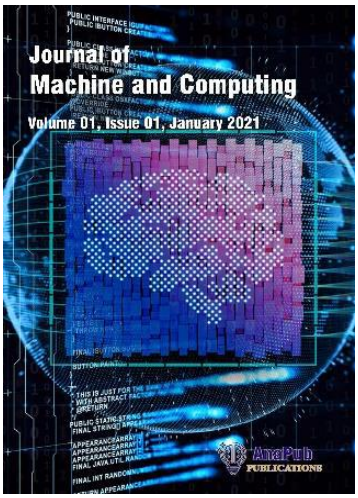
Journal Pre-proof

A Deep Learning Approach to Smart Waste Classification for Sustainable Environments



Vishnu Tej Y, Ashwitha A, Lakshmi H N, Vuppala Balaji, Suryanarayana G, and Sirish Kumar M

DOI: 10.53759/7669/jmc202505119

Reference: JMC202505119

Journal: Journal of Machine and Computing.

Received 24 November 2024

Revised form 02 March 2025

Accepted 25 May 2025

Please cite this article as: Vishnu Tej Y, Ashwitha A, Lakshmi H N, Vuppala Balaji, Suryanarayana G, and Sirish Kumar M, ‘‘A Deep Learning Approach to Smart Waste Classification for Sustainable Environments”, Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505119.

AnaPub
PUBLICATIONS

A Deep Learning Approach to Smart Waste Classification for Sustainable Environments

[1]Y. Vishnu Tej, [2]Ashwitha A, [3]Lakshmi H N, [4]Vuppala Balaji, [5,*]G Suryanarayana, [6]M Sirish Kumar

[1]Associate Professor, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering (A), Karakambadi Road, Tirupati, India.
[2]School of Computer Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education,Manipal, India.
[3]Professor, Department of Computer Science and Engineering (AI&ML), CVR College of Engineering, Hyderabad.
[4]Associate Professor, Department of CSE(AI&ML), Vardhaman College of Engineering, Hyderabad, India.
[5]Symbiosis Institute of Technology, Hyderabad Campus, Symbiosis International University, Pune, India.
[6]Associate Professor, Department of Data Science, School of Computing, Mohan Babu University, Tirupati, India.
[1]vishnutej.br@gmail.com, [2]ashwitha.a@manipal.edu, [3]hn.lakshmi@cvr.ac.in,
[4]vuppalabalaji802@gmail.com, [5,*]surya.aits@gmail.com, [6]drmsk21@gmail.com

[*]Corresponding Author: surya.aits@gmail.com

Abstract – A key element of sustainable development is efficient trash classification, which aims to minimize environmental damage and expedite recycling procedures. In addition, being time-consuming, traditional human sorting methods are prone to mistakes, which makes waste management systems less effective. Automated garbage classification has attracted so much attention as AI, especially ML and DL, has grown. However, because they frequently rely on small-scale datasets and traditional architectures, many of the models that are now in use have issues with generalization, poor performance, and high error rates. This work presents a hybrid deep learning system that combines an autoencoder with a vision transformer (ViT) to address these issues. By efficiently capturing local and global data, our design improves classification robustness and accuracy across various waste types. Our model was trained and assessed using a sizable and varied dataset to enhance generalization to real-world scenarios. According to experimental data, the suggested model achieves a precision of 96.72%, a recall of 96.21%, an F1-score of 96.46%, and a balanced accuracy of 96.48%, outperforming some cutting-edge CNN-based architectures. Furthermore, sophisticated measures like Cohen's Kappa (95.90%) and Matthews Correlation Coefficient (MCC = 94.91%) confirm the dependability of our solution. Lastly, by successfully implementing the model in an inference pipeline, we show that it is ready for real-world deployment and that it has the potential to promote sustainable development goals through scalable, intelligent waste management.

Keywords - Sustainable development, Waste classification, Vision transformer, Autoencoder, Hybrid model

I. INTRODUCTION

The increase in solid waste output caused by rapid urbanization, industrialization, and population growth worldwide poses a severe danger to environmental sustainability. If these patterns continue, the World Bank predicts worldwide waste generation will surpass 3.40 billion tones' annually by 2050 [1]. Implementing intelligent waste management systems has emerged as a critical priority for creative and sustainable urban infrastructure as countries work to achieve the UN's SDGs, particularly objectives eleven (Sustainable Cities and Communities) and twelve (Responsible Consumption and Production). Traditional rubbish management systems lean heavily on manual sorting and fundamental mechanical separation, which are inefficient, labor-intensive, error-prone, and often economically unsustainable. Moreover, the complexity of modern garbage streams, including plastics, metals, glass, paper, and organic materials, requires sophisticated categorization techniques that can adapt to variability in appearance, contamination, and composition. Automated waste classification using computer vision and artificial intelligence has gained substantial traction in this context. Deep learning algorithms show great promise for reliably recognizing and classifying waste materials from photos, allowing for more inventive recycling systems and lowering the environmental impact of unmanaged trash [2], [3].

Convolutional Neural Networks (CNNs) have been the cornerstone of image classification tasks in recent years, achieving notable success in medical imaging, autonomous driving, and industrial automation [4]. However, CNNs exhibit certain limitations when applied to waste classification. Their reliance on local receptive fields and translation invariance makes

capturing long-range spatial dependencies and contextual relationships challenging, especially in cluttered or occluded images often found in real-world waste environments [5]. Additionally, many waste classification datasets are small or imbalanced, hindering the performance of CNNs trained from scratch. Transfer learning from generic datasets such as ImageNet is commonly used to alleviate this issue, but domain mismatch frequently leads to suboptimal generalization [6]. This work proposes a novel technique based on the Vision Transformer Autoencoder (ViT-AE) architecture to address these challenges. ViTs have emerged as a powerful alternative to CNNs, successfully employing self-attention mechanisms to describe the global context and long-range interdependence [7]. Unlike CNNs, ViTs divide input images into fixed-size patches processed sequentially, allowing the model to comprehend the links between distant portions of the image. This property is particularly advantageous for waste classification tasks where distinguishing features may be spatially distant or subtle.

The proposed ViT-Autoencoder architecture integrates ViTs' strengths with an unsupervised learning paradigm through masked image modeling, qualifying the system to learn meaningful visual representations even without large labeled datasets. By pretraining the model to reproduce masked image patches, we give it a good concept of the waste domain before fine-tuning it using labeled data. This two-stage learning strategy improves classification accuracy, generalization, and data efficiency [8]. To validate our approach, we used a high-resolution custom dataset comprising 15,000 labeled images spanning 30 diverse waste categories, including plastic bottles, glass shards, cardboard boxes, aluminum cans, food waste, and e-waste. The dataset incorporates significant intra-class variation, lighting differences, background noise, and occlusions, making it a realistic benchmark for evaluating waste classification models. Extensive experiments demonstrate that our ViT-AE model achieves superior accuracy and robustness compared to baseline CNN architectures such as ResNet, MobileNet, and DenseNet. The contributions of this study are summarized as follows:

- A Novel Model Architecture: We propose a hybrid Vision Transformer Autoencoder model that leverages unsupervised pretraining and fine-tuning for robust waste image classification.
- Custom Dataset Creation: We construct a large-scale, multi-class, high-resolution dataset tailored for real-world waste scenarios, supporting future research in sustainable AI.
- Comprehensive Evaluation: We conduct detailed experiments comparing our model with several state-of-the-art CNN baselines using precision, recall, F1-score, and inference time metrics.
- Practical Implications: We show how the model can be used in smart bins, automatic sorting conveyors, and recycling facilities, helping to promote ecologically responsible urban life.

