Multi-label Waste Classification using Super Resolution and Image Segmentation

A Project Report Submitted
in Partial Fulfilment of the Requirements
for the Degree of

Bachelor of Technology(Honours)

in

Computer Science and Engineering

by

Emmanuel George P
(Roll No. 2022BCS0104)

to

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
KOTTAYAM-686635, INDIA

April 2025

DECLARATION

I, **Emmanuel George P** (**Roll No: 2022BCS0104**), hereby declare that, this report entitled **"Multi-label Waste Classification using Super Resolution and Image Segmentation"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science and Engineering** is an original work carried out by me under the supervision of **Dr. Cinu C Kiliroor** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam-686635 **Emmanuel George P**

April 2025 **2022BCS0104**

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Multi-label Waste Classification using Super Resolution and Image Segmentation"** submitted by **Emmanuel George P** (**Roll No: 2022 BCS0104**) to the Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **Computer Science and Engineering** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635 — Dr. Cinu C Kiliroor

April 2025 — Project Supervisor

ABSTRACT

This work presents a multi-label waste classification system combining super-resolution and image segmentation. Super-resolution enhances image clarity, aiding in detailed feature extraction, while segmentation isolates waste objects to reduce background interference. A Deep Learning model is trained to classify multiple waste types within a single image. The integrated approach improves accuracy, especially in complex, mixed-waste scenarios, making it suitable for smart waste sorting and automated recycling applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Effective identification and segregation of waste are crucial for sustainable waste management. With the rise in population, urbanization, and consumption, the volume and complexity of waste have grown significantly. Proper segregation not only facilitates efficient recycling but also reduces environmental pollution, conserves resources, and supports a circular economy. Manual sorting, while still widely used, is often inefficient, inconsistent, and costly, highlighting the need for intelligent, automated waste classification systems that can operate at scale.

Traditional single-label classification systems assign only one label to each image, which is often insufficient in practical scenarios where an image may contain multiple waste items belonging to different categories. Multi-label classification solves this problem by enabling the model to detect and classify multiple waste types within a single image. For instance, if an image contains both a plastic bottle and a paper wrapper, a multi-label model can accurately assign both "plastic" and "paper" labels. This approach better reflects the real-world nature of mixed waste and significantly enhances the reliability of automated classification systems.

To achieve this, our project leverages advanced techniques such as image super-resolution and image segmentation. Super-resolution is used to enhance low-quality or blurry images, ensuring that small or unclear waste components are more visible and identifiable. Image segmentation allows us to detect and isolate individual waste objects within an image, enabling more precise labeling. Combined with multi-label classification, these techniques form a powerful framework for accurately identifying and categorizing different types of waste within complex, cluttered scenes. This multi-step pipeline aims to improve the overall performance of automated waste segregation systems and contribute to smarter, cleaner waste management solutions.

Chapter 2

Literature Review

As part of this study, we conducted a comprehensive review of existing literature across the domains of super-resolution, image segmentation, and waste classification. In the super-resolution space, we examined models such as DRCT and HMANet which are widely used for enhancing image quality and preserving fine details crucial for object recognition. For image segmentation, we explored both classical and deep learning-based approaches, including YOLO and Mask R-CNN, all of which have shown strong performance in accurately isolating objects within cluttered scenes. On the classification front, we investigated various architectures tailored for classification tasks, including adaptations of ResNet, Inception, and MobileNet. These models provided valuable insights into the strengths and trade-offs of different approaches, helping us design an integrated pipeline that combines super-resolution, segmentation, and multi-label classification for efficient and accurate waste identification.

2.1 Super Resolution

2.1.1 Hybrid Multi-Axis Aggregation Network

- Overcomes the limited receptive field in Transformers through a hybrid model that integrates both local and global attention mechanisms[1].
- Incorporates Residual Hybrid Transformer Blocks (RHTBs) and Grid Attention Blocks (GABs) to enhance spatial dependencies and capture hierarchical feature similarities effectively.
- Demonstrates superior performance over state-of-the-art models such as SwinIR and HAT on standard benchmarks, while utilizing fewer computational resources.
- Achieves efficiency and performance gains through a customized pre-training strategy tailored to the model architecture.

