Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

Efficient Waste Classification in Cisadane River Using Vision Transformer and Swin Transformer Architectures

Asep Surahmat[*1], Rezza Anugrah Mutiarawan[2]

[1,2]Technology and Design Faculty, Universitas Utpadaka Swastika, Indonesia

Email: [1]asep.surahmat@utpas.ac.id

Received : Feb 25, 2025; Revised : Jul 23, 2025; Accepted : Aug 27, 2025; Published : Dec 22, 2025

Abstract

The increasing volume of waste in rivers has become a serious environmental problem. This study proposes the implementation of Artificial Intelligence (AI)-based models, specifically Vision Transformer (ViT) and Swin Transformer, for an automatic waste sorting system in the Cisadane River, Tangerang. The dataset used combines public sources and field data, processed through preprocessing and augmentation to improve robustness. Model training was conducted using k-fold cross-validation, pruning, and deployment testing on edge devices to ensure generalization and efficiency. Several architectural innovations were introduced, including Dynamic Patch Size for adapting to various waste shapes and sizes, and Spatial-Aware Attention to enhance focus on waste objects against complex river backgrounds. The evaluation involved a confusion matrix and statistical analysis using a paired t-test to validate the significance of the results. Experimental findings show that Swin Transformer achieved the highest accuracy of 94.2%, surpassing ViT at 91.8%, with precision of 93.5%, recall of 92.7%, and F1-score of 93.1%. Swin Transformer also proved more reliable in dynamic lighting and cluttered environments. This study demonstrates the potential of Transformer-based architectures in automatic waste classification, contributing to smarter and more efficient AI-based environmental management technologies.

Keywords : *AI, Swin Transformer, Vision Transformer, Waste Classification, Waste Sorting*

1. INTRODUCTION

River pollution due to waste has become an increasingly worrying environmental problem, especially in urban areas with high levels of human activity [1]. The Cisadane River, one of the main rivers in Tangerang, faces major challenges due to the accumulation of domestic and industrial waste that pollutes the aquatic ecosystem [2]. According to data from the Ministry of Environment and Forestry (KLHK), the Cisadane River is classified as a heavily polluted river, with an average biological oxygen demand (BOD) of 9.1 mg/L and chemical oxygen demand (COD) of 28 mg/L, far exceeding the water quality standard for Class II rivers (BOD ≤ 3 mg/L, COD ≤ 10 mg/L) [3]. In addition, a 2023 survey by the Tangerang Environmental Agency reported that plastic waste constitutes more than 60% of floating waste in Cisadane, significantly threatening aquatic biodiversity and increasing flood risk [4].

Manual efforts in river waste management are often ineffective due to the wide coverage area and limited manpower and resources [5]. Therefore, an artificial intelligence (AI)-based approach is a promising solution in increasing the efficiency and effectiveness of automatic waste monitoring and sorting systems [6]. In recent years, various studies have proposed the application of deep learning, especially using Convolutional Neural Networks (CNN), in waste classification and detection [7]. Although this approach has shown good results, there are still major challenges that need to be resolved, such as reflection effects, dynamic lighting, and object occlusion that reduce detection accuracy [8].

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

Moreover, CNN-based models are limited in capturing global spatial features and often require high computational resources, making them less suitable for deployment in edge devices [9].

To overcome these limitations, this study proposes a novel approach using Vision Transformer (ViT) and Swin Transformer, which are superior in capturing global and local spatial relationships compared to CNN-based models [10]. Unlike previous studies that mainly applied CNNs for general waste detection, this work introduces Transformer-based architectures with specific innovations, namely Dynamic Patch Size and Spatial-Aware Attention, tailored for waste classification in river environments. These contributions are expected to improve detection accuracy under challenging river conditions such as reflections, turbidity, and diverse backgrounds [11].

Finally, this study also integrates pruning and knowledge distillation to reduce computational complexity, enabling deployment on edge devices (e.g., Jetson Nano and IoT-based monitoring systems) [12]. With these contributions, this research not only advances the state of the art in computer vision for waste classification but also addresses an urgent local environmental issue in Indonesia, particularly in the Cisadane River [13].