II. Literature Review

In recent years, research has increasingly leveraged deep learning models for intelligent waste classification to support environmental sustainability. In 2025, Qi et al.[9] proposed an enhanced EfficientNetV2 model incorporating CE-Attention and SAFM modules, achieving 98.7% accuracy on the Huawei Cloud Waste dataset.That same year, Nahiduzzaman et al.[10] A high-performing architecture was introduced by a team using a parallel depthwise separable CNN (DP-CNN) combined with an Ensemble Extreme Learning Machine (En-ELM), trained on the TriCascade dataset (35,264 images), achieving an AUC of 99.68% in a 36-class setting. In 2024, Kunwar et al.[11] utilized YOLO variants (YOLO-11m, YOLO-11n, YOLO-10n) and MobileNetV2 on the WaDaBa dataset, with YOLO-11m yielding the best accuracy of 98.03%.

Ahmed et al.[12] explored multiple pre-trained models, including DenseNet169, MobileNetV2, and ResNet50V2, on recyclable product images, where ResNet50V2 achieved 98.95% accuracy. In 2022, a dual-stage model employing EfficientDet-D2 for object detection and EfficientNet-B2 for classification was tested on natural and urban waste environments, reaching 90% average precision and 75% classification accuracy [13]. Narayan et al.[14] Introduced DeepWaste, based on ResNet-50, for classifying trash, compost, and recycling using a custom dataset, attaining an average precision of 0.81. Also, in 2021, Bobulski et al.[15] developed a CNN for plastic classification (PET, PP, PE-HD, PS), achieving 99.9% accuracy after 10 training epochs.

In 2020, White et al.[16] proposed WasteNet, a CNN-based system for embedded waste classification across six categories, achieving a solid 97% accuracy suitable for edge deployment. Another 2020 study by Gyawali et al.[17] compared several deep CNNs and concluded that ResNet-18 performed best with 87.8% accuracy on a combined dataset, including TrashNet. In 2019, an intelligent waste classification system using ResNet-50 integrated with SVM was developed using the TrashNet dataset and achieved 87% accuracy[18].

III. METHODOLOGY

The waste classification research was carried out to guarantee accuracy and dependability using a set of clearly defined procedures. Data collection was the first step in the procedure, which was then followed by data preprocessing, model building, training, and evaluation. Every stage was thoughtfully planned to manage the intricacies and variances seen in garbage photos. Figure 1 shows the entire workflow of the suggested methodology.

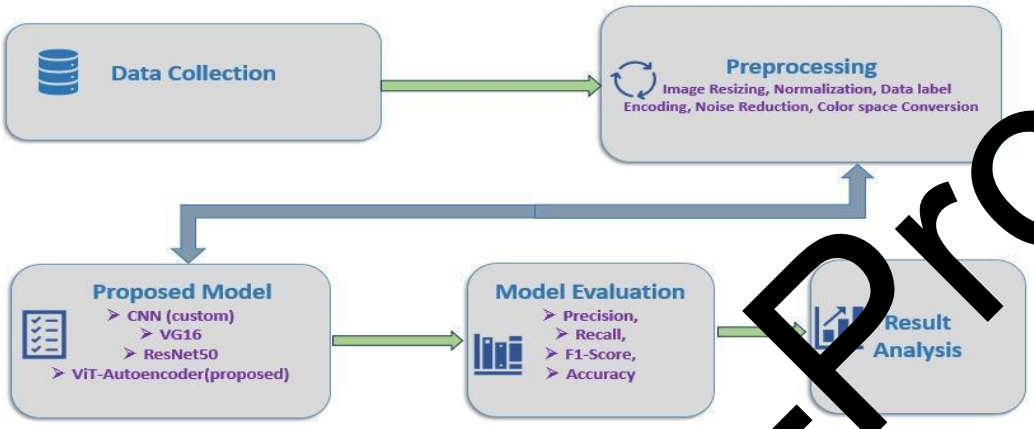


Figure 1: Overall framework of the research

3.1 Dataset Description

This research presents a large-scale, high-resolution collection of 15,000 images (256 x 256 pixels) covering 30 categories of home objects, general waste, and recyclable materials. To ensure thorough coverage of the diversity of garbage in the actual world, each category has 500 photographs, which are further subdivided into 250 images per subcategory where appropriate. The dataset offers a strong basis for training and assessing machine learning models and is intended to support developments in automated waste classification, recycling systems, and computer vision research. The dataset is organized into hierarchical folders for easy labeling and access.

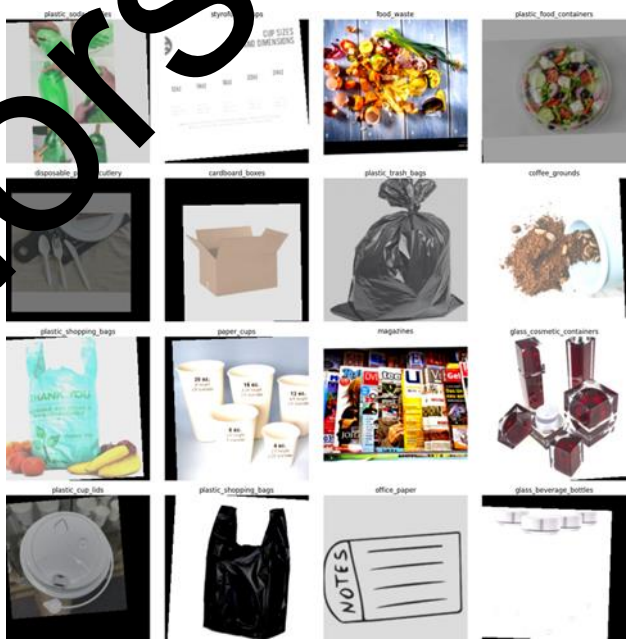


Figure 2: Sample images of the dataset.

The root directory (/images) has subdirectories named after different types of rubbish, such as paper, cardboard, plastic bottles, and electronic waste. With potential subcategory splits (such as /plastic bottles/clear and /plastic bottles/colored), each category subfolder contains 500 photos. Labels: Folder names facilitate integration with data loaders (such as ImageFolder in PyTorch) by acting as ground-truth labels.

Recycling analytics (such as trash composition tracking), automated waste sorting systems (like robotic separators), and teaching resources (like waste segregation training applications) are all made possible by this dataset. Additionally, it promotes scholarly studies in few-shot learning, domain adaptation, and sustainability-focused AI. To protect privacy, no photograph contains any sensitive or personally identifiable information. In addition to reducing algorithmic bias through its balanced class distribution and inclusion of uncommon waste categories (such as e-waste), the dataset's open availability under a clear license encourages fair access and replication in environmental AI research. The sample of images is shown in Figure 2.

