

# **University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)**

**ISSN: 2714-3627**

*A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria*

**Volume 17 No. 1, June 2026**

**[journals.ui.edu.ng/uijslictr](http://journals.ui.edu.ng/uijslictr)**

**<http://uijslictr.org.ng/>**

**[uijslictr@gmail.com](mailto:uijslictr@gmail.com)**



## Waste Classification Using Pre-Trained Vision Transformer and CNN Models

<sup>1</sup>Sawyer B. A., <sup>2</sup>Fasina E. P., <sup>3</sup>Afolabi S. P. and <sup>4</sup>Iyekolo O. G.

<sup>1</sup>Department of Artificial Intelligence and Robotics, University of Lagos

<sup>1</sup>bsawyer@unilag.edu.ng, <sup>2</sup>efasina@unilag.edu.ng

<sup>3</sup>Department of Computer Sciences, University of Lagos

<sup>3</sup>sundaypeterafolabi2308@gmail.com, <sup>4</sup>iyekolobusayomi@gmail.com

### Abstract

Efficient waste classification is crucial for sustainable waste management, particularly in rapidly developing areas. In this paper, the performance of the Vision Transformer (ViT) and Convolutional Neural Network (CNN) models (VGG19, ResNet50V2, InceptionResNetV2, and DenseNet121) was evaluated on a publicly available dataset comprising a total of ten waste categories. All models were trained using the same preprocessing and data augmentation techniques, and their accuracy, precision, recall, F1-score, and loss values were tested. Overall, ViT outperformed all CNN baselines with a test accuracy of 96.24%, an F1 Score of 0.96, a recall of 95% and a precision of 96%. VGG and ResNet50 remained competitive among the CNN models, but ViT demonstrated a better generalisation on the test set, which opens a potential use: automated waste classification. However, there is a need for more testing on real-world and domain-specific datasets to improve recycling accuracy and to provide cleaner and more efficient waste management.

**Keywords:** Waste Classification, Vision Transformer, Convolutional Neural Networks, Transfer Learning, Deep Learning.

### 1. Introduction

Rapid population growth and urbanisation in Nigeria have led to a substantial increase in waste generation, particularly in densely populated urban areas. Daily human activities produce large volumes of waste, and recent statistics [1] indicate that Nigeria generates over 30 million tonnes annually, with projections rising to 107 million tonnes by 2050. In many cases, waste disposal is not properly carried out because waste is scattered on streets, dumped in bushes, or deposited at unregulated sites and landfills. In addition, waste is incinerated without proper classification, thereby posing risks to municipalities.

Poor waste management predisposes public health to harm, and it contributes to the spread of disease, degradation of the environment, and the creation of unpleasant odours in the community. Therefore, efficient waste disposals

are important for improving both public and environmental health [2]. Several existing waste management systems in Nigeria are inadequate because of the low awareness and limited understanding of proper waste classification rules by the general public. Therefore, effective solutions to waste disposal require automated and reliable classification systems that support the appropriate handling of different waste categories [3], [4].

In Nigeria, accurate classification of landfill and recyclable waste can reduce material contamination, landfill overflow, and avoidable environmental harm. Studies on municipal solid waste management in Nigeria highlight persistent challenges, including weak regulation, inadequate funding, limited infrastructure, poor strategic planning, and insufficient professional involvement [5]. In this context, deep learning offers a practical approach to image-based waste recognition, as pre-trained models can be adapted to new classification tasks with less training data and reduced training time. However, the comparative value of transformer-based models

for waste classification remains less established than that of CNNs. In this study, four CNN architectures—VGG19, ResNet50V2, DenseNet121, and InceptionResNetV2—are compared against a Vision Transformer (ViT) on a labelled multiclass waste-image dataset.

