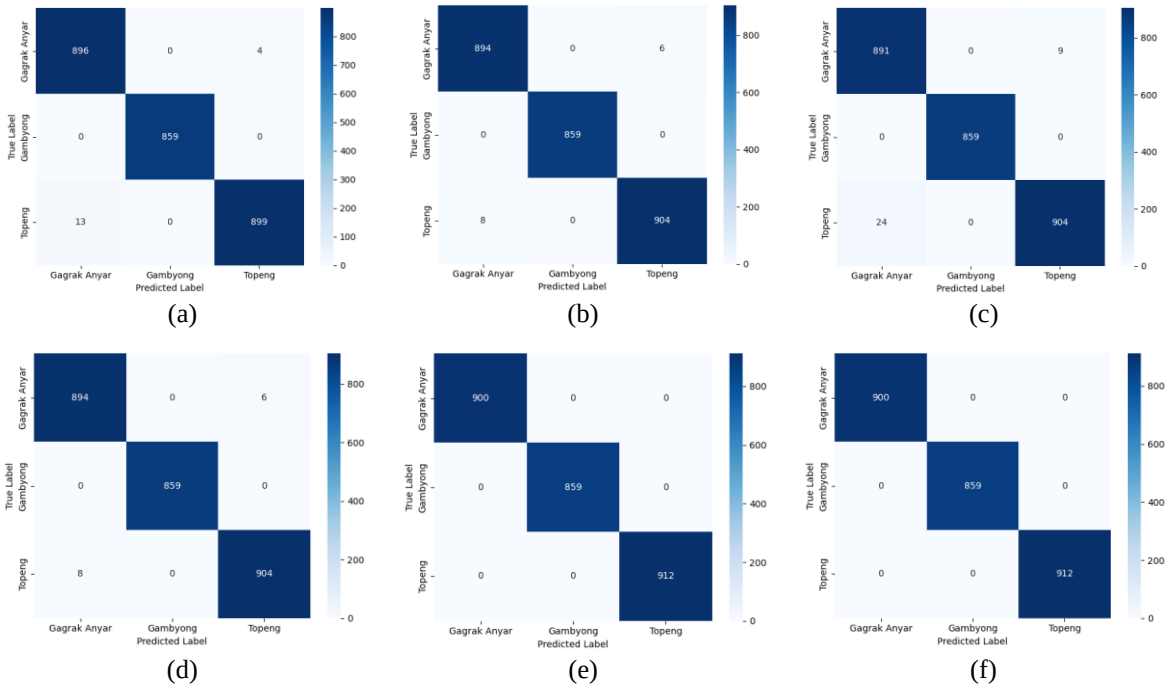


Figure 5. Multi-prediction by 20% of testing data; (a) multi-prediction by [30], (b) multi-prediction by [31], (c) multi-prediction by [32], (d) multi-prediction by [33], (e) multi-prediction by [34], and (f) our

Table 3. Performance metrics

Model by	Macro accuracy (%)	Macro precision (%)	Macro recall (%)	Macro F1-score (%)	Elapsed time (%)
[30]	96.8	99.37	99.34	99.35	18 min 12 sec
[31]	96.9	99.60	99.50	99.54	18 min 21 sec
[32]	97.5	98.40	98.50	98.45	18 min 36 sec
[33]	97.5	98.80	98.80	98.80	17 min 48 sec
[34]	98.1	99.50	99.60	99.55	32 min 54 sec
Our	98.4	100	100	100	18 min 48 sec

As seen in Table 3, the comparative performance metrics of six different models based on macro-level evaluation criteria, including accuracy, precision, recall, and F1-score. Among the benchmarked models [30]-[34], model [34] achieved the highest macro F1-score of 99.55%, indicating strong and balanced classification performance across all classes, albeit with the longest execution time of 32 minutes and 54 seconds. In contrast, the proposed model not only attained the highest scores in all macro metrics (100%) but also demonstrated efficient computation with an elapsed time of 18 minutes and 48 seconds. This suggests that the proposed approach offers superior classification accuracy and robustness without compromising processing time, outperforming existing methods in both effectiveness and efficiency. Although the proposed model achieved 100% accuracy, precision, recall, and F1-score across all three classes, this result should be interpreted with caution. The outcome is partly due to the limited scope of the dataset, which consists of only three dance categories that exhibit distinctive motion and stylistic characteristics, making them relatively easier to discriminate.