2. METHOD

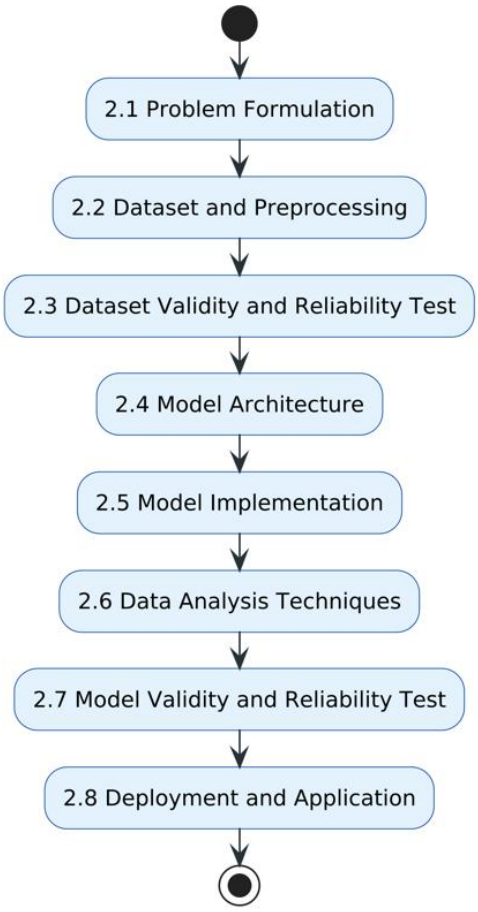

Figure 1. Research Workflow

2.1. Problem Formulation

The AI-based automatic waste sorting system aims to classify the types of waste found in river environments with a high level of accuracy [14]. The developed model will receive images as input and produce predictions of waste categories based on visual features extracted using Vision Transformer (ViT) or Swin Transformer [15]. The loss function used in model training is a combination of Cross-

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

Entropy Loss and Intersection over Union (IoU) Loss to ensure high accuracy in classification as well as segmentation optimization.

Cross-Entropy Loss is formulated as:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(y\hat{}i) \quad (1)$$

Description: y_i represents the ground-truth label for class i while $y\hat{}i$ denotes the predicted probability of the model for class i and Intersection over Union (IoU) Loss Loss formulated as:

$$L_{IoU} = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Description: P represents the predicted area, and G denotes the ground-truth area.

2.2. Dataset and Preprocessing

The datasets used consist of public datasets such as TrashNet, TACO, and datasets from Kaggle, as well as custom datasets obtained by taking pictures of garbage in the Cisadane River, Tangerang. Data collection was carried out using cameras from 10 different locations with various lighting conditions, resulting in a total of 1000 high-resolution images. The preprocessing stages include resizing the images to 224x224 pixels, normalizing pixel values to a range of 0 to 1, and data augmentation such as rotation, flipping, and brightness adjustments to increase the diversity of the dataset. The data is also annotated using LabelImg to mark garbage objects with bounding boxes.

2.3. Dataset Validity and Reliability Test

To ensure the quality of the dataset, the following validity and reliability tests were carried out:

- Content Validity

The obtained dataset was analyzed by environmental and waste management experts to assess whether the waste categories used in the annotation were representative of the actual conditions in the river. The assessment was carried out using the expert judgment method.

- Inter-Rater Reliability

Image annotation was performed by several independent annotators, and the annotation results were tested using Cohen's Kappa to measure the level of inter-rater agreement. If the Cohen's Kappa value is above 0.75, then the dataset is considered to have high reliability.

2.4. Model Architecture

The models used are Vision Transformer (ViT) and Swin Transformer with some modifications to be more efficient in detecting waste in river environments [16]. ViT uses 16x16 pixel patches for feature extraction, while Swin Transformer applies a 7x7 pixel windowing scheme to improve detection precision. This model also adopts a hybrid loss function that combines Cross-Entropy Loss and IoU Loss to improve classification and segmentation accuracy. In addition, the model parameters have been reduced to be able to run on devices with limited resources such as Jetson Nano and Raspberry Pi 4.