This paper makes three specific contributions. It first provides a direct comparison between ViT and four widely used CNN baselines under the same experimental conditions. It then evaluates performance across ten waste categories using both aggregate metrics and confusion-matrix evidence. Finally, it relates the observed classification errors to practical sorting concerns, particularly where visually similar materials may affect recycling quality

## 2. Related Works

Automated waste classification has made strides with deep learning, particularly using convolutional neural networks (CNNs) and pre-trained models because of the ability of the models to learn visual patterns that are difficult to describe using hand-crafted rules. A variety of architectures, datasets and optimisation techniques have been explored before to improve performance.

The study in Shah and Kamat [6] focused on two types of household wastes, namely the organic and recyclable classes. Two models of image classifiers were created, one based on low-resolution images and the other on high-resolution images. Both models employed a five-layer 2D CNN, and they were tested for 50 epochs using accuracy and cross-entropy metrics, with a random-guess classifier as the baseline. The aim was to minimise the computation time by comparing the resolutions of the images. The low-resolution model gave better results with 80.88 % accuracy compared to the high-resolution model with 76.19 % accuracy. The study, however, concentrated only on organic and recyclable waste, exclusive of transfer learning.

In Abu-Qdais, Shatnawi, and Al-Alamie [7], image processing and machine learning techniques were combined to build intelligent systems for solid waste classification. The performance of traditional models like Random Forest and Support Vector Machine (SVM) was compared and found to be less effective than the CNN model, which achieved 92.7% accuracy.

Using images with multiple objects improved model prediction performance, although some objects were still misclassified.

Amin, Sami, and Hassan [8] showed that transfer learning algorithms are able to achieve higher precision accuracy than standard machine learning algorithms by benchmarking Xception, ResNet50, DenseNet121, MobileNetV2, and EfficientNetB7. They were able to show that DenseNet121 performed best among the models. Islam, Hassan, and Akhtar in [9] also demonstrated that transfer learning can be very effective in solid waste classification. They also showed that model selection can significantly affect the accuracy of waste classification. Data augmentation was used to increase the variety of images. Similarly, Alalibo, Oborkhale, and Omijeh [10] introduced a fine-tuned ResNet50 model to improve solid waste classification, and training on the TrashNet dataset produced better results than other existing CNN architectures.

A CNN-based classifier for waste management was introduced by Poudel and Poudyal in [11] to conduct a comparative analysis of pre-trained models in classifying waste items into seven distinct categories: trash, plastic, paper, organic, metal, glass, and cardboard. The work involved fine-tuning several pre-trained CNN models - including Inception V3, InceptionResNetV2, MobileNet, DenseNet201, Xception, ResNet50, and VGG19 - on a waste image dataset, with images resized specifically for each model. Post-tuning, classification was performed using the softmax activation function. DenseNet201, InceptionResNetV2, and ResNet50 emerged as the top performers, exhibiting minimal discrepancies between training and validation accuracy. Conversely, InceptionV3 attained higher training accuracy but demonstrated signs of overfitting due to marked differences between its training and validation results. The study acknowledges the limitations of the TrashNet dataset and notes that efficient model training necessitates the availability of larger datasets.

In a subsequent study, Srivatsan, Dhiman and Jain [12] used the larger CompostNet dataset to study transfer learning for classifying waste into seven categories. The three pre-trained models that were used, DenseNet121, ResNet34 and MobileNetV2, were trained using 60 epochs with a batch size of 16. DenseNet performed best with a classification accuracy of 96.24%.

Furthermore, two CNN models developed by Rayhan and Rifai [3] were tested using both full learning and transfer learning strategies. Full learning was carried out with a sequential approach, and transfer learning was done with MobileNetV2 and DenseNet121. The results indicated that the accuracy of transfer learning was higher. The data was diverse and it caused class imbalance, which could have created bias during testing. To reduce the imbalance, the quantities of images in larger classes were restricted to give a more uniform distribution throughout the categories.