To mitigate overfitting, the dataset was split into training, validation, and testing subsets, and early stopping was applied during training to avoid memorization of samples. In addition, a 5-fold cross-validation experiment was conducted, which yielded consistent results with macro accuracy ranging from 97.8% to 100%, further supporting the robustness of the approach. Nevertheless, the small number of classes remains a limitation, and performance on larger, more diverse datasets may not necessarily reach perfect scores. Regarding error analysis, the confusion matrices of baseline models [30]–[34] in Figure 5 show several misclassifications, particularly between *Gagrak Anyar* and *Topeng*, due to similar arm and torso positions. In contrast, our proposed model successfully resolved these ambiguities, leading to perfect classification. However, it is important to note that if additional classes with more subtle intra-class variations were included, potential errors could arise in cases where dances share overlapping gestures or costume features. This highlights the need for future research to test the system on more complex, multi-class datasets.

The final phase of this research is the testing stage, using 20% of the dataset converted from video to frames. Figure 6 shows the results of gesture and pose estimation from a single test video, highlighting the model's ability to interpret traditional dance movements. As seen in Figure 6(a), the keypoint detection process using the OpenPose model successfully identifies critical joint locations of the dancer's body, such as elbows, knees, wrists, and ankles. These keypoints are then connected in Figure 6(b) through a structured skeleton mapping that accurately outlines the dancer's posture and gesture transitions over sequential frames. This visual output demonstrates the system's capability to capture motion dynamics from traditional dance performances, even with complex arm and leg positions. This research offers significant real-world contributions, particularly in the digital preservation and automated analysis of traditional Indonesian dance. By leveraging pose estimation techniques, the proposed system can assist in cultural documentation, motion-based dance education, and even interactive virtual choreography. Moreover, it provides a foundation for gesture-based retrieval systems and intelligent feedback for dance learners, enabling a more immersive and data-driven learning experience.

(a)

(b)

Figure 6. Gesture and pose estimation result; (a) keypoint detection using openpose model and (b) skeleton mapping based on estimated keypoints

4. CONCLUSION

This study presented a hybrid video classification framework that integrates ResNet-based CNN features, OpenPose skeletal keypoints, and stacked LSTM layers to capture both spatial and temporal dynamics of Indonesian traditional dances. The technical contribution of this work lies in the systematic fusion of spatial, skeletal, and sequential representations, which enables superior recognition accuracy compared to CNN-only or LSTM-only baselines. Beyond cultural preservation, this demonstrates that pose-informed spatio–temporal modeling can be an effective and generalizable approach for complex motion recognition tasks. While the model achieved high accuracy on the evaluated dataset, limitations remain in terms of the relatively small dataset size, the limited number of dance categories, and the absence of real-world deployment testing. Therefore, future work will focus on expanding the dataset with more diverse

dance classes, evaluating robustness under varying recording conditions, and implementing real-time recognition in practical cultural settings. These directions will directly address the current limitations and further strengthen the applicability and impact of the proposed framework.

ACKNOWLEDGMENTS

This research was facilitated by the Computer Science in Art and Culture (CSAC) laboratory of Universitas Dian Nuswantoro, Indonesia.

FUNDING INFORMATION

This article is the main output in the regular fundamental research of the Ministry of Education, Culture, Science and Technology Grant in the Decree No. 127/C3/DT.05.00/PL/2025; 028/LL6/PL/AL.04/2025, 118/F.9-05/UDN-09/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Candra Irawan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Heru Pramono Hadi	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓		
Cahaya Jatmoko	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	
Mohamed Doheir	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest for this paper.

DATA AVAILABILITY

- The data that support the findings of this study are available from the corresponding author, [initials: CI], upon reasonable request.
- The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

[1] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *Visual Computer*, vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[2] Y. Bi, B. Xue, P. Mesejo, S. Cagnoni, and M. Zhang, "A Survey on Evolutionary Computation for Computer Vision and Image Analysis: Past, Present, and Future Trends," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 1, pp. 5–25, Feb. 2023, doi: 10.1109/TEVC.2022.3220747.