2.1.2 Dense Residual Connection Transformer

- Identifies spatial information loss in deeper network layers and introduces the Dense-Residual Connection Transformer (DRCT) to stabilize feature flow and prevent information bottlenecks.
- DRCT combines dense-residual connections with Swin Transformer layers to enhance receptive fields and retain high-frequency spatial information[4].
- Achieves state-of-the-art performance on benchmark datasets, surpassing more complex models while using fewer parameters and requiring

lower computational resources.

2.2 Image Segmentation

2.2.1 Using YOLO for Instance Segmentation

- The paper proposes Insta-YOLO, a novel one-stage end-to-end deep learning model for real-time instance segmentation.
- Instead of pixel-wise prediction, their model predicts instances as object contours represented by 2D points in Cartesian space[8].
- The Insta-YOLO model achieves competitive accuracy on the Carvana and Cityscapes datasets, with results showing improvements over baseline methods and comparable performance to state-of-the-art models, but at a faster speed (e.g., 2.4 times the speed of YOLACT).

2.2.2 Mask R-CNN for Image Segmentation

- The paper seeks to develop a simple, flexible, and fast framework for instance segmentation, combining the tasks of object detection and semantic segmentation[2].
- Mask R-CNN extends Faster R-CNN by adding a branch that predicts object masks in parallel with bounding box recognition. It also introduces RoIAlign to improve mask accuracy by addressing misalignment issues in RoIPool.

- Mask R-CNN achieves state-of-the-art results on the COCO dataset and outperforms previous single-model entries and runs efficiently.

2.3 Waste Classification

2.3.1 ECCDN-Net model for Waste Classification

- The study was conducted on a dataset containing 25,000 waste images, categorized into two classes: organic and plastic.
- The proposed model, named ECCDN-Net, was a hybrid architecture combining DenseNet201 and ResNet18.
- The ECCDN-Net model achieved a high accuracy of 96% on the binary classification task[5].
- The study concluded that future work should extend the model to multi-class classification in order to recognize a broader range of waste categories such as plastic, metal, paper, and more.

2.3.2 Adadelta Optimizer for Waste Classification

- The dataset used consists of 10,000 images, divided into six waste categories[7].
- Classification techniques explored include CNN, SVM, and HOG (Histogram of Oriented Gradients) features.
- The highest classification accuracy of 93% was achieved using a hybrid model combining CNN and HOG features.

- The Adadelta optimizer provided the best performance on this dataset.
- Background clutter in images significantly reduced the model's accuracy, indicating the need for effective segmentation or background removal.

2.3.3 Preprocesssing steps in Garbage Classification

- The study utilized a dataset containing 2,500 images, comprising multiple waste classes such as plastic, glass, and metal.
- Preprocessing steps applied to the dataset included resizing, rotation, and image flipping.
- Classification models used in the study included SVM, CNN, DenseNet, and ResNet.
- The highest classification accuracy of 93% was achieved using the ResNet model[9].
- The study noted a limitation in that the models were unable to detect multiple labels within a single image.

2.3.4 Garbage Classification with Deep Learning Techniques

- The study utilised a dataset containing 2500 images consisting of multiple classes like plastic, glass, metal etc.

- The preprocessing done on the dataset included resize, rotation, flips etc.
- Models utilised were SVM, CNN, DenseNet and ResNet with a top accuracy of 93% for ResNet
- The study explains that the model could not detect multiple labels from the same image

Chapter 3

Proposed Methodology

The proposed methodology for multi-label waste classification integrates super-resolution and image segmentation to enhance image clarity and isolate relevant waste regions. This hybrid approach boosts classification accuracy by refining input data before label prediction. The pipeline consists of data preprocessing, super-resolution enhancement, segmentation and multi-label classification using a deep learning framework. A block diagram explaining this workflow is shown in Figure 3.1.

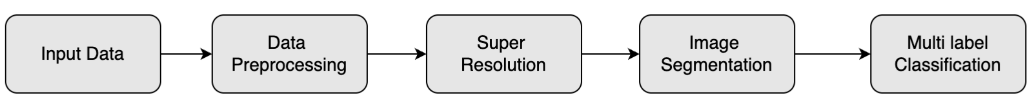


Figure 3.1: Block diagram showing proposed methodology

3.1 Super Resolution

Super-resolution is a technique in image processing that enhances the resolution of an image by reconstructing high-resolution (HR) images from low-resolution (LR) inputs. It uses deep learning models to predict the fine details and textures that are typically lost in low-resolution versions. In classification, super-resolution significantly improves performance by providing models with clearer and more detailed image inputs.