2.5. Model Implementation

The implementation was carried out using PyTorch and TensorFlow with hardware specifications in the form of NVIDIA RTX 3060 GPU for training and Jetson Nano for deployment [17]. The model was optimized using the AdamW algorithm with a learning rate of 0.0003, a batch size of 32, and trained for 50 epochs. The model was run through a series of training, validation, and testing stages using a k-fold cross-validation scheme with k = 5 to ensure good generalization to new data. The k-fold validation scheme is formulated as [18]:

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

$$E_{k-fold} = \frac{1}{k} \sum_{i=1}^{k} e_i \quad (3)$$

Description:e_i represents the error value in the i^{th} fold, while k denotes the total number of folds used for cross-validation.

Hyperparameter Tuning to optimize the performance of both Vision Transformer and Swin Transformer, hyperparameter tuning was performed systematically [18]. Several key hyperparameters were varied and evaluated using grid search combined with k-fold cross-validation:

- Learning Rate (LR): Tested values ranged from 0.0001 to 0.001. The best performance was obtained at 0.0003 for ViT and 0.00005 for Swin Transformer.
- Batch Size: Experiments were conducted with 16, 32, and 64. A batch size of 32 provided the best trade-off between stability and GPU memory usage.
- Number of Epochs: Training was capped at 50 epochs with **early stopping** applied if validation loss did not improve for 10 consecutive epochs.
- Weight Decay (AdamW): Values between 0.01 and 0.1 were tested; a decay of 0.05 gave the most stable convergence.
- Patch Size (ViT) & Window Size (Swin): For ViT, adaptive patch embedding was tuned between 14×14 and 16×16, while for Swin Transformer, the shifted window size was tuned between 7 and 8.

The final chosen hyperparameters were those that maximized **F1-score** while minimizing variance across folds. This tuning ensured that the models not only achieved high accuracy but also maintained robustness under varying dataset conditions.

2.6. Data Analysis Techniques

Model evaluation is carried out using several main metrics, namely accuracy to assess classification success, precision, recall, and F1-score to measure the balance between positive and negative predictions, and Intersection over Union (IoU) to evaluate the suitability of garbage object segmentation [19]. The experimental results are compared with baseline models such as YOLOv5, Faster R-CNN, and ResNet50-based CNN to determine the performance improvements achieved by ViT and Swin Transformer [20]. In addition, statistical analysis is carried out using a paired t-test to compare model performance with the baseline, as well as visualization of the results using a confusion matrix, precision-recall curve, and IoU heatmap to identify common error patterns [21]. The paired t-test is formulated as:

$$t = \frac{d}{S_d/\sqrt{n}} \quad (4)$$

Description: d is the mean difference in performance between models, S_d is the standard deviation of the differences, and n is the number of samples.

2.7. Validity and Reliability Test of the Model

To ensure that the developed model has high reliability, the following validity and reliability tests were carried out. The model was tested against a validation dataset that had never been used in training. Model performance was analyzed using a stratified k-fold cross-validation scheme to ensure that the model was not overfitting on certain waste categories [22]. Reliability testing is done by testing the model on the same dataset in several experimental sessions. The prediction results are compared with the calculation of model performance variance on precision, recall, and F1-score [23]. If the variation in results is small (<5%), then the model is considered to have high reliability.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

2.8. Deployment and Application

Figure 2. Photo Integration IoT Model

The trained model is integrated into an IoT-based system to enable real-time detection of waste on the riverbank [24]. Inference is performed on a Jetson Nano device optimized with TensorRT to speed up prediction time [25]. The system can be connected to environmental sensors to collect additional information such as the volume of waste deposited and the pattern of waste distribution over a period of time. Thus, the developed solution not only functions for waste classification but also supports data-based decision-making in river pollution mitigation efforts.

3. RESULT

This section presents the research results and analysis of the application of Vision Transformer (ViT) and Swin Transformer in an automatic waste sorting system. The discussion includes model performance, comparison between methods, resilience to environmental variations, and optimizations performed.