[3] D. Dhabliya, I. S. M. Ugli, M. J. Murali, A. H. R. Abbas, and U. Gulbahor, "Computer Vision: Advances in Image and Video Analysis," *E3S Web of Conferences*, vol. 399, p. 04045, Jul. 2023, doi: 10.1051/e3sconf/202339904045.

[4] I. Nevliudov, V. Yevsieiev, S. Maksymova, N. Demska, K. Kolesnyk, and O. Miliutina, "Object Recognition for a Humanoid Robot Based on a Microcontroller," in *2022 IEEE XVIII International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, Polyana (Zakarpattya), Ukraine, Sep. 2022, pp. 61–64, doi: 10.1109/MEMSTECH55132.2022.10002906.

[5] P. Pareek and A. Thakkar, "A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.

[6] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video Processing Using Deep Learning Techniques: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 139489–139507, 2021, doi: 10.1109/ACCESS.2021.3118541.

[7] W. Qin and J. Meng, "The research on dance motion quality evaluation based on spatiotemporal convolutional neural networks," *Alexandria Engineering Journal*, vol. 114, pp. 46–54, Feb. 2025, doi: 10.1016/j.aej.2024.11.025.

[8] B. Ni *et al.*, "Expanding Language-Image Pretrained Models for General Video Recognition," in *Computer Vision – ECCV 2022 (ECCV 2022)*, 2022, pp. 1–18, doi: 10.1007/978-3-031-19772-7_1.
[9] Y. Tian, Y. Yan, G. Zhai, G. Guo, and Z. Gao, "EAN: Event Adaptive Network for Enhanced Action Recognition," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2453–2471, Oct. 2022, doi: 10.1007/s11263-022-01661-1.
[10] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, no. 2, pp. 690–712, Feb. 2021, doi: 10.1007/s10489-020-01823-z.
[11] H. Wu, X. Ma, and Y. Li, "Spatiotemporal Multimodal Learning With 3D CNNs for Video Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1250–1261, Mar. 2022, doi: 10.1109/TCSVT.2021.3077512.
[12] I. Ghosh, S. R. Ramamurthy, A. Chakma, and N. Roy, "Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 5, Sep. 2023, doi: 10.1002/widm.1496.
[13] A. G. B. Cruz, Y. Seo, and D. Scaraboto, "Between Cultural Appreciation and Cultural Appropriation: Self-Authorizing the Consumption of Cultural Difference," *Journal of Consumer Research*, Apr. 2023, doi: 10.1093/jcr/ucad022.
[14] M. Shoji, Y. Takafuji, and T. Harada, "Behavioral impact of disaster education: Evidence from a dance-based program in Indonesia," *International Journal of Disaster Risk Reduction*, vol. 45, May 2020, doi: 10.1016/j.ijdrr.2020.101489.
[15] A. E. Odefunso, E. G. Bravo, and Y. V. Chen, "Traditional African Dances Preservation Using Deep Learning Techniques," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 4, Sep. 2022, doi: 10.1145/3533608.
[16] I. Arici and R. Niewiadomski, "Positive Emotion Recognition—A Survey of Computational Models," *IEEE Access*, vol. 13, pp. 119131–119156, 2025, doi: 10.1109/ACCESS.2025.3585338.
[17] T. F. N. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujally, and A. Jalal, "A review of video-based human activity recognition: theory, methods and applications," *Multimedia Tools and Applications*, vol. 84, no. 17, pp. 18499–18545, Jul. 2024, doi: 10.1007/s11042-024-19711-w.
[18] R. J. Raj, S. Dharan, and T. T. Sunil, "Optimal feature selection and classification of Indian classical dance hand gesture dataset," *Visual Computer*, vol. 39, no. 9, pp. 4049–4064, Sep. 2023, doi: 10.1007/s00371-022-02572-5.
[19] R. A. Smith and E. S. Cross, "The McNorm library: creating and validating a new library of emotionally expressive whole body dance movements," *Psychological Research*, vol. 87, no. 2, pp. 484–508, Mar. 2023, doi: 10.1007/s00426-022-01669-9.
[20] H. Li and X. Huang, "Intelligent Dance Motion Evaluation: An Evaluation Method Based on Keyframe Acquisition According to Musical Beat Features," *Sensors*, vol. 24, no. 19, p. 6278, Sep. 2024, doi: 10.3390/s24196278.
[21] W. Li, "Comment on 'Philosophical manifestation in dance: bridging movement and thought,'" *Trans/Form/Ação*, vol. 48, no. 3, 2025, doi: 10.1590/0101-3173.2025.v48.n3.e025065.