When waste images are blurry or captured in low quality, important features like textures, edges, and object boundaries may be obscured, leading to misclassification. By enhancing these visual details, super-resolution allows the model to extract more accurate and distinctive features, which improves the recognition and labeling of multiple waste categories.

3.1.1 Evaluation Metrics for Super Resolution

The evaluation metrics used to evaluate super resolution models are PSNR and SSIM.

Peak Signal-to-Noise Ratio (PSNR):

PSNR is a metric used to evaluate the reconstruction quality of images, especially in super-resolution tasks. It is derived from the Mean Squared Error (MSE) between the original high-resolution image I and the reconstructed image $\hat{I}$. The equation for calculating PSNR value is shown in Equation 3.1 and 3.2.

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left[I(i,j) - \hat{I}(i,j) \right]^2 \tag{3.1}$$

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{3.2}$$

Where:

- MAX_I is the maximum possible pixel value (usually 255 for 8-bit images),
- $m \times n$ is the image dimension,
- Lower MSE implies higher PSNR, indicating better image quality.

Structural Similarity Index (SSIM):

SSIM measures the perceived quality of images by evaluating structural similarity, considering luminance, contrast, and structure between the original and super-resolved images. The equation for SSIM index is shown in Equation 3.3

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.3}$$

Where:

- μ_x, μ_y are the mean intensities,
- σ_x^2, σ_y^2 are the variances,
- σ_{xy} is the covariance,

- C_1 and C_2 are constants to stabilize the division.

SSIM ranges from 0 to 1, with 1 indicating perfect similarity. It better reflects human perception compared to PSNR.

3.1.2 Models

This study evaluated super resolution models using PSNR and SSIM in reference with Set14 dataset. The models DRCT-L and HMANet came out to be the best models with the values as given in Table 3.1. DRCT-L is a super-resolution model that leverages dynamic residual channel attention and lightweight architecture to effectively enhance image detail while maintaining low computational complexity. HMANet (Hierarchical Multi-scale Attention Network) employs multi-scale attention mechanisms to capture rich contextual information, enabling high-fidelity image reconstruction in super-resolution tasks.

Table 3.1: Comparison of PSNR and SSIM for DRCT-L and HMANet

Model	**PSNR (dB)**	**SSIM**
DRCT-L	29.54	0.8025
HMANet	29.51	0.8019

3.2 Image Segmentation

Image segmentation is a technique in computer vision that divides an image into distinct regions by grouping pixels based on features like color, texture, or intensity. It helps to localize objects by assigning each pixel to a specific class. Segmentation can be semantic, grouping similar classes, or instance-based, which separates individual objects of the same class. Deep learning models like Mask R-CNN are widely used for precise segmentation.

In multilabel classification, segmentation is used to isolate multiple objects within a single image, allowing each to be identified independently. This is crucial when different waste items overlap or are closely positioned. Using object boundaries and reducing background interference, segmentation enhances feature extraction and supports better label prediction.

3.2.1 Evaluation Metrics for Image Segmentation

Intersection over Union (IoU)

IoU measures the overlap between predicted and ground truth masks. It is calculated as the ratio of intersection to union. Higher IoU indicates better segmentation accuracy. The equation for IoU is shown in Equation 3.4

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{3.4}$$

- Area of Overlap: The region where the predicted segmentation mask and the ground truth mask overlap.

- Area of Union: The total region covered by either the predicted mask or the ground truth mask

3.3 Multi-label Classification

Classification is a supervised machine learning task where a model learns to assign predefined labels or categories to input data based on learned patterns. Here images after being segmented would be divided into multiple images with each image being of a different class. Further the classification model will classify each of these segmented images, thus predicting multiple classes of from the same image.

Chapter 4

Dataset and Experimental Setup

4.1 Dataset Description

The dataset used here contains real world images of waste categorized into 10 classes: battery, biological, cardboard, clothes, glass, metal, paper, plastic, shoes, and trash [6]. The dataset contains 19,762 images and some of the sample images are shown in the Figure 4.1.