While previous work demonstrated that CNN and transfer learning can be successful for waste classification, there are still many areas that have not been addressed. Most of the research tends to be carried out on CNN-based models, with only a small amount of work on transformer-based models like Vision Transformers (ViTs). Moreover, many studies are based on a small number of cases and are therefore not generalisable. Very few studies offer comparative evaluations under fixed experimental conditions. To tackle these problems, this paper compares four pre-trained CNN models with a Vision Transformer using a single framework for multiple-class waste classification.

### 3. Methodology

#### 3.1 Datasets

The dataset used in this study was obtained from the Kaggle Garbage Dataset [13]. It contains images from ten waste categories: trash, shoes,

plastic, paper, metal, glass, clothes, cardboard, biological, and battery. The original dataset comprises 19,762 images. The dataset was downsampled to a maximum of 1,100 images per category to reduce the risk of overfitting during training, because some categories contained more samples than others. After downsampling, the dataset contained 10,506 images. The dataset was split into training (70%), validation (20%) and testing (10%) datasets.

#### 3.2 Data Preprocessing and Augmentation

The training, validation, and test datasets were preprocessed to satisfy the input format of each CNN model. This process included pixel scaling and channel reordering to ensure compatibility with the pre-trained architectures. Figures 2 and 3 show sample dataset images and examples of augmented images.

For most models, images were resized from 256×256 to 224×224 pixels, whereas InceptionResNetV2 required 299×299 inputs because of its architectural design. Standard preprocessing, including normalising pixel values by dividing each pixel value by 1/255. It was applied to all datasets that the models will use as input. To improve generalisation, the training set was augmented using rotation (20°), horizontal flipping, zoom (0.2), shear (0.2), height and width shifts (0.2), brightness adjustment (0.9–1.1), and nearest fill mode. The accepted image formats were JPG, JPEG, and PNG.