[22] Y. Chen, S. Wang, D. S. Ametefe, and D. John, "Impact of Dancing on Physical and Mental Health: A Systematic Literature Review," *Dance Research*, vol. 42, no. 2, pp. 220–256, Nov. 2024, doi: 10.3366/drs.2024.0432.
[23] W. Liu, H. Xue, and Z. Y. Wang, "A systematic comparison of intercultural and indigenous cultural dance education from a global perspective (2010–2024)," *Frontiers in Psychology*, vol. 15, Nov. 2024, doi: 10.3389/fpsyg.2024.1493457.
[24] G. Loupas, T. Pistola, S. Diplaris, K. Ioannidis, S. Vrochidis, and I. Kompatsiaris, "Comparison of Deep Learning Techniques for Video-Based Automatic Recognition of Greek Folk Dances," *MultiMedia Modeling (MMM 2023)*, 2023, pp. 325–336, doi: 10.1007/978-3-031-27818-1_27.
[25] S. Badaruddin, J. Masunah, and R. Milyartini, "Two Cases of Dance Composition Learning Using Technology in Dance Education Study Program in Indonesia," *Advances in Social Science, Education and Humanities Research*, 2024, pp. 549–561, doi: 10.2991/978-2-38476-100-5_70.
[26] J. R. Challapalli and N. Devarakonda, "A novel approach for optimization of convolution neural network with hybrid particle swarm and grey wolf algorithm for classification of Indian classical dances," *Knowledge and Information Systems*, vol. 64, no. 9, pp. 2411–2434, Sep. 2022, doi: 10.1007/s10115-022-01707-3.
[27] N. Jain, V. Bansal, D. Virmani, V. Gupta, L. Salas-Morera, and L. Garcia-Hernandez, "An enhanced deep convolutional neural network for classifying indian classical dance forms," *Applied Sciences*, vol. 11, no. 14, Jul. 2021, doi: 10.3390/app11146253.
[28] Y. Abdillah, S. Supriono, and B. Supriyono, "Change and innovation in the development of Balinese dance in the garb of special interest tourism," *Cogent Social Sciences*, vol. 8, no. 1, Dec. 2022, doi: 10.1080/23311886.2022.2076962.
[29] S. Guo, X. Yang, N. H. Farizan, and S. Samsudin, "The analysis of teaching quality evaluation for the college sports dance by convolutional neural network model and deep learning," *Heliyon*, vol. 10, no. 16, p. e36067, Aug. 2024, doi: 10.1016/j.heliyon.2024.e36067.
[30] J. Chen, J. Wang, Q. Yuan, and Z. Yang, "CNN-LSTM Model for Recognizing Video-Recorded Actions Performed in a Traditional Chinese Exercise," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 351–359, 2023, doi: 10.1109/JTEHM.2023.3282245.
[31] S. Wang, J. Li, T. Cao, H. Wang, P. Tu, and Y. Li, "Dance Emotion Recognition Based on Laban Motion Analysis Using Convolutional Neural Network and Long Short-Term Memory," *IEEE Access*, vol. 8, pp. 124928–124938, 2020, doi: 10.1109/ACCESS.2020.3007956.
[32] P. Malavath and N. Devarakonda, "Natya Shastra: Deep Learning for Automatic Classification of Hand Mudra in Indian Classical Dance Videos," *Revue d'Intelligence Artificielle*, vol. 37, no. 3, pp. 689–701, Jun. 2023, doi: 10.18280/ria.370317.
[33] F. Gîrbacia, "An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage," *Electronics (Basel)*, vol. 13, no. 18, p. 3738, Sep. 2024, doi: 10.3390/electronics13183738.
[34] A. Permana, T. K. Shih, A. Musdholifah, and A. K. Sari, "Error Action Recognition on Playing The Erhu Musical Instrument Using Hybrid Classification Method with 3D-CNN and LSTM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 3, p. 313, Jul. 2023, doi: 10.22146/ijccs.76555.
[35] V. Vivilia, "Tari Remo Gagrak Anyar." [Online]. Available: https://youtu.be/RbeyfjuA8ew?si=tv29qVk7WdPvQmKn. (Accessed: Jul. 23, 2025).
[36] J. Febri, "Tari Gambyong Ayun Ayun." [Online]. Available: https://youtu.be/7TJcyUpFvwk?si=ajJqnLb_pad2jSvT. (Accessed: Jul. 23, 2025).
[37] I. N. Cahyono, "Tari Topeng Bapang." [Online]. Available: https://youtu.be/vwbkdN9eXY4?si=ZU_j2hxN8ZNHEiV6. (Accessed: Jul. 23, 2025).
[38] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning," *BMC Medical Imaging*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12880-023-00964-0.