4.1.1 Class Imbalance

Class imbalance refers to a situation in machine learning where some classes have significantly more samples than others, potentially biasing the model's predictions. The dataset we used here also has irregular distribution of images among all the classes. The 'clothes' class in the dataset has way more images compared to all other classes. A pie chart showing the distribution

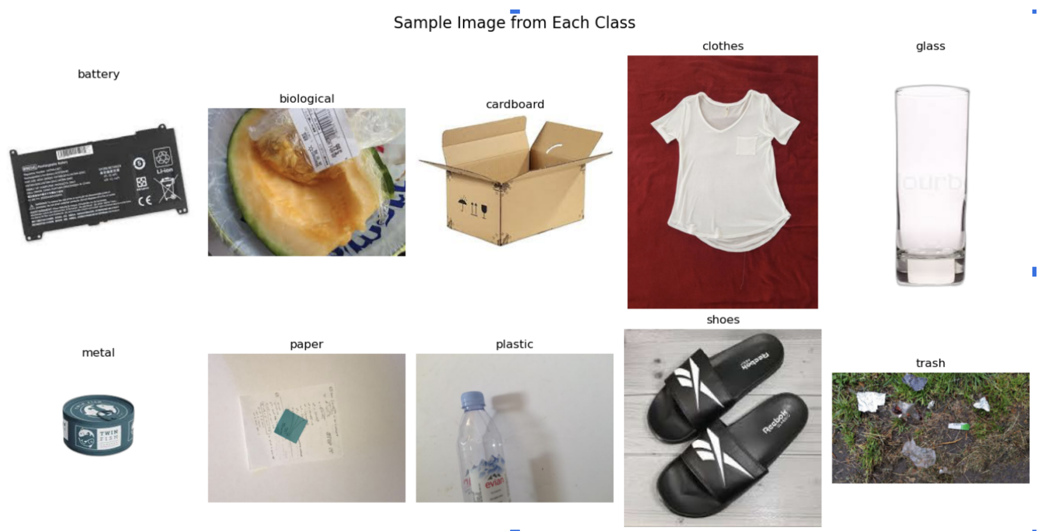


Figure 4.1: Samples of images from the dataset

of images among the classes is shown in Figure 4.2.

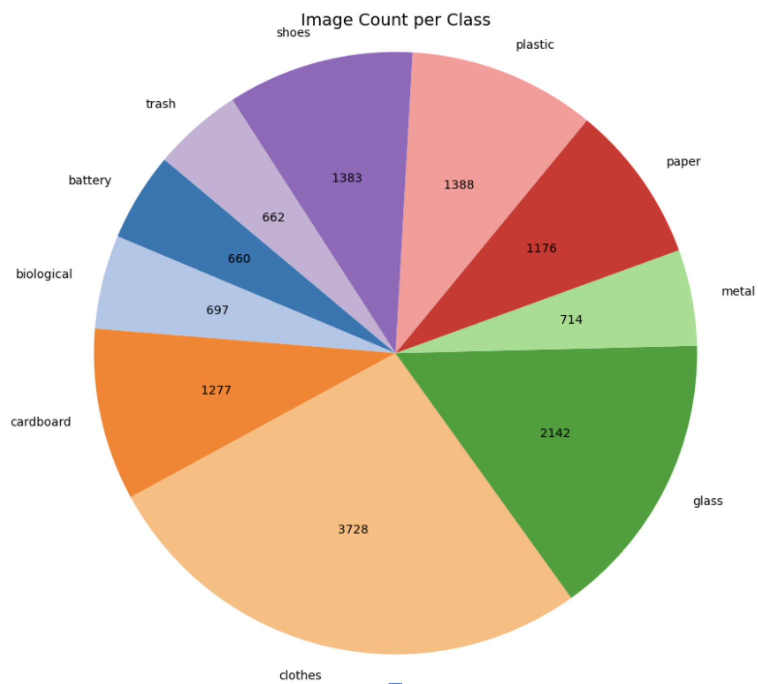


Figure 4.2: Pie chart showing distribution of images to each of the classes

4.1.2 Data Preprocessing

Class imbalance would lead to overfitting of the model on a single class which would lead to hindrance in performance. Thus we need to implement certain preprocessing steps to tackle class imbalance. Inorder to tackle this problem we need standardize the number of images in a single class to a constant value. Here number of images have been initialized to 1300 in a single class. The classes with more than 1300 images, the extra images were removed from the dataset. In the classes with less than 1300 images data augmentation steps like rotating, zooming and flipping were applied to increase the number of images to 1300.