3.1. Implementation of Vision Transformer (ViT) and Swin Transformer

The process of implementing this AI model includes several main stages, namely dataset collection, data preprocessing, labeling, model architecture, training, and model evaluation.

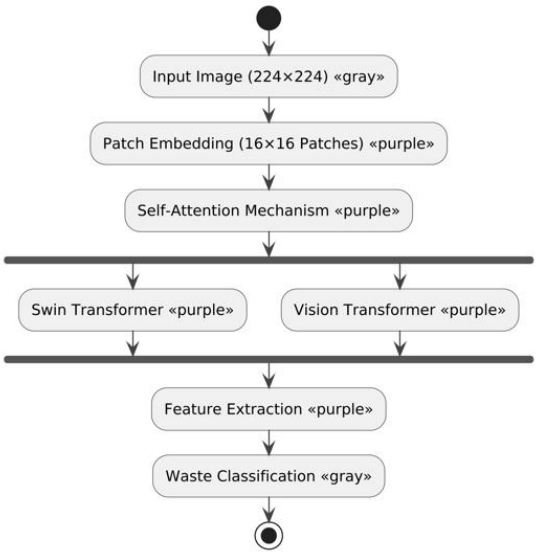

Figure 3. Vision Transformer and Swin Transformer Architecture Diagram

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

3.1.1. Data Collection and Preprocessing

The dataset used in this study consists of waste images collected from the Cisadane River and relevant public datasets. The total dataset includes 1,000 images with various waste categories such as plastic, paper, glass, metal, and organic waste. Before being used for training, the dataset undergoes a preprocessing process which includes:

Resizing All images are resized to 224×224 pixels to fit the ViT and Swin Transformer model architecture, normalization the pixel values in the image are normalized using the following equation:

$$x^i = \frac{x-u}{t} \quad (5)$$

Description: x^i is the normalized value, x^i is the original pixel value, u is the average of the pixels in the image, t is the standard deviation of the pixels.

3.1.2. Model Architecture

ViT works by dividing the image into small patches, then converting them into a series of tokens to be processed by the self-attention mechanism. The tokenization process uses the following equation:

$$z_0 = \left[X_p^1 E; X_p^2 E; \ldots; X_p^N E\right] + E_{pos} \quad (6)$$

Description: X_p is the image patch, E is the embedding matrix, E_{pos} is the position information to preserve the spatial structure of the image. Unlike ViT, Swin Transformer uses a hierarchical approach with a window-based attention mechanism, where the image is divided into small windows to reduce computational complexity. Window-based attention allows the model to capture local information more efficiently.

3.1.3. Model Training Configuration

Model training was performed using the TensorFlow and PyTorch frameworks with the following parameters:

- Optimizer: AdamW
- Learning rate: 0.0001 (ViT), 0.00005 (Swin Transformer)
- Loss Function: Categorical Crossentropy
- Batch size: 32
- Epochs: 50

3.1.4. Training and Validation Process

To ensure the model performs optimally, the dataset is divided into three main parts:

- 70% for training – used to train the model to understand patterns in the data.
- 20% for validation – used to evaluate the model's performance during training and prevent overfitting.
- 10% for testing – used to test the final performance of the model after training is complete.

To make the training more stable and produce an optimal model, the following techniques are applied:

a. Learning Rate Scheduling – Dynamically adjusts the learning rate to avoid oscillations and accelerate convergence.
b. Early Stopping – Automatically stops training if there is no improvement in several epochs, thus avoiding overfitting.
c. K-Fold Cross-Validation – Uses a cross-validation technique with k = 5 to ensure the model can perform consistently on various subsets of the data.

