PROJECT REPORT

TEAM: TRASH TALKERS

ORGANIC / INORGANIC WASTE CLASSIFICATION

TEAM MEMBERS

Abdullah Danish
Pushkar Desai
Rama Krishnan
Abdul Kalam Azad
Joe Daniel

VIABILITY OF PROBLEM

The Problem: The Recycling Headache

- **The Core Issue:** People frequently put trash into the wrong bins (e.g. food waste in the plastics bin). This cross-contaminates entire batches of recyclable materials.
- **The Consequence:** A