4.2 Experimental Setup

4.2.1 Tools and Framework used

Visual Studio Code (VS Code) is a popular, lightweight code editor that provides robust support for Python development, making it an ideal environment for machine learning and data analysis projects. In this setup, Python was used as the primary programming language due to its simplicity and extensive ecosystem of libraries. Key frameworks included TensorFlow for building and training machine learning models, NumPy for efficient numerical computations and array manipulations, and Pandas for data handling and preprocessing tasks.

4.2.2 Models Used

Transfer learning models leverage pre-trained neural networks, such as CNNs, to solve new tasks with limited data. Transfer learning boosts performance and speeds up training by reusing learned features from large datasets like ImageNet. Convolutional Neural Networks (CNNs) excel in image-related tasks by extracting hierarchical features. Thus Efficient Net and ResNet transfer learning models have been finetuned to train the model, additionally we have also implemented a CNN model.

4.2.2.1 Efficient Net

EfficientNet is a deep learning model developed by Google that optimizes both accuracy and computational efficiency. It introduces a compound scaling method to balance network depth, width, and resolution, starting from a baseline model (EfficientNet-B0) and scaling up to B7[10]. By using techniques like depthwise separable convolutions and squeeze-and-excitation blocks, EfficientNet achieves high performance with fewer parameters and lower computational cost. It is widely used for image classification tasks, offering a powerful balance between speed and accuracy.

In this implementation, an EfficientNetB3 model is used as the base architecture, preloaded with ImageNet weights and configured to exclude its original top (classification) layers. The base model's layers are frozen to retain the pretrained feature extraction capabilities without updating its weights during initial training. On top of this, two custom dense layers with 128 and 64 units respectively (both using ReLU activation) are added for further feature transformation. The final output layer uses a *softmax* activation function with 10 neurons, suitable for multi-class classification. This design enables transfer learning by leveraging the strong visual feature extraction of EfficientNetB3 while customizing the classifier for a specific 10-class problem.

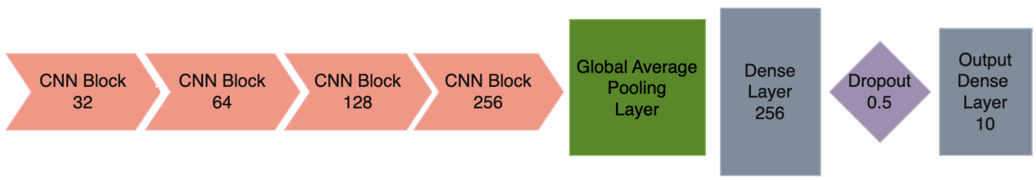
4.2.2.2 ResNet

ResNet, or Residual Network, is a deep convolutional neural network introduced by Microsoft that addresses the vanishing gradient problem in very deep architectures[3]. It uses residual connections, or "skip connections," which allow the model to learn identity mappings by bypassing one or more layers. This enables the training of extremely deep networks, such as ResNet-50 or ResNet-152, without degradation in performance. ResNet has significantly advanced image recognition tasks and is widely used in computer vision for its accuracy and training stability.

In this implementation, the ResNet50 model is utilized as the base model, loaded with ImageNet weights and configured without its original top classification layer. This setup enables the model to produce a compact, global feature vector through average pooling. The base model is frozen, ensuring its pretrained convolutional layers are not updated during training. On top of this, a new classification head is constructed using two dense layers with 128 and 64 units respectively, both employing the ReLU activation function. A final dense layer with 10 neurons and a softmax activation function is added to perform multi-class classification. This approach leverages the powerful feature extraction capabilities of ResNet50 and adapts them for a specific 10-class problem through transfer learning.

4.2.2.3 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are specialized deep learning models designed for image and spatial data analysis. They use layers of convolution, pooling, dropout, and activation to automatically extract and learn hierarchical features, making them highly effective for visual recognition tasks. The configuration of the CNN model used here is shown in the Figure 4.3



(a) Block diagram showing the architecture of the CNN model

(b) Block diagram showing the architecture of CNN block in diagram (a)

Figure 4.3: Two vertically stacked figures: (a) and (b).