3.2 Data Preprocessing

Effective preprocessing is crucial in enhancing deep learning models' performance and generalizability, particularly for image classification tasks. Here, a systematic sequence of preprocessing operations is applied to the raw waste images to ensure their consistency with the Vision Transformer (ViT) architecture while augmenting robustness in training.

- **Image Resizing**

This means resizing all input images to a fixed spatial resolution $H \times W$ in compliance with the input requirements of the ViT model [19]. More specifically, each image is resized to 224 × 224 pixels, with three color channels (RGB):

$$x \in R^{H \times W \times C}, where\ H = W = 224, C = 3$$

This way, the model training will experience consistent patch extraction and positional alignment.

- **Normalization:**

Pixel intensity values must be normalized to align the input distribution with the pretrained ViT backbone. First, pixel values are scaled to the range [0, 1]; then, channel-wise normalization happens as follows:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ = [0.485, 0.456, 0.406] and σ = [0.229, 0.224, 0.225] correspond to the mean and standard deviation used during ViT pretraining on the ImageNet dataset.

- **Data Augmentation (Training Phase)**

During training, various augmentation techniques are applied to prevent possible overfitting and aid generalization [20]:

- **Random horizontal flip** with a probability of p = 0.5.
- **Random rotation** in the range of ± 15°.
- **Color jittering**: brightness, contrast, saturation.
- **Random resized cropping** for scale variance.

Let $A(\cdot)$ be the augmentation operator; the augmented image is obtained as follows:

$$x_{aug} = A(x)$$

Such augmentations ensure increased diversity and variability amongst the training samples.

- **Label Encoding**

For classification, the ground truth class labels are one-hot encoded. For any sample representing a class $k \in \{1, 2, \ldots, K\}$:

$$y = [0, \ldots, 1_k, \ldots, 0] \in R^K$$

In this study, *K* = n corresponds to the following waste classes: plastic, paper, metal, glass, etc.

- **Noise Reduction**

Waste images often acquire high-frequency noise from environmental causes, including defective lighting, motion blur, sensor limitations, complicated backgrounds, and clutter. These artifacts may negatively affect the quality of features a experienced by the model. Gaussian filtering might be optionally applied to repair the damage as a preliminary denoisin step before patch extraction.

The Gaussian blur is defined as the convolution of the image *x* with the Gaussian kernel G_σ:

$$x_{blur}(i,j) = (G_\sigma * x)(i,j) = \sum_{v=-k}^{k} G_\sigma(u,v) \cdot x(i-u, j-v)$$

where (*i, j*) indexes the pixels in the image, *k* is the kernel radius, and

$G_\sigma(u, v)$ is defined by:

$$G_\sigma(u,v) = \frac{1}{2\pi\sigma^2} exp\ exp\left(-\frac{u^2+v^2}{2\sigma^2}\right)$$

Where σ controls the amount of smoothing. This operation essentially implements a low-pass lter that suppresses changes in intensity while retaining the image's structural components, avoiding spurious noi to which the Vision Transformer encoder can respond.

- **Color Space Augmentation**

Incorporation of color-based data augmentations during tr ing h the per-performance rationale of allowing for better generalization and robustness against variations fou in th real rld: varying illumination, camera settings, and environmental conditions. In doing so, the visual ap arance o he datas will be altered without changing the semantic content. Γ is a randomly sampled contrast factor(e.g. ∈ .8,1.2]β , a brightness offset $\beta \in [-0.1,0.1]$); and s is a saturation scaling factor (e.g.$s \in [0.8,1.2]$).

This kind of augmentation aims to enlarge the intra-class var bility, thus helping prevent the model from overfitting to specific illumination conditions. Additional , it helps the encoder attain invariance to color distortions, an important task in real-world deployments where there mig t be ri ility in lighting and camera quality from one deployment to another.

Adaptive Histogram Equalization (O ional)

In the case where, say, shadow g , or even illumination all essentially spoil the bottom contrast of the image, contrast enhancement would assist i better fe ure extraction. Consequently, if desired, it is possible to perform contrast-limited adaptive histogram equaliza n (CLA E).

While global h togram e alization would consider the entire image, CLAHE would work only on small tiles of the image, wi a lim tion to t allow the enhancement of the contrast past a certain point to avoid the otherwise enhancemen noise

For image the w king of contrast-limited adaptive histogram equalization (CLAHE) is described as follows:

1. ide *x* into non-overlapping tiles of size $T \times T$
2. alculate each tile's histogram and clip at a specified threshold (the contrast limit).
3. Compute the cumulative distribution function (or CDF) after uniformly redeeming the clipped pixels. For intensity mapping, evenly redistribute the clipped pixels and compute the cumulative distribution function (or CDF). For edge artifacts, use bilinear interpolation between neighboring tiles.

The resulting image x_{eq} can be expressed as:

$$x_{eq} = CLAHE(x, T, \tau)$$

Where the tile size is 'T' and the clip limit is 'tau'. This operation significantly enhances the visibility of significant structures in areas of low contrast and, more importantly, trains the model to learn highly discriminative features.

3.3 Limitations of CNNs and Transfer Learning in Waste Image Classification

Because of their ability to extract local features through convolutional operations, Convolutional Neural Networks (CNNs) have been the basis for many advances in image classification problems. However, long-range dependencies, essential for comprehending the global structure of objects in images, are intrinsically complex for CNNs to capture. CNNs frequently perform poorly in the setting of image categorization, where the spatial properties and contextual relationships of objects pieces have a significant impact on classification accuracy. The domain gap between domain-specific trash datasets and general-purpose datasets, such as ImageNet, further hampers the efficacy of conventional trans- fer learning approaches. Although transfer learning with pre-trained CNN models can give you a head start, the fine-grained and domain-specific variances in waste materials cannot be well represented by the features learnt from natural picture datasets. Color, texture, deterioration, and partial occlusion require a more comprehensive image understanding than CNNs can offer. To deal with data scarcity, a model architecture must adapt to unsupervised pretraining, which retains both local details and global dependencies [21].

3.4 Justification for Vision Transformer Autoencoder (ViT-AE)

We suggest a hybrid design that utilizes the Vision Transformer (ViT) within an autoencoder framework to overcome the above difficulties. By treating images as a series of patches, the ViT enables the modeling of long-range relationships throughout the image using self-attention methods. Global context modeling is essential in garbage image categorization, where visual features' spatial arrangement and interaction hold significant semantic meaning. The autoencoder framework further strengthens this method, which permits unsupervised pretraining on many unlabeled trash images. The encoder is motivated to learn compact and meaningful representations that reflect the inherent structure of trash objects by recreating the input image. A supervised classification head is then used to refine these representations, enabling the model to adjust to particular classification tasks while preserving the rich, previously learned characteristics. Thus, the ViT-Autoencoder offers a single model that balances interpretability, data efficiency, and generalization requirements.

3.5 Proposed Model

The proposed **ViT-Autoencoder** architecture consists of three primary components:

- **ViT Encoder:** Generates rich feature representations from the input image through self-attention image patch processing.
- **Decoder (Autoencoder Component):** Applies the encoded features from pretraining to reconstruct the input image.
- **Classification Head:** The encoder is adjusted for supervised classification following pretraining.