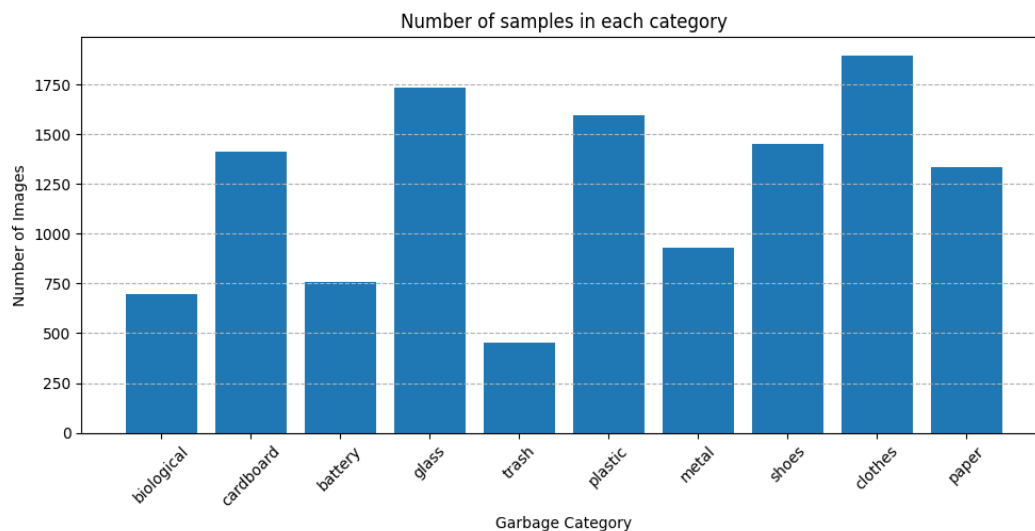


**Figure 1.** Waste class distribution



**Figure 2:** Images from the dataset with their respective classes [13]

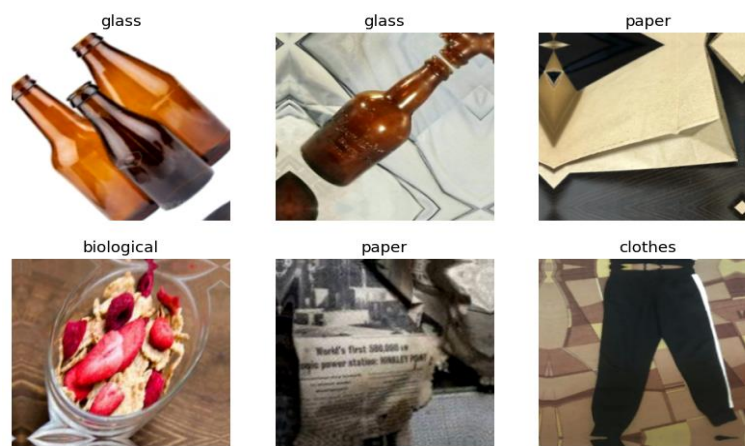


**Figure 3:** Random Augmented Images

### 3.3 Model Architectures

Five pre-trained models were evaluated: VGG19, ResNet50V2, DenseNet121, InceptionResNetV2, and ViT. The CNN models were initialised with ImageNet weights and adapted to the waste-classification task through transfer learning. ViT was implemented as a pre-trained image-classification transformer using 224×224 input images, patch-based tokenisation, transformer encoder blocks, multi-head self-attention, and a task-specific softmax output layer for the ten classes. The same data partitions were used for all models so that performance differences would reflect model behaviour rather than differences in data exposure.

### 3.4 Transfer Learning and Fine-Tuning

Transfer learning uses pre-trained models as the basis for new neural network architectures, eliminating the need to build a convolutional neural network from scratch. In this study, the

pre-trained layers were used as a starting point, and the convolutional layers were first frozen, and additional classification layers were added for the classes of waste of interest. The custom layers consisted of Global Average Pooling (GAP), dense layers with ReLU activation, a dropout rate of 0.5, and a softmax output layer for multiclass classification. The performance of pre-trained models has been very good in image classification tasks to date [14].

Fine-tuning is the addition of custom layers on top of the pre-trained network after having frozen the layers of the base network. The model is initialised with ImageNet weights, and base layers are not trained during the training. Further output layers used in the Keras Sequential API were Global Average Pooling 2D, dense layers with ReLU and softmax activations, and a dropout layer with a dropout rate of 0.5. The combination of the frozen base layers with the

output task-specific layers was employed to boost the performance of classification.

### 3.5 Training Configuration

1. Optimiser: Adam
2. Loss Function: Sparse Categorical Cross-Entropy
3. Batch Size: 32
4. Epochs: 10
5. Seed: 42
6. Hardware accelerator: T4 GPU

Model selection was based on validation performance, with the model achieving the highest validation accuracy or the lowest validation loss chosen for final testing and comparison.

### 3.6 Feature Extraction

Feature extraction is the process of identifying features that are unique to an image that assist in distinguishing it from other images. In a pre-trained neural network, the features are generally learned by the convolutional layers, which are the fundamental components of the network. One popular method is to remove fully connected layers, freeze the rest of the layers, and introduce new fully connected layers as the output layer to train the network on the target data.

### 3.7 Training

Sparse categorical cross-entropy was used as the loss function to measure the average divergence between the true and predicted class probability distributions. The overall model performance was assessed using accuracy, and the Adam optimiser was employed to train the model efficiently. The models were trained for 20 epochs over the augmented training set, where each epoch is a single complete cycle over the training data. The performance was monitored by validating after each epoch. The accuracy and loss of models were measured, and the model with the highest validation accuracy was chosen for further analysis.

## 4. Results and Discussion

In this paper, the performance of a Vision Transformer and four pre-trained CNN models

is evaluated on the Garbage Dataset. The dataset contains ten classes, and the data was divided into 8581 training samples, 2453 validation samples and 1225 test samples. The following models were evaluated: VGG19, ResNet50V2, DenseNet121, InceptionResNetV2, and Vision Transformer (ViT). All were trained with transfer learning with similar experimental conditions so as to make a fair comparison. For architectural reasons, InceptionResNetV2 was fed with an input size of 299×299, while the other models had an input size of 224×224.