Explanation of each of these layers shown in Figure 4.3 is as given below and the different configurations of these models is also explained below.

1. **Convolutional Layer (kernel size = 3×3):**
   Extracts spatial features using 3×3 filters that scan the input image, capturing patterns like edges, textures, and shapes while preserving spatial relationships in the data. The number of filters used in each CNN block is give in the Figure 4.3 a.

2. **Batch Normalization Layer:**
   Normalizes outputs of the previous layer across the batch, stabilizing

learning, reducing internal covariate shift, and accelerating convergence with learnable scale and shift parameters for each feature map.

3. **Max Pooling Layer (pool size = 2×2):**

   Reduces spatial dimensions by selecting the maximum value in each 2×2 region, helping to retain dominant features and reduce computational load while adding spatial invariance.

4. **Average Pooling Layer:**

   Reduces feature map dimensions by computing the average in each pooling region, preserving overall information and creating smoother representations, often used before classification layers.

5. **Dropout Layer (rate = 0.5):**

   Randomly deactivates 50% of neurons during training to prevent overfitting, forcing the network to learn redundant, generalized representations that improve model robustness and accuracy.

6. **Dense Layer:**

   Fully connected layer that maps high-level features to output classes or targets, learning complex relationships by weighting each input's contribution through trainable parameters and nonlinear activation functions.

Chapter 5

Results

5.1 Evaluation Metrics

Evaluation metrics are essential for assessing the performance of machine learning models. They provide quantitative measures to compare predictions with actual outcomes, helping to determine model accuracy, precision, recall, and overall effectiveness. The choice of metric depends on the problem type, such as classification, regression, or multi-label tasks. The equations for each of these evaluation metrics are shown in Equations 5.1-5.4.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

5.2 Results

The model had been tested on a dataset and these images were taken from the original dataset using a 70-30 train test split. The accuracy and weighted average of precision, recall and f1-score were taken to evaluate the models. The Table 5.1 shows the values of these evaluation metrics on different models.

Table 5.1: Performance Comparison of Models

Models	**Precision**	**Recall**	**F1-score**	**Accuracy**
Efficient Net	0.93	0.92	0.92	0.92
Res Net	0.94	0.93	0.93	0.93
CNN	0.69	0.65	0.65	0.65

From the table we conclude that the ResNet model performed best for the dataset and the confusion matrix for Res Net model is shown in the Figure 5.1. Both the transfer learning models are pre-trained on a large image dataset and thus results in a way better accuracy than the traditional CNN model.

The class wise accuracy, precision and recall is also important to understand the performance of the model for each of the classes. The plot showing the class wise evaluation metrics are shown in Figure 5.2.