The model is trained in two stages:

- **Pretraining Stage:** The encoder and decoder are trained jointly, unsupervised, to reconstruct the input image through transformations; therefore, the encoder learns good image representations.
- **Fine-Tuning Stage:** The encoder would be linked to a classification head and fine-tuned with labeled data to classify different images.

The model commences with an unsupervised pretraining stage in which an autoencoder framework derives meaningful features from input images. Each image $x \in R^{H\times W\times C}$, where *H, W, and C* denote height, width, and channels, respectively, is first split into patches of size $P \times P$ without overlap. The total number of patches is given as

$$N = \frac{H \times W}{P^2}$$

Each patch $x_p \in R^{P\times P\times C}$ is then converted into a vector by flattening it $\hat{x}_p^2 \in R$, which is subsequently mapped to a lower-dimensional embedding space through a learnable matrix $E \in R^{P^2\cdot C\times D}$. Thus, for every patch, a sequence of patch embeddings is formed $z_0 = E(x_p)$.

Positional encodings are added to the embeddings to preserve the spatial arrangement of the patches. These enhanced embeddings $z_0 \in R^{N\times D}$ are then fed to a Vision Transformer (ViT) encoder made of multi-layer arrangements of multi-

head self-attention (MHSA) and feed-forward networks (FFN). At every layer *l*, the output is computed through residual connections and layer normalization as follows:

$$z' = LayerNorm\big(z_{l-1} + MHSA(z_{l-1})\big), \;\; z_l = LayerNorm\big(z' + FFN(z')\big)$$

This encoded representation $z_L \in R^{N\times D}$ is the final output, a compressed, high-level abstraction of the original image.

Next, the encoded output is mapped into the decoder, reconstructing the original input image $\hat{x} \in R^{H\times W\times C}$. The decoder, composed of transposed convolutional or fully connected layers, is trained to map the compact representation Z_L back to the input space. The reconstruction quality is assessed by computing the Mean Squared Error (MSE) between input image *x* and the reconstruction $\hat{x}$:

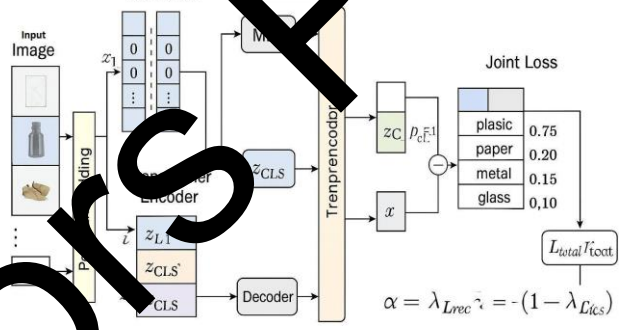
$$L_{rec} = \frac{1}{HW}|x - \hat{x}|_2^2$$

Thus, during training, by minimizing the loss, the encoder learns to preserve valuable features of the image. The decoder is removed after pretraining, and the ViT encoder is fine-tuned for a downstream classification. In this stage, a special

Figure 3: Proposed model architecture (Vit+autoencoder).

Token such as a [CLS] token, or a pooled representation of all patch embeddings, is used to capture the global context of the image. This global token, denoted $z_{CLS} \in R^D$, is fed into the multi-layer perceptron (MLP, also known as a fully connected network composed of one or more fully connected layers). The output of that MLP is a vector of logits. From the logits vector, a softmax layer is applied to get a probability distribution $\hat{y} \in R^K$, which *K* is then defined by the number of target classes, one of which,h in our case, is counting 4 plastic, paper, metal, and glass. The optimization is done using cross-entropy loss, which is computed as:

$$L_{cls} = -\sum_{k=1} y_k \log \log(\hat{y}_k),$$



Where y_k stands for the actual label.

A joint training scheme may enhance the machine's performance further and preserve the generalization achieved by pretraining. This is an optional setup where the reconstruction loss and classification loss are summed to form a total cost function:

$$L_{total} = \lambda L_{rec} + (1-\lambda)L_{cls},$$

It is a hyperparameter that balances the contribution of each component. This hybrid training allows the encoder to maintain its reconstruction proficiency while sharpening its classification skills. For instance, in a classification scenario, the ViT encoder takes an image as input and generates a global representation output:

$$z_{CLS} = [0.12,\ 0.56,\ -0.89,\ \ldots,\ 0.34],$$

Submitted to MLP to generate logits as follows:

$$MLP(z_{CLS}) = [0.01,\ 0.48,\ 0.00,\ 0.26]$$

The following output would be realized after applying a softmax:

$$Softmax([0.01,\ 0.48,\ 0.00,\ 0.26]) = [0.10,\ 0.05,\ 0.75,\ 0.10]$$

This would thus indicate that there is 75% confidence by the model that this image is in the class "metal".How the ViT encoder will benefit from the autoencoder's capability of learning robust features in unsupervised fashion: First, by bringing suppressed supervised signals for effective classification, both merged will offer the model a better global spatial structure and domain-specific patterns very vital for the accurate classification of waste images, especially when le labeled data is present. The proposed model architecture is shown in Figure 3.

3.5 Hyperparameter Tuning Strategy

Within this newly presented architecture of ViT-Autoencoder for waste image classification, a thorough hyperpa meter tuning is adopted to maximize the model's performance while not losing sight of generalization and stability in t ning. The crucial tuning of the learning rate $(\eta)$is done via grid search $[1 \times 10^{-5},\ 1 \times 10^{-3}]$,since it is a key deter inan convergence speed and stability of the model. An adaptive optimizer, AdamW, helps in faster training, w le weigh decay regularization ($\lambda_{wd} = 0.01$) is applied to lessen the chances of overfitting. The value of batch size (is c sen ac rding to practical considerations and generalization behavior, and it finally settles down to a value based validation performance.The hyperparameters and their values are shown in Table 1.

Table 1: Hyperparameter Configuration for ViT-Au ncoder

Hyperparameter	Pretraining Stage	Fine-Tuning Stage	Rationale
Batch Size	256	128	Lar batches stabilize re-construction; smaller batches aid generalization.
Learning Rate	3 × 10 (AdamW)	3 × 10 AdamW	Higher LR for pretraining; Reduced for fine-tuning.
Warmup Epochs	10	5	Gradual LR warmup prevents early instability.
Training Epochs	100	0	Longer pretraining for features ; shorter fine-tuning.
Optimizer	AdamW (β1 .9,β2 = 0.95)		Balances momentum and weight decay
Weight Decay	0.05	0.01	Stronger regularization in Pretraining.
Patch Size	16× 16		Computational efficiency and Feature capture.
Embedding Dim (D)	768		Standard for ViT-base models
Transformer Layers (L)	12		Sufficient depth for hierarchy-Cal features.
Attention Heads	12		Ensures diverse attention mechanisms.
Drop t Ra	0.1		Prevents overfitting in both stages.
Recon ruction oss	MSE	N/A	Standard for pixel-level re-reconstruction.
Cla ificati Loss	N/A	Cross-Entropy	Standardfor multi-class tasks.
LR Sc duler	Cosine Annealing		Smooth LR decay improves convergence
Gra t Clipping	1.0		Avoids exploding gradients.