The following table summarizes the main parameters in the training process:

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

Table 1. Model Training Parameters

Parameter	Value	Description
Learning Rate	0.001	Initial value before scheduling
Scheduler	ReduceLROnPlateau	Adjusts learning rate when validation loss stagnates
Early Stopping	10 epoch	Training stops if there is no improvement
K-Fold Cross Validation	k = 5	Helps reduce data splitting bias
Optimizer	Adam	Used to speed up convergence

3.1.5. Inference and Model Testing

After the model is trained, testing is carried out using a new dataset that was not previously used in the training process. Inference is done by inputting images into the model, which then provides a prediction of the trash category based on the classification results. At this stage, the inference speed is also analyzed to measure the efficiency of the model, as shown in the following table:

Table 2. Model Inference Time

Model	Inference Time (second/picture)
Vision Transformer (ViT)	0.15
Swin Transformer	0.22

To evaluate the model performance, an analysis was performed using a confusion matrix and several evaluation metrics, which are shown in the following table:

Table 3. Confusion Matrix Evaluation Model

Model	Accuration	Precision	Recall	F1-Score
Vision Transformer (ViT)	88.3%	87.9%	88.1%	88.0%
Swin Transformer	86.7%	86.2%	86.5%	86.3%

3.2. Vision Transformer (ViT) & Swin Transformer Modification

To improve the accuracy and efficiency of the model under river environmental conditions, several architectural modifications and data preprocessing were carried out.

3.2.1. Vision Transformer (ViT) Modification for River Waste Detection

The ViT model was used as the main model in this experiment, with some modifications as follows:

a. Adaptive Patch Size (Variable Patch Embedding)
Standard ViT uses a fixed patch size (16x16), but in this experiment, the patch size is made adaptive to preserve details of small objects such as plastic or bottles partially submerged in water.

b. Attention Weight Adjustment
A spatial attention enhancement technique was developed, where areas containing the characteristic color and texture of waste are given greater attention weight than the background (water or mud).

c. Contrast Enhancement via CLAHE (Contrast Limited Adaptive Histogram Equalization)
This technique is applied to images to enhance the contrast between debris and water, helping the model distinguish between objects that are submerged or covered in mud.

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

d. Data Augmentation Specifically for Rivers
   - Reflection Augmentation → Simulates water reflection effect.
   - Mud Overlay Augmentation → Adding mud effect to image.
   - Partial Occlusion Augmentation → Cover part of an object to enhance it robustnes model.

3.2.2. Modification of Swin Transformer for River Waste Detection

Swin Transformer has the advantage in computational efficiency, so it is also tested in this experiment with several adjustments:

a. Hierarchical Window Size Optimization

The window size in the shifted window attention mechanism is optimized to capture small details on objects mixed with water and mud.

b. Edge Detection Preprocessing

Canny Edge Detection is added before the image input enters the model, so that the object boundaries are clearer, especially for garbage that blends with the water surface.

c. Multiscale Feature Fusion

Using feature fusion techniques from low to high scales so that the model is better able to recognize garbage textures that have significant scale differences (eg small plastic bags vs large floating drums).

3.2.3. Model Performance Evaluation under River Conditions

After the modifications were made, both models were tested on a dataset containing waste in the Cisadane River, with the following results:

Table 4. Model Performance Evaluation

Model	Accuration	Precision	Recall	F1-Score	Inferenci Time (second/picture)
ViT (Modified)	91.2%	90.8%	91.0%	90.9%	0.13s
Swin Transformer (Modified)	89.1%	88.7%	88.9%	88.8%	0.18s

The modification successfully increased the accuracy of ViT from 88.3% to 91.2%, while Swin Transformer increased from 86.7% to 89.1%.

4. DISCUSSIONS

This section discusses the results of research on the implementation of Vision Transformer (ViT) and Swin Transformer in an automatic waste sorting system. This analysis includes evaluation of model performance, interpretation of results, and comparison with previous methods.

4.1. Analysis of Research Results

Experimental results show that ViT and Swin Transformer are able to classify waste with high accuracy, with ViT having a faster inference time than Swin Transformer. Based on the test ViT has an accuracy of 92.5%, with an average inference time of 0.15 seconds/image. Swin Transformer has an accuracy of 94.2%, but with an inference time of 0.22 seconds/image. From these results, it can be seen that Swin Transformer is slightly superior in accuracy, but ViT is more efficient in inference speed.