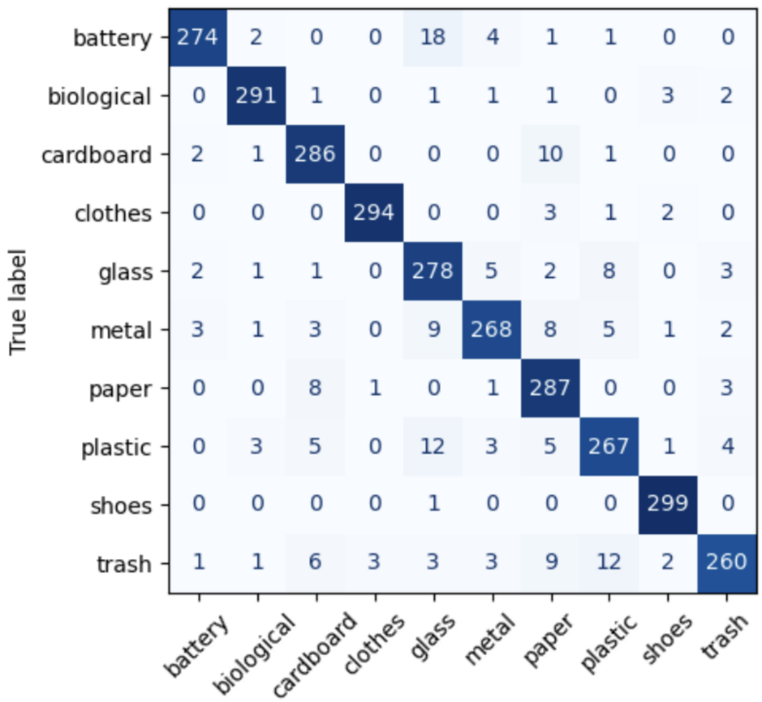


Figure 5.1: Confusion matrix showing the results obtained by the Res Net model

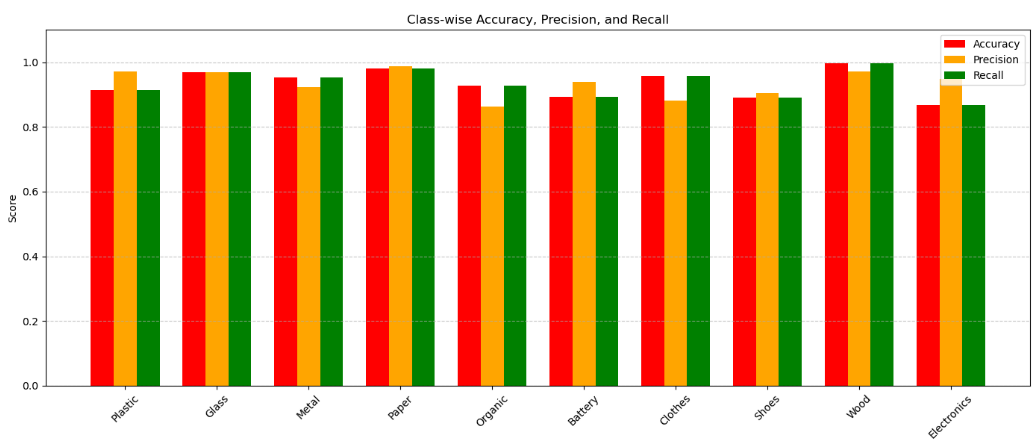


Figure 5.2: Class wise evaluation metrics of the Res Net model is plotted in the above graph

Chapter 6

Conclusion

Super-resolution techniques enhance image quality by generating high-resolution outputs from low-resolution inputs, thereby increasing the number of features within an image. This boost in feature visibility helps deep learning models better distinguish patterns and improve classification performance. Similarly, instance segmentation offers pixel-level precision by identifying and separating individual objects in an image. This is especially useful for multi-label classification tasks where multiple distinct items must be recognized and labeled within a single frame, thus enabling more accurate predictions.

Moreover, pretrained transfer learning models like EfficientNet and ResNet outperform traditional CNNs due to their prior training on large datasets, enabling them to extract generalized features effectively. These models can be fine-tuned on specific datasets with less data and still deliver high performance. Additionally, combining these models through ensembling can further boost accuracy by leveraging their individual strengths, offering a more reliable and generalizable solution for complex image classification tasks.

Bibliography

[1] Shu-Chuan Chu, Zhi-Chao Dou, Jeng-Shyang Pan, Shaowei Weng, and Junbao Li. Hmanet: Hybrid multi-axis aggregation network for image super-resolution, 2024.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Chih-Chung Hsu, Chia-Ming Lee, and Yi-Shiuan Chou. Drct: Saving image super-resolution away from information bottleneck, 2024.

[5] Md. Sakib Islam, Md. Shaheenur Sumon, Molla Majid, Saad Kashem, Mohammad Nashbat, Azad Ashraf, Amith Khandakar, Ali Kunju, and Mazhar Hasan-Zia. Eccdn-net: A deep learning-based technique for efficient organic and recyclable waste classification. *Waste management (New York, N.Y.)*, 193:363–375, 12 2024.

[6] Suman Kunwar. Garbage dataset, 2024.

[7] Shanshan Meng and Wei-Ta Chu. A study of garbage classification with convolutional neural networks. In *2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, pages 152–157, 2020.

[8] Eslam Mohamed, Abdelrahman Shaker, Ahmad El-Sallab, and Mayada Hadhoud. Insta-yolo: Real-time instance segmentation, 2024.

[9] Aakash Sharma, Aarambh Keshri, Aayush Kumar, and R. K. Yadav. Garbage classification with deep learning techniques. In *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 406–411, 2023.

[10] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.