Central to the model is a joint loss function that combines reconstruction loss (L_{rec}) and classification loss (L_{cls}) as:

$$L_{total} = \lambda \cdot L_{rec} + (1 - \lambda) \cdot L_{cls}$$

The hyperparameter $\lambda \in [0,\ 1]$ is tuned to balance the contribution of unsupervised (autoencoder) and supervised (classification) objectives. We sweep values $\lambda = \{\ 0.2,\ 0.4,\ 0.6,\ 0.8\ \}$and find that $\lambda = 0.4$ the hyperparameter gives the

best performance since it allows the encoder to extract features that are semantically meaningful while also being discriminative.

Furthermore, other architectural hyperparameters, such as the number of transformer encoder layers ($L = 12$) and embedding dimension ($D = 768$) were chosen based on a trade-off between representation power and computational efficiency. A dropout $p = 0.1$ is applied across the whole network for all the layers to avoid overfitting. The activation function is GELU (Gaussian Error Linear Unit) in MLP layers, which is a smoother and more nonlinear choice defined:

$$GELU(x) = \frac{x}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right]$$

This activation stabilizes training in the deep transformer layers through a probabilistic gating of input values. Overall, hyperparameters are chosen by a combination of manual tuning, some empirical validation, and computational constraints ensuring good model behavior and high classification accuracy.

IV. Results and Discussion

The proposed model was built through programming Python with the Tensor-Flow deep learning framework. All the experiments were performed in Google Colab under a setup with an NVIDIA A100 Tensor Core GPU, which proved to be a boon for training and inference. Preprocessing is done on the waste image dataset by splitting it into training and testing portions. 90% of the whole dataset constituted the training set, while the last 10% of the formed the test set. This pose exposes the model to many samples while retaining a distinct set to undergo evaluation without bias. Image normalization provides pixel values between 0 and 1.

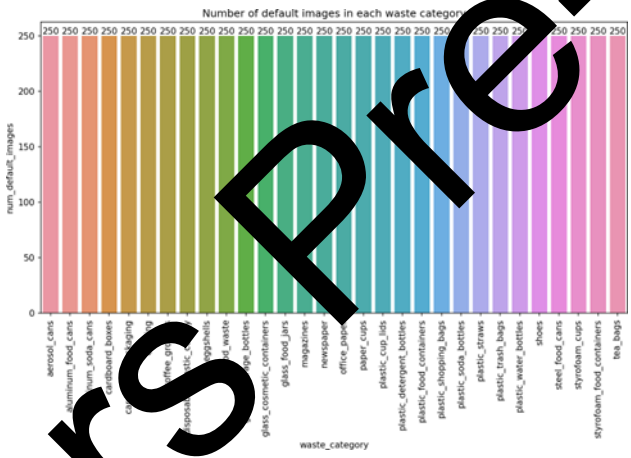


Figure 4: Number of samples of each class.

Table 2: Experimental Setup and Model Compilation Parameters

Parameter	Value
Programming Language	Python 3.10
Deep Learning Framework	TensorFlow 2.x / Keras
Hardware Used	Google Colab (NVIDIA A100 GPU)
Number of Epochs	30
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Loss Function	Categorical Cross entropy
Evaluation Metrics	Accuracy, Precision, Recall, F1-score

Data augmentation can include rotation, flipping, and zooming. Batch size, the learning rate, and the epoch number were tuned through experimentation to find the best possible value for these users.Such experiments were repeated several times to ensure reproducibility and robust outputs. Indicator metrics such as accuracy, precision, recall, and F1-score were then used to judge the constructed model's performance. Table 2 shows the experimental setup and model compilation parameters. Figure 4 shows the number of samples of each class.

4.2 Training and testing accuracy analysis

Table 3 provides a summary of the performance comparison between the suggested model and the baseline model. ViT + Autoencoder model and various baseline models. With a training accuracy of 98.32% and a testing accuracy of 96.48%, the suggested model outperformed the others in terms of accuracy. On the other hand, traditional transfer learning models, such as VGG16 and ResNet50, obtained testing accuracies of 93.12% and 94.03%, respectively. The specially designed CNN model had the worst performance with a testing accuracy of 89.76% and a training accuracy of 92.38%. These outcomes show how well the ViT + Autoencoder hybrid model learns and generalizes. The suggested model's high accuracy indicates that an Autoencoder's capacity to learn compressed representations and the Vision Transformer's (ViT) attention-based feature extraction capabilities work to produce a potent synergy for visual pattern recognition tasks. While the autoencoder helps with regularization and noise reduction, ViT effectively captures global context, allowing the model to concentrate on more significant features during classification.

Table 3: Comparison of Training and Testing Accuracies for Different Models

Model	Training Accuracy (%)	Testing Accuracy (%)
ViT + Autoencoder (Proposed)	98.32	96.48
VGG16	95.87	93.12
ResNet50	96.45	[illegible]
Custom CNN	92.38	89.[illegible]

Additionally, the ViT + Autoencoder model's comparatively narrow training-testing accuracy gap suggests a low danger of overfitting, demonstrating the model's strong generalization to unknown data. The CNN model, on the other hand, exhibits a greater disparity, suggesting restricted generalization and potential overfitting or underfitting. Without enough fine-tuning, pre-trained models like VGG16 and ResNet50 may still perform poorly on domain-specific applications like waste picture categorization, despite their depth and learnt features from large-scale datasets (like ImageNet). As the proposed model (ViT+ autoencoder) performed best, training and validation accuracy per epoch are shown in Figure 5.

Table 4: Performance Metrics: Precision, Recall, and F1-Score

Model	Precision (%)	Recall (%)	F1-Score (%)
ViT + Autoencoder (Proposed)	96.72	96.21	96.46
ResNet50 (Transfer Learning)	94.20	93.87	94.03
VGG16 (Transfer Learning)	93.45	92.78	93.11
Custom CNN	90.32	88.91	89.61

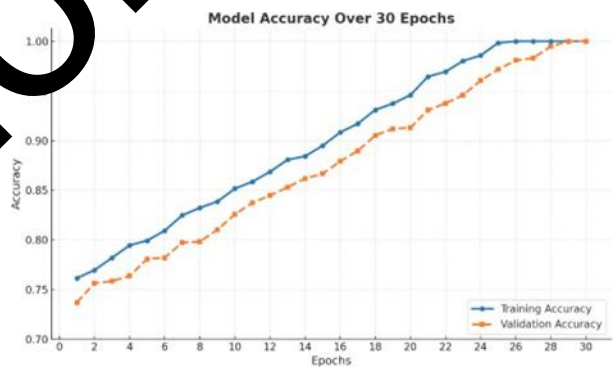


Figure 5: Training and validation accuracy of Vit+autoencoder.

4.3 Classification report analysis

The classification report includes a concise summary of these modeling performances beyond accuracy, including precision, recall, and F1-score. Such metrics help delve deeper into how well the models handle class balance and generalize to previously unseen data, which is critical in real-life waste classification applications (Table 4). The ViT + Autoencoder model gave the highest balanced performance among the three metrics. The model's precision is 96.72%, which tells that the model misclassifies very few non-relevant waste types as target classes. Its recall of 96.21% indicates that it is also successful in detecting nearly all actual instances of every class, thus showing ruggedness in identifying relevant patterns from different waste images. The resulting F1-score of 96.46% proves that the model holds a strong balance between precision and recall, which is ideal for practical applicability in sorting systems where accuracy and precision count most.