### 4.1 Evaluation Metrics

The following metrics were used to evaluate the performance of the model:

1. Test Accuracy: It shows the number of correctly predicted images of waste divided by the total number of images and can give a general idea of the model's effectiveness.
2. Test Loss: This shows the accuracy of a model when applied to unseen data. Loss decreases indicate greater generalisation and fewer errors in predictions.
3. Confusion Matrix: This gives a class-wise distribution of predictions and shows misclassifications by classes for 10 waste classes.

From the confusion matrix, the following metrics were computed:

1. Precision is the proportion of correct positive predictions relative to the total number of predicted positives.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

2. Recall is the ratio of correct positive predictions to the total number of actual positives.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. F1-score is a balanced measure of precision and recall metrics, which is the harmonic mean.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Where:

TP (True Positive) = diagonal value

FP (False Positive) = Column sum - TP

FN (False Negative) = Row sum - TP

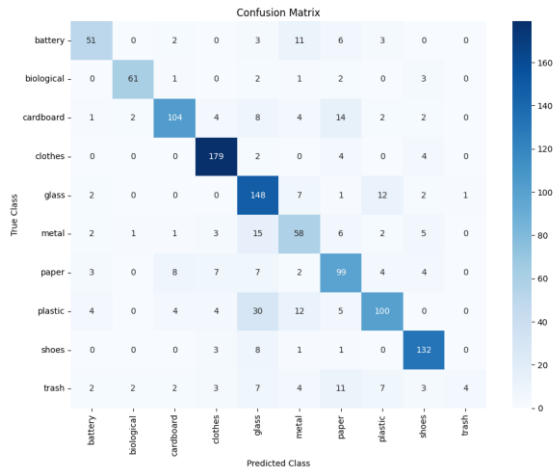


Figure 4: InceptionResNetV2 Confusion Matrix

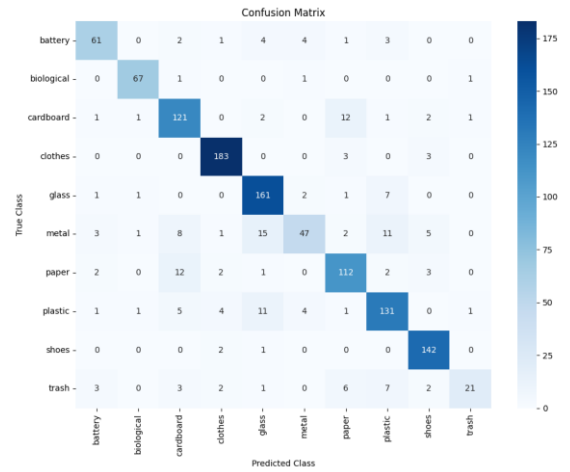


Figure 5: DenseNet121 Confusion Matrix

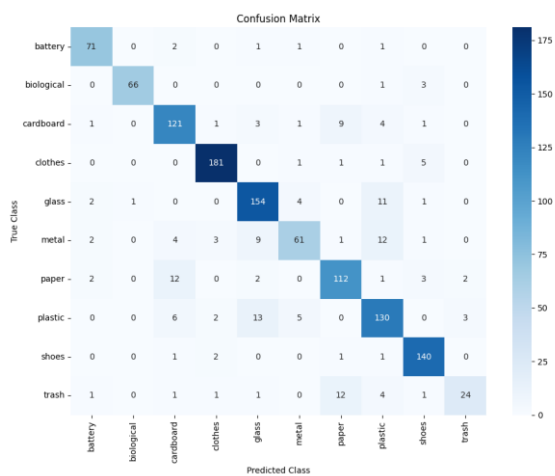


Figure 6: ResNet50V2 Confusion Matrix

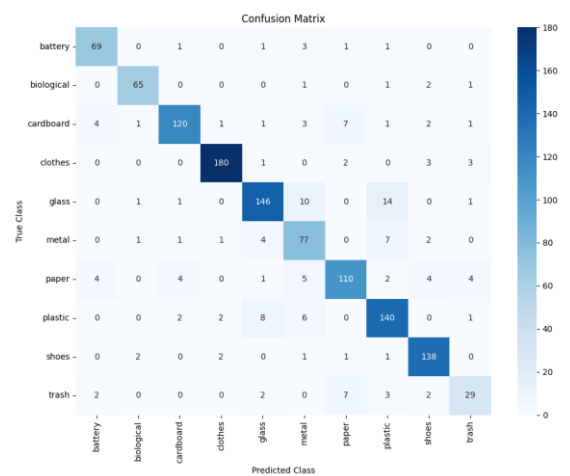


Figure 7: VGG19 Confusion Matrix

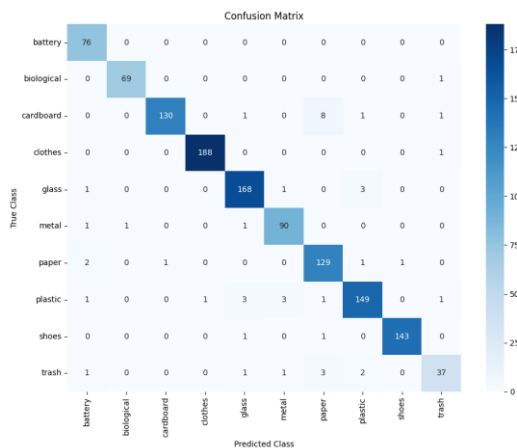


Figure 8: ViT Confusion Matrix