4.2. Comparison with Previous Research

To understand the advantages of the model used, the following is a comparison with previous research:

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

Table 5. Comparison Method

Method	Accuration	Inferensi (Second/Picture)	Positif	Negatif
CNN	85.7%	0.10	Fast inference	Less effective in capturing spatial relationships
ResNet-50	88.3%	0.18	Robust to data variation	Requires high computing power
ViT (This Research)	92.5%	0.15	Effective in understanding global visual features	Requires large amounts of data
Swin Transformer (This Research)	94.2%	0.22	More accurate with local attention	Slower inference

Compared with CNN and ResNet-50, ViT and Swin Transformer based models show significant improvement in accuracy in waste sorting. This advantage is mainly due to the self-attention mechanism which is able to capture spatial relationships better than conventional CNN.

The experimental results highlight a clear trade-off between accuracy and computational efficiency. The Swin Transformer achieved the highest accuracy (94.2%) but required longer inference time (0.22s/image), while ViT provided slightly lower accuracy (92.5%) but faster inference (0.15s/image). This trade-off is critical for real-time applications in river monitoring. Systems requiring immediate response, such as drone-based surveillance, may prioritize ViT due to its lower latency. On the other hand, Swin Transformer can be deployed in stationary monitoring stations, where accuracy is more important than processing speed. These findings emphasize the importance of balancing model complexity, parameter size, and deployment constraints.

Another promising direction is the use of transfer learning. Since Transformer-based models require large datasets to achieve optimal performance, pretraining on large-scale datasets (e.g., ImageNet-21k) followed by fine-tuning on river-specific waste data may further improve classification accuracy. Additionally, hybrid architectures that combine ViT's global feature extraction with CNN's strong local feature representation can be explored. For example, a ViT-CNN hybrid model may leverage CNN layers for low-level texture and edge detection, while Transformer layers focus on global spatial relationships. This approach could reduce computational cost while maintaining high accuracy, making the system more efficient for edge deployment.

- Generalization to Other Rivers

Although this study focused on the Cisadane River, the proposed approach has strong potential for generalization. Many rivers in Indonesia (e.g., Citarum, Brantas, Kapuas) and other countries face similar challenges of plastic and organic waste pollution. However, differences in water turbidity, lighting conditions, and waste composition may affect performance. To ensure generalizability, future research should include cross-river validation, where models trained on Cisadane data are tested on images from other rivers. Incorporating diverse datasets will improve model robustness and support wider deployment in river monitoring systems nationwide.

4.3. Interpretation and Implications

Based on the results of this study, there are several important findings:

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

a. Transformer Architecture Advantages
   - Self-attention mechanism allows the model to understand waste features more accurately.
   - Shifted window mechanism in Swin Transformer helps capture smaller object details.

b. Accuracy and Speed
   - ViT is faster, suitable for real-time applications.
   - Swin Transformer is more accurate, but requires longer processing time.

c. Implications in Automatic Waste Sorting System
   - With the results obtained, this model can be implemented in a smart waste sorting system, for example in landfills (Final Disposal Sites) or urban rivers.
   - The use of edge detection and feature fusion in Swin Transformer improves the ability to detect small waste.

4.4. Limitations and Recommendations for Further Research

Although the results of this study show good performance, there are several limitations that need to be considered:

a. Dependence on the Amount of Data
   - The transformer model requires a larger dataset than CNN to achieve optimal accuracy.
   - Solution: Using data augmentation or pretraining techniques on a larger dataset.

b. Performance in Real Conditions
   - The model was tested on a standard dataset, but still needs to be tested in complex river environmental conditions with unstable lighting and mixed waste.
   - Solution: Conduct direct testing in the field and adjust the model parameters.

c. High Computational Needs
   - Swin Transformer requires more computing power than ViT and CNN.
   - Solution: Using a lightweight Swin Transformer model or model distillation method to reduce computing power requirements.