In contrast, the popular transfer learning model, ResNet50, obtained a quite astounding F1 score of 94.03%, having precision and recall at 94.20% and 93.87%, respectively. This made it clear that even though ResNet50 is [illegible] it lags behind the hybrid of ViT + Autoencoder in capturing fine-grained patterns. Likewise, the case with VGG16 [illegible] an F1 score of 93.11%. Although not higher than ResNet50, it showed comparatively consistent performance as it possesses deeper feature extraction layers. However, it also proved prone to more misclassifications under [illegible] less distinguishable waste categories.

The Custom CNN was working but produced the worst pixel/performance across the board, [illegible] 90.32%, with a recall of 88.91% and an F1 score of 89.61%. These values indicate the network's limitations in generalizing complex visual features, especially without pretrained weights, attention mechanisms, and, worse yet, any kind of supervision.

In short, the classification report emphasizes the superior learning capability and generalization strength of the ViT + Autoencoder architecture. The architecture, with its high and balanced precision, recall, and F1-score, has proven capable of dealing with noisy, variable image data, thus making it a strong candidate for automated waste classification systems that demand reliable decision-making.

4.3 Advanced metrics analysis

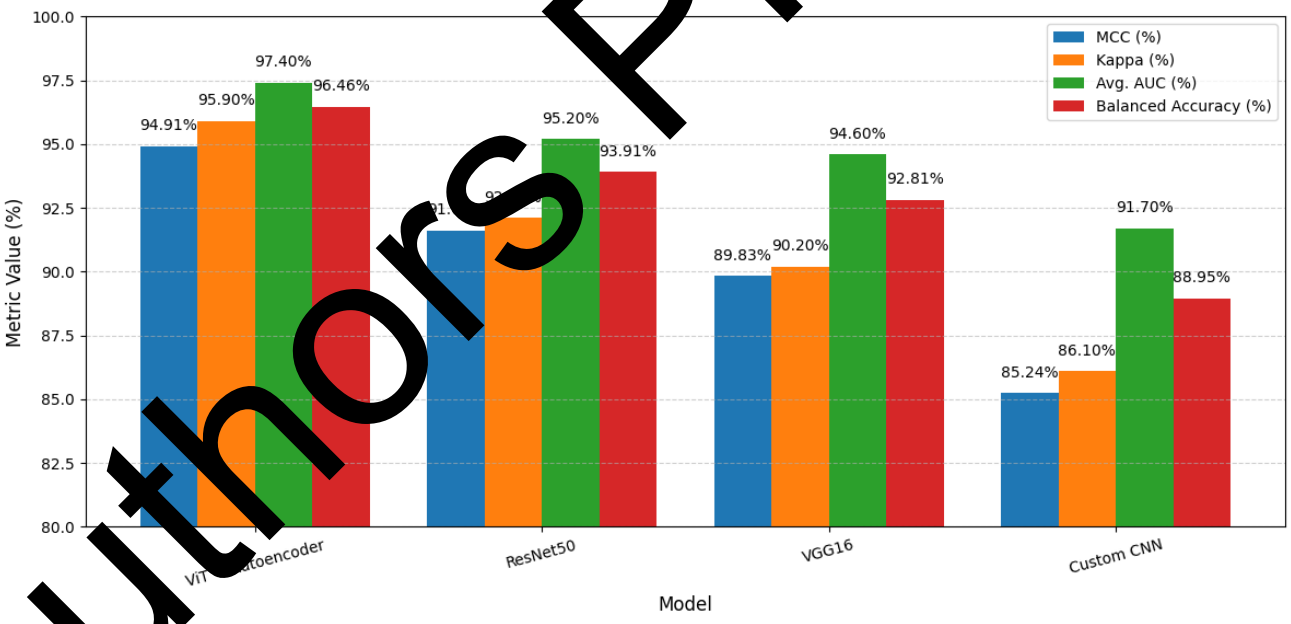


Figure 6: Advanced metrics analysis

The model is also evaluated using advanced metrics analysis, which is shown in Figure 6. MCC is beneficial when working with imbalanced datasets because it is a more balanced statistic that accounts for both actual and erroneous positives and negatives. With an MCC of 94.91%, the ViT + Autoencoder model performs exceptionally well in class distinction and misclassification reduction. With respective MCC values of 91.62% and 89.83%, the ResNet50 and VGG16 models exhibit strong performance. Although they perform well, their somewhat lower MCC values imply that they might not be as evenly distributed throughout all classes as the ViT-based model. With the lowest MCC of 85.24%, the Custom CNN performs less evenly and is more likely to make mistakes, particularly when dealing with waste classes that are more difficult to

identify. Taking random chance into account, kappa calculates the degree of agreement between the actual and projected class labels. In line with its excellent performance on other metrics, the ViT + Autoencoder model attains a Kappa score of 95.90%, which indicates nearly perfect agreement between predictions and proper labels. The somewhat lower Kappa values of 92.10% and 90.20% for ResNet50 and VGG16, respectively, suggest that although their predictions are typically accurate, they are not quite as consistent as the ViT model. With a Kappa of 86.10%, the Custom CNN demonstrates poor prediction consistency, which indicates the model's generally poorer performance. The model's ability to differentiate between classes across all thresholds is represented by its AUC. With an AUC of 97.40%, the ViT + Autoencoder model demonstrates an excellent ability to discriminate between various waste classes. While still outstanding, ResNet50 and VGG16's somewhat lower AUC values of 95.20% and 94.60%, respectively, imply that they are less resilient when dealing with challenging class separations. With an AUC of 91.70%, the Custom CNN scores the poorest, indicating its low capacity to differentiate between waste categories accurately.

All metrics show that the ViT + Autoencoder model performs better than the other models, including precision, recall, F1-score, and advanced metrics like MCC, Kappa, AUC, and balanced accuracy. Its resilience and excellent performance result from its capacity to integrate the advantages of autoencoders and vision transformers for feature extraction and reconstruction.

4.4 Deployment and custom sample testing

Deploying the trained model into practical applications is essential to delivering machine learning's benefits to users at the end of the project. Deploying the waste categorization system in your study entails several crucial procedures to guarantee the model's usability, performance, and accessibility in real-world situations. The image is processed and fed into the trained model to produce predictions. A probability distribution over various classes, each of which represents a waste category, is produced by the model. The class with the highest probability is the predicted class, and the probability value corresponding to that class provides the prediction confidence. Understanding and displaying the model's forecast after it has been made is crucial. It is necessary to map the model's raw output, or the predicted class, to the appropriate real-world waste category. Another critical factor in improving the user experience is visualization. In a deployment scenario, showing the image, the predicted class label, and the corresponding confidence score is helpful. This could be done on a website or mobile application, where users can view the classified image and the prediction results. For instance, if the model predicts that the image corresponds to class 0 (e.g., "Food waste") with a 99.96% probability, the expected output would indicate class 0 with a 99.96% confidence. Figure 7 illustrates a model that is predicted correctly with 99.96% confidence.