In Figures 4 - 8, the confusion matrices show that ViT was able to separate the waste classes more neatly and better than all the CNN models. The CNN models displayed more visible confusion among categories with similar visual features, especially plastic, glass, metal, cardboard, and paper. In a real-world sorting system, these errors can affect the performance of the sorting system

because misclassification of biodegradable or harmful substances can contaminate material streams and diminish the efficiency of downstream recycling decisions.

The most notable class-level errors occurred among materials with similar colours, shapes, textures, or packaging features. Plastic, glass,

metal, cardboard, and paper can look alike in uncontrolled images, especially when items are crushed, dirty, partially obscured, or captured against cluttered backgrounds. These findings suggest that future deployment would benefit from locally collected images, better lighting

control, and the possible use of additional metadata or multi-view imaging.

#### 4.2 Discussion

The comparative performance of the evaluated models is presented in Table 1

Table 1: Comparative performance of the models

| Model             | Test Acc. | Test Loss | Precision | Recall | F1-Score |
|-------------------|-----------|-----------|-----------|--------|----------|
| InceptionResNetV2 | 0.7641    | 0.6919    | 0.78      | 0.71   | 0.71     |
| DenseNet121       | 0.8539    | 0.6298    | 0.86      | 0.81   | 0.82     |
| ResNet50V2        | 0.8653    | 0.4884    | 0.87      | 0.84   | 0.85     |
| VGG19             | 0.8767    | 0.3545    | 0.86      | 0.86   | 0.86     |
| ViT               | 0.9624    | 0.1291    | 0.96      | 0.95   | 0.96     |

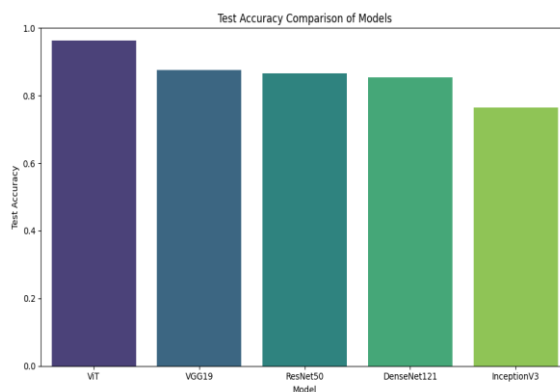


Figure 9: Test Accuracy

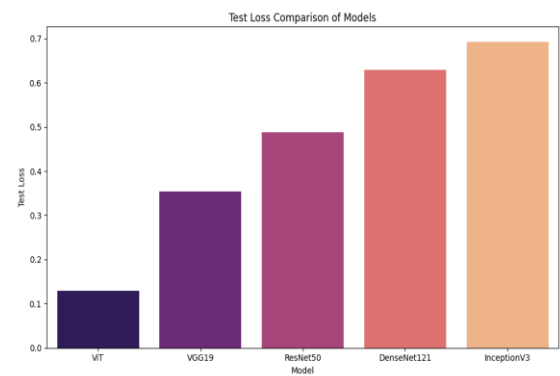


Figure 10: Test Loss

The results in Table 1 and Figures 9–10 show a consistent ranking across the evaluation metrics. ViT achieved the highest accuracy, lowest loss, and strongest F1-score, which means its performance advantage was not limited to a single measure. The CNN results also show that VGG19 and ResNet50V2 remained competitive, while InceptionResNetV2 performed least effectively under the present experimental conditions.

#### 4.3 Discussion

The experimental results obtained show that the Vision Transformer (ViT) outclassed all CNN-based models by achieving a peak test accuracy of 96.24% across 1,225 test cases. In comparison, VGG19, ResNet50V2, InceptionResNetV2 and DenseNet121 recorded accuracies of 87.67%, 86.53%, 76.41% and 85.39%, respectively. The self-attention mechanism in ViT is one of the key reasons for its success, as it is able to capture long-range dependencies and global context information with greater accuracy than

convolutional architectures, which rely on local features. This led to ViT achieving the lowest test loss and the highest F1-score, showing a good generalisation to a wide range of waste categories.

Within the CNN models, VGG19 and ResNet50V2 exhibited good and stable results, suggesting that deep hierarchical feature extraction is still useful for multiclass waste classification. InceptionResNetV2 was the least effective due to its complexity and seeming sensitivity to input variability and class overlaps. Furthermore, the confusion matrices revealed that the CNN models tend to often confuse between similar categories, while the ViT model can correctly classify them with lower false positive and false negative rates.

The relationship between test loss and accuracy is clearly inverse, meaning that the more accurately one generalises, the more accurately it predicts

unseen images. ViT obtained the best F1-score, which showed that it has a good balance between precision and recall, therefore indicating that it can be used for automated waste management systems. The results obtained in this study require further evaluation on a more diverse and real-world local dataset to establish the generalisability and robustness of the results for practical recycling applications.

Deployment of the ViT model for waste sorting requires more performance measurements than a high classification accuracy. A practical system would also require acceptable inference speed, manageable memory, reliable camera input, and robustness in handling shadows, motion blur, occlusion, and mixed waste. ViT showed a promising performance on the present dataset, but its suitability for edge devices or embedded sorting systems would need to be tested.

#### 4.4 Limitations

There are several limitations that should be noted, such as the use of a publicly sourced dataset rather than one collected locally from the Nigerian waste-management environments. If a locally sourced dataset is used, there is likely going to be a domain shift if the model is deployed locally. In addition, the study did not examine inference time, memory consumption, energy use, or hardware feasibility. Future work will include repeated experiments, external datasets, locally collected images, and pilot testing in real-life sorting conditions.

#### 5. Conclusion

The Vision Transformer (ViT) achieved the highest performance in terms of all the evaluation metrics. ResNet50V2 and VGG19 performed better than the InceptionResNetV2 model. Furthermore, the F1 score confirmed that the performance of ViT was the best in terms of accuracy, error reduction, and reliability for class recognition.

The results indicate that transformer models like ViT can achieve better performance in the garbage classification task than conventional CNNs. While CNNs are useful, they are not as good as the attention mechanism of ViT at capturing the global context. In summary, the accuracy, generalisation and classification capabilities of ViT make it a promising solution for smart waste management systems.

#### References

- [1] Agbo, M. Jr, "The Problem with Solid Waste Management in Nigeria's Low-Income Neighbourhoods," Earth.org. <https://earth.org/the-problem-with-solid-waste-management-in-nigerias-low-income-neighbourhoods/>. Aug. 2023.
- [2] Etim, E., Choedron, K. T. and Ajai, O. "Municipal solid waste management in Lagos State: Expansion diffusion of awareness," *Waste Management*, vol. 190, pp. 261–272, Dec. 2024, doi: 10.1016/j.wasman.2024.09.032.
- [3] Rayhan, Y. and Rifai, A. P. "Multi-class Waste Classification Using Convolutional Neural Network," *Applied Environmental Research*, Jun. 2024, doi: 10.35762/AER.2024021.
- [4] Nnamoko, N, Barrowclough, J. and Procter, J. "Solid Waste Image Classification Using Deep Convolutional Neural Network," *Infrastructures (Basel)*, vol. 7, no. 4, p. 47, Mar. 2022, doi: 10.3390/infrastructures7040047.
- [5] Onungwe, I., Hunt, D. V. L. and Jefferson, I. "Transition and implementation of circular economy in municipal solid waste management system in Nigeria: A systematic review of the literature," *Sustainability*, vol. 15, no. 16, p. 12602, Aug. 2023, doi: 10.3390/su151612602.
- [6] Shah J. and Kamat, S. "A Method for Waste Segregation using Convolutional Neural Networks," in *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, Apr. 2022, pp. 1–5. doi: 10.1109/ICAECT54875.2022.9807969.
- [7] Abu-Qdais, H., Shatnawi, N. and Al-Alamie, E. "Intelligent system for solid waste classification using a combination of image processing and machine learning models," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 37, no. 7, pp. 1063–1074, Oct. 2025, doi: 10.1080/0952813X.2024.2323043.
- [8] Amin, Z. M. A., Sami, K. N. and Hassan, R. "An approach of classifying waste using transfer learning method," *International Journal on Perceptive and Cognitive Computing*, vol. 7, no. 1, pp. 41–52, 2021.
- [9] Islam, M. M., Hassan, S. and Akhtar, M. N. "Transfer learning-based solid waste classification using organic and recyclable waste images," in *Proceedings of the Waste Safe 2023: 8th International Conference on Integrated Solid Waste and Faecal Sludge Management*, M. Alamgir, S. M. T. Islam, and J. A. Saju, Eds., Khulna, Bangladesh, Feb. 2023, pp. 1–8.
- [10] Alalibo, T.-O. J., Oborkhale, L. I. and Omijeh, B. O. "Leveraging transfer learning with a modified ResNet-50 model for enhanced solid

- waste classification,” *Global Scientific Journal*, 2023.
- [11] Poudel, S. and Poudyal, P. “Classification of Waste Materials using CNN Based on Transfer Learning,” in *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, New York, NY, USA: ACM, Dec. 2022, pp. 29–33. doi: 10.1145/3574318.3574345.
- [12] Srivatsan, K., Dhiman, S. and Jain, A. “Waste Classification using Transfer Learning with Convolutional Neural Networks,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 775, no. 1, p. 012010, Jun. 2021, doi: 10.1088/1755-1315/775/1/012010.
- [13] Kunwar, S. “Garbage Dataset,” Kaggle, 2024. doi: 10.34740/KAGGLE/DSV/10182596.
- [14] Du, X., Sun, Y., Song, Y., Sun, H., and Yang, L. “A comparative study of different CNN models and transfer learning effect for underwater object classification in side-scan sonar images,” *Remote Sensing*, vol. 15, no. 3, p. 593, Jan. 2023, doi: 10.3390/rs15030593.