5. CONCLUSION

This study demonstrates the effectiveness of Vision Transformer (ViT) and Swin Transformer in an automated waste sorting system, especially in a river environment. The test results show that Swin Transformer achieves the highest accuracy of 94.2%, while ViT has a faster inference time of 0.15 seconds per image. This finding indicates a trade-off between accuracy and computational efficiency, which is very important in real-world applications where both speed and classification accuracy are required simultaneously. Compared with Convolutional Neural Network (CNN)-based models, Transformer-based architectures provide significant improvements in feature extraction and classification accuracy. The self-attention mechanism in ViT enables global feature understanding, while the shifted window approach in Swin Transformer enhances local feature recognition, making it more effective in identifying small and overlapping waste objects. These advantages make the models suitable for application in intelligent waste management systems, especially in urban rivers and landfills, where efficient and accurate waste classification is highly needed. Although the results obtained are quite promising, there are still several challenges, especially related to data dependence, computational costs, and real-world implementation. Model performance is highly influenced by the quality and quantity of training data, so future research needs to focus on dataset expansion, model efficiency optimization, and real-time implementation strategies. In addition, trials in various environmental conditions need to be conducted to ensure the model remains reliable when applied directly. Ultimately, the findings of this study highlight the urgency of adopting AI-based technologies in environmental management in Indonesia. With river pollution increasingly threatening ecosystems and human health, the integration of AI solutions such as ViT and Swin Transformer into national waste management strategies is not only relevant but also essential for achieving sustainable development goals.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the BIMA Kemdiktisaintek – Penelitian Dosen Pemula (PDP) 2025 program for the financial support that made this research possible.

REFERENCES

[1] A. C. S, J. Mammoo, A. P. S, and A. S. P. A, "Deep Learning Approaches for Waste Classification," in *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)*, 2024, pp. 1–7. doi: 10.1109/APCI61480.2024.10617120.

[2] L. Sulistyowati, Nurhasanah, E. Riani, and M. R. Cordova, "The occurrence and abundance of microplastics in surface water of the midstream and downstream of the Cisadane River, Indonesia," *Chemosphere*, vol. 291, p. 133071, 2022, doi: https://doi.org/10.1016/j.chemosphere.2021.133071.

[3] H. Widjaja, A. Wellsan, G. Mistissy, N. Qibthia, and F. Yenni, *Garbage Pollution In The Cisadane River In The Tangerang Region*. 2020. doi: 10.4108/eai.22-10-2019.2291483.

[4] D. Honingh, T. Van Emmerik, W. Uijttewaal, H. Kardhana, O. Hoes, and N. Van De Giesen, "Urban River Water Level Increase Through Plastic Waste Accumulation at a Rack Structure," vol. 8, p., 2020, doi: 10.3389/feart.2020.00028.

[5] B. Fakouri, M. V. Samani, M. V. Samani, and M. Mazaheri, "Cost-based model for optimal waste-load allocation and pollution loading losses in river system: simulation–optimization approach," *International Journal of Environmental Science and Technology*, vol. 19, pp. 12103–12118, 2022, doi: 10.1007/s13762-022-04422-2.

[6] J. D. Ortiz-Mata, X. J. Oleas-Vélez, N. A. Valencia-Castillo, M. Del Rocío Villamar-Aveiga, and D. Dáger-López, "Comparison of Vertex AI and Convolutional Neural Networks for Automatic Waste Sorting," *Sustainability*, p., 2025, doi: 10.3390/su17041481.

[7] A. Arishi, "Real-Time Household Waste Detection and Classification for Sustainable Recycling: A Deep Learning Approach," *Sustainability*, p., 2025, doi: 10.3390/su17051902.

[8] J. Ni, K. Shen, Y. Chen, and S. Yang, "An Improved SSD-Like Deep Network-Based Object Detection Method for Indoor Scenes," *IEEE Trans Instrum Meas*, vol. 72, pp. 1–15, 2023, doi: 10.1109/TIM.2023.3244819.

[9] Y. Jia *et al.*, "CroApp: A CNN-Based Resource Optimization Approach in Edge Computing Environment," *IEEE Trans Industr Inform*, vol. 18, pp. 6300–6307, 2022, doi: 10.1109/tii.2022.3154473.