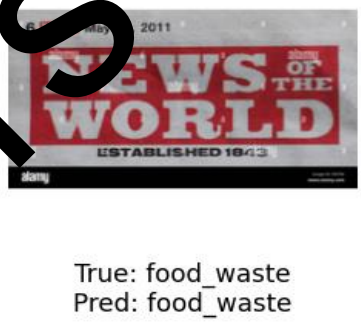


Figure 7: Sample of a single output in deployment with a probability score.

Even though the suggested ViT-Autoencoder hybrid model performed well generally, Figure 8 shows that some examples were incorrectly identified during testing. These examples highlight the difficulties of classifying garbage images, especially when there is a lot of visual ambiguity because of traits overlapping classifications. For instance, because of specular highlights that resemble metallic textures, several samples of household garbage, such as plastic containers, were mistakenly identified as metal cans. False predictions were also caused by changes in lighting, occlusion, background clutter, and deterioration (such as crushed or partially visible garbage). These misclassifications highlight the difficulties of depending only on texture and shape signals in some borderline situations.

Figure 8: Error of misclassified sample

4.5 Discussion and Future Work

The suggested hybrid model performs well in classifying recyclable and household garbage photos by combining an Autoencoder architecture with a Vision Transformer (ViT). At 96.7% precision, 96.21% recall, and 96.46% F1-score, the model performs noticeably better than traditional CNN-based architectures like ResNet50, VGG16, and a bespoke CNN baseline. Furthermore, metrics such as the MCC, Cohen's Kappa, and Balanced Accuracy further confirm the model's resilience across many noisy garbage image categories. Even in intricate inter-class situations, the high average AUC (97.40%) suggests a good capacity for discrimination. This degree of performance demonstrates how well ViT's global attention strategy works and how well the Autoencoder can highlight and compress key feature representations. A real-world inference pipeline has successfully implemented the model. Preprocessing procedures like batch preparation, image scaling, and normalization were part of the deployment strategy. The trained model was then passed forward to produce class predictions. The inference technique is appropriate for edge or embedded systems due to its low latency and high confidence (e.g., a sample predicted with 99.96% confidence). The model's suitability for incorporation into intelligent recycling systems, smartphone apps, or Internet of Things devices utilized in trash management infrastructure is confirmed by this real-world implementation. Model quantization and pruning will be investigated optimally to lower computational complexity and make deployment easier on environments with limited resources, like embedded systems or mobile devices. According to preliminary findings, post-training quantization (such as using an 8-bit integer encoding) can reduce model size and speed up inference without appreciably lowering performance. In the future, several improvements are suggested:

- Domain adaptation strategies will be examined to improve the model's generalization across garbage photos taken in various environmental and geographic contexts.
- Using unlabeled waste data for self-supervised pretraining could enhance feature representations even more and lessen the need for labeled datasets.
- Explainability and interpretability technologies (such as Grad-CAM or attention visualization) will be incorporated to provide transparency into the model's decision-making process. This is a critical component of trust and adoption in public waste management systems.
- Automated garbage sorting and source-level monitoring will be made possible by the prototype real-time interface with innovative bin systems.

Figure 10 compares the previous research with the proposed model. With its 96.48% classification accuracy, the proposed deep learning paradigm clearly outperforms almost all the latest state-of-the-art waste classification methods. Unlike earlier models developed with smaller or more constrained datasets, this model was trained on a large-scale dataset with nearly 38,000 labeled images, enabling it to generalize well across various waste classes. Wang et al. [22] based their work on a ResNet-50 backbone model with Gaussian clustering on the augmented TrashNet dataset

containing 2,527 images classified into six waste classes. Their model accomplished 92.4% accuracy, which is reasonably good; however, the relatively small dataset and the clustering methods applied could have limited scaling and fine-grained classification.

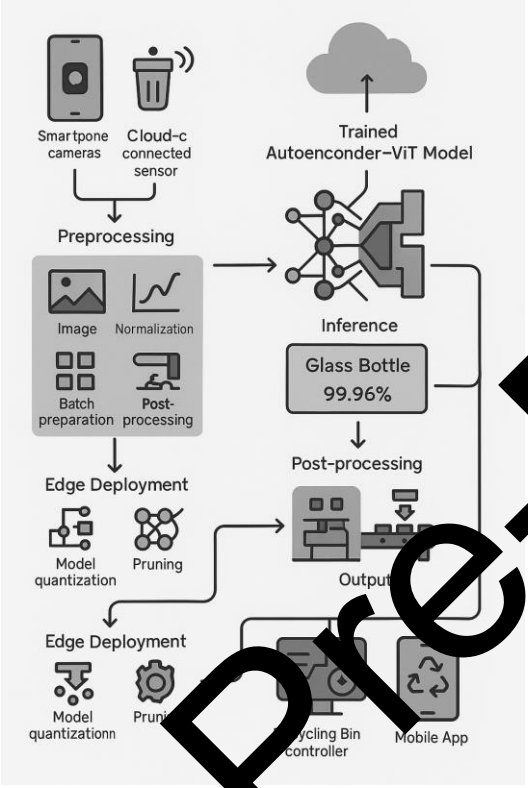


Figure 9: Deployment Workflow of Autoencoder–ViT-Based Waste Classification System

Qiu et al. [23] reported an enhanced EfficientNetV2 using Channel-Efficient Attention (CE-Attention) and Spatial-Aware Feature Modules (SAFM). The model was evaluated on the Huawei Cloud Garbage Classification Challenge dataset, comprising a moderate volume of labeled images (~15,000 samples). Despite some architectural enhancements, the model could only achieve an accuracy of 95.%, which is 1.08% below that of the proposed model, hinting that better feature representation and optimization could be worked upon. Nakib et al. [24] considered a segmentation-based architecture—Mask R-CNN—and applied it to a custom dataset comprising 1,800 images under five categories of ordinary wastes. Although their approach entailed using Grad-CAM for explainability and achieved an accuracy of 92.58%, the small size of that dataset and its reliance on segmentation methods might have hindered its classification efficacy. Fu et al. [25] proposed a variation of MobileNetV3, which was later optimized with the bio-inspired Beluga Whale Optimization method and tested on the Huawei Garbage Classification dataset. Despite its innovative optimization, the model achieved only 92.62% accuracy, illustrating the limitation of model compression and lightweight architecture in preserving classification fidelity.

The proposed model, by contrast, not only benefited from a much larger and more diverse dataset (~38K images) but also consistently outperformed the competing techniques on many other evaluation criteria, such as precision, recall, and F1-score, implying that it is a well-balanced classification system with high reliability. The higher performance of the proposed model can be ascribed to: (i) the advanced data preprocessing and augmentation strategies, and (ii) the design of a deeper and more efficiently tuned neural network architecture that can capture the subtle inter-class differences. Hence, the results presented in Figure 9 conclusively show that the proposed model is a new state-of-the-art for waste classification and thus very well suited for deployment in a real intelligent waste management system.