[39] M. Abdullah, M. Ahmad, and D. Han, "Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification," in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, Barcelona, Spain, 2020, pp. 1-3, doi: 10.1109/ICEIC49074.2020.9051332.
[40] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, "Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM," *IET Image Process*, vol. 14, no. 7, pp. 1227–1232, May 2020, doi: 10.1049/iet-ipr.2019.1188.
[41] Y. M. Chen, W. T. Huang, W. H. Ho, and J. T. Tsai, "Classification of age-related macular degeneration using convolutional-neural-network-based transfer learning," *BMC Bioinformatics*, vol. 22, Nov. 2021, doi: 10.1186/s12859-021-04001-1.
[42] A. Ziaee and E. Çano, "Batch Layer Normalization A new normalization layer for CNNs and RNNs," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2022, pp. 40–49, doi: 10.1145/3571560.3571566.
[43] C. Irawan, E. H. Rachmawanto, and H. P. Hadi, "An Ensemble Learning Layer for Wayang Recognition using CNN-based ResNet-50 and LSTM," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 10, no. 1, Feb. 2025, doi: 10.22219/kinetik.v10i1.2053.
[44] X. X. Li, D. Li, W. X. Ren, and J. S. Zhang, "Loosening Identification of Multi-Bolt Connections Based on Wavelet Transform and ResNet-50 Convolutional Neural Network," *Sensors*, vol. 22, no. 18, Sep. 2022, doi: 10.3390/s22186825.

BIOGRAPHIES OF AUTHORS

Candra Irawan received the Bachelor of Computer and Master of Computer form Universitas Dian Nuswantoro respectively in 1999 and 2004. Actively serves as a reviewer for various national conferences and journals, and is a member of the professional certification body at Dian Nuswantoro University. He has been actively teaching since 2005. In 2023, he joined the Computer Science in Art and Culture Research Laboratory to develop scientific knowledge in machine learning and deep learning to optimize the functionality of cultural data. His research interests include machine learning, deep learning for culture. He can be contacted at email: candra.irawan@dsn.dinus.ac.id.

Heru Pramono Hadi is a senior lecturer at Dian Nuswantoro University, having joined in 1994. He is active in research and community service, focusing on marketing-based village empowerment. He currently serves as Head of the Quality Assurance Unit of the Faculty of Computer Science at Universitas Dian Nuswantoro. His research areas of interest are machine learning and deep learning optimization. He can be contacted at email: heru.pramono.hadi@dsn.dinus.ac.id.

Cahaya Jatmoko received Bachelor's and Master's degrees from the University of Dian Nuswantoro respectively in 2002 and 2011. Has been a lecturer since 2012. Until now, he has been active in researching and giving oral presentations in international conferences and international journals. Research interest include image processing, computer vision, pattern recognition, machine learning, and computing data security. He can be contacted at email: cahayajatmoko@dsn.dinus.ac.id.

Mohamed Doheir received his doctorate in Healthcare Management in 2020 from the University Teknikal Malaysia Melaka (UTeM). He received his master's from the University Teknikal Malaysia Melaka (UTeM) in 2012. His current research such as cloud computing, information technology, and system management. In 2022, he served as dean of the Faculty of Technopreneurship, the University Teknikal Malaysia Melaka, Malaysia. He can be contacted at email: doheir@utem.edu.my.