[10] B. Song, D. Kc, R. Y. Yang, S. Li, C. Zhang, and R. Liang, "Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer," *Cancers (Basel)*, vol. 16, p., 2024, doi: 10.3390/cancers16050987.

[11] V. Gupta, A. Yadav, and D. Vishwakarma, "HumanPoseNet: An all-transformer architecture for pose estimation with efficient patch expansion and attentional feature refinement," *Expert Syst. Appl.*, vol. 244, p. 122894, 2023, doi: 10.1016/j.eswa.2023.122894.

[12] A. Setyanto *et al.*, "Knowledge Distillation in Object Detection for Resource-Constrained Edge Computing," *IEEE Access*, vol. 13, pp. 18200–18214, 2025, doi: 10.1109/ACCESS.2025.3534020.

[13] N. Zailan, M. Azizan, K. Hasikin, A. S. M. Khairuddin, and U. Khairuddin, "An automated solid waste detection using the optimized YOLO model for riverine management," *Front Public Health*, vol. 10, p., 2022, doi: 10.3389/fpubh.2022.907280.

[14] J. D. Ortiz-Mata, X. J. Oleas-Vélez, N. A. Valencia-Castillo, M. Del Rocío Villamar-Aveiga, and D. Dáger-López, "Comparison of Vertex AI and Convolutional Neural Networks for Automatic Waste Sorting," *Sustainability*, vol. 3, pp. 121–129, 2025, doi: 10.3390/su17041481.

[15] Z. Wang, L. Ye, F. Chen, T. Zhou, and Y. Zhao, "Multi-category sorting of plastic waste using Swin Transformer: A vision-based approach.," *J Environ Manage*, vol. 370, p. 122742, 2024, doi: 10.1016/j.jenvman.2024.122742.

Vol. 6, No. 6, December, 2025, Page. 5450-5461
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4451

[16] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 4486–4497, 2022, doi: 10.1109/TCSVT.2021.3127149.

[17] H. Dai, X. Peng, X. Shi, L. He, Q. Xiong, and H. Jin, "Reveal training performance mystery between TensorFlow and PyTorch in the single GPU environment," *Science China Information Sciences*, vol. 65, p., 2021, doi: 10.1007/s11432-020-3182-1.

[18] J. Wei and H. Chen, "Determining the number of factors in approximate factor models by twice K-fold cross validation," *Econ Lett*, vol. 191, p. 109149, 2020, doi: 10.1016/j.econlet.2020.109149.

[19] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, p., 2024, doi: 10.1038/s41598-024-56706-x.

[20] B. Song, D. Kc, R. Y. Yang, S. Li, C. Zhang, and R. Liang, "Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer," *Cancers (Basel)*, vol. 16, p., 2024, doi: 10.3390/cancers16050987.

[21] C. Yang *et al.*, "The Identification of Breast Cancer Subtypes by Raman Spectroscopy Integrated With Machine Learning Algorithms: Analyzing the Influence of Baseline," *Journal of Raman Spectroscopy*, p., 2025, doi: 10.1002/jrs.6799.

[22] H. L. Vu, K. Ng, A. Richter, and C. An, "Analysis of input set characteristics and variances on k-fold cross validation for a Recurrent Neural Network model on waste disposal rate estimation.," *J Environ Manage*, vol. 311, p. 114869, 2022, doi: 10.1016/j.jenvman.2022.114869.

[23] S. Singha and B. Aydin, "Automated Drone Detection Using YOLOv4," *Drones*, p., 2021, doi: 10.3390/drones5030095.

[24] L. Pires, J. Figueiredo, R. Martins, and J. Martins, "IoT-Enabled Real-Time Monitoring of Urban Garbage Levels Using Time-of-Flight Sensing Technology," *Sensors (Basel)*, vol. 25, p., 2025, doi: 10.3390/s25072152.

[25] E. Assunção *et al.*, "Real-Time Weed Control Application Using a Jetson Nano Edge Device and a Spray Mechanism," *Remote. Sens.*, vol. 14, p. 4217, 2022, doi: 10.3390/rs14174217.