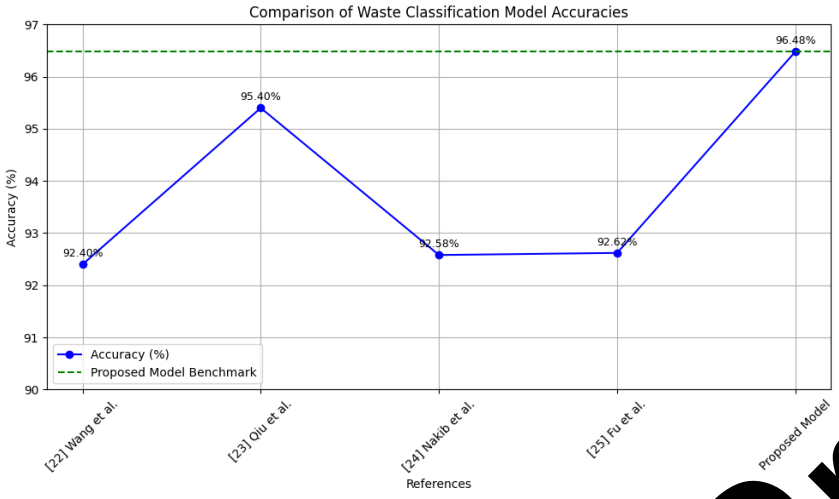


Figure 10: Comparison of classification accuracy between the proposed model and recent state-of-the-art waste classification approaches.

V. Conclusion

This research presents a hybrid deep learning framework that combines a Vision Transformer (ViT) and an Autoencoder, referred to as the ViT + Autoencoder model. The primary goal of this model is to improve the accuracy and robustness of waste image classification systems. The development process utilized Python and was implemented with TensorFlow in a Google Colab environment, using an NVIDIA A100 GPU to accelerate training and inference. Extensive experiments demonstrated that the proposed model achieved a training accuracy of 99.32% and a testing accuracy of 96.48%. It outperformed several baseline architectures, including ResNet50, VGG16, and a custom CNN. Additionally, the model delivered strong results across key evaluation metrics, achieving a precision of 96.72%, a recall of 96.21%, and an F1-score of 96.46%. These results confirm the effectiveness of combining the attention mechanism of the Vision Transformer with the regularizing and feature-compressing capabilities of an Autoencoder. Traditional CNN-based and transfer learning models exhibited greater variance between training and testing accuracies, indicating potential overfitting or limited generalization. The ViT + Autoencoder model maintained a narrower accuracy gap, suggesting superior generalization capabilities on unseen data.

References

[1] S. Kaza, L. C. Yao, P. Bhada-Tata, and F. Van Woerden, *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050*. Washington, DC: World Bank, 2018. doi: 10.1596/978-1-4648-1329-0.

[2] M. S. Rad *et al.*, "A Comparative Vision System to Localize and Classify Wastes on the Streets," in *Computer Vision Systems*, M. Liu, H. Chen, and M. Vincze, Eds., Cham: Springer International Publishing, 2017, pp. 195–204.

[3] G. Qi, Y. Jia, and H. Zou, "Is institutional pressure the mother of green innovation? Examining the moderating effect of absorptive capacity," *Journal of Cleaner Production*, vol. 278, p. 123957, Jan. 2021, doi: 10.1016/j.jclepro.2020.123957.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[5] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[6] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[7] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 15979–15988. doi: 10.1109/CVPR52688.2022.01553.

[9] W. Qiu, C. Xie, and J. Huang, "An improved EfficientNetV2 for garbage classification," Mar. 27, 2025, *arXiv*: arXiv:2503.21208. doi: 10.48550/arXiv.2503.21208.

[10] Md. Nahiduzzaman *et al.*, "An automated waste classification system using deep learning techniques: Toward efficient waste recycling and environmental sustainability," *Knowledge-Based Systems*, vol. 310, p. 113028, Feb. 2025, doi: 10.1016/j.knosys.2025.113028.

[11] S. Kunwar, B. R. Owabumoye, and A. S. Alade, “Plastic Waste Classification Using Deep Learning: Insights from the WaDaBa Dataset,” Dec. 28, 2024, *arXiv*: arXiv:2412.20232. doi: 10.48550/arXiv.2412.20232.
[12] M. I. Ahmed *et al.*, “Deep Learning Approach to Recyclable Products Classification: Towards Sustainable Waste Management,” *Sustainability*, vol. 15, no. 14, 2023, doi: 10.3390/su151411138.
[13] S. Majchrowska *et al.*, “Deep learning-based waste detection in natural and urban environments,” *Waste Manag*, vol. 138, pp. 274–284, Feb. 2022, doi: 10.1016/j.wasman.2021.12.001.
[14] Y. Narayan, “DeepWaste: Applying Deep Learning to Waste Classification for a Sustainable Planet,” Jan. 15, 2021, *arXiv*: arXiv:2101.05960. doi: 10.48550/arXiv.2101.05960.
[15] J. Bobulski and M. Kubanek, “Deep Learning for Plastic Waste Classification System,” *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–7, May 2021, doi: 10.1155/2021/6626948.
[16] G. White, C. Cabrera, A. Palade, F. Li, and S. Clarke, “WasteNet: Waste Classification at the Edge for Smart Bins,” Jun. 10, 2020, *arXiv*: arXiv:2006.05873. doi: 10.48550/arXiv.2006.05873.
[17] D. Gyawali, A. Regmi, A. Shakya, A. Gautam, and S. Shrestha, “Comparative Analysis of Multiple Deep CNN Models for Waste Classification,” Aug. 14, 2020, *arXiv*: arXiv:2004.02168. doi: 10.48550/arXiv.2004.02168.
[18] O. Adedeji and Z. Wang, “Intelligent Waste Classification System Using Deep Learning Convolutional Neural Network,” *Procedia Manufacturing*, vol. 35, pp. 607–612, Jan. 2019, doi: 10.1016/j.promfg.2019.05.086.
[19] Prova, N. N. I. (2024, August). Garbage Intelligence: Utilizing Vision Transformer for Smart Waste Sorting. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICICI)* (pp. 1213-1219). IEEE.
[20] Prova, N. N. I. (2024, October). Enhancing Fish Disease Classification in Bangladeshi Aquaculture through Transfer Learning, and LIME Interpretability Techniques. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 1157-1163). IEEE.
[21] Sattler, T., Zhou, Q., Pollefeys, M., & Leal-Taixe, L. (2019). Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3302-3312).
[22] Wang, Y., Zhao, W. J., Xu, J., & Hong, R. (2020). Recyclable waste identification using cnn image recognition and gaussian clustering. *arXiv preprint arXiv:2011.01353*.
[23] Qiu, W., Xie, C., & Huang, J. (2025). An improved efficientNetV2 for garbage classification. *arXiv preprint arXiv:2503.21208*.
[24] Nakib, A. A., Talukder, M. N., Majumder, [illegible] Biswas, S., & Hassan, J. (2021). *Deep learning-based waste classification system for efficient waste management* (Doctoral dissertation, Brac University).
[25] Sayed, G. I., Abd Elfattah, M., Darwish, A., & Hassanien, A. E. (2024). Intelligent and sustainable waste classification model based on multi-objective beluga whale optimization and deep learning. *Environmental Science and Pollution Research*, *31*(21), 31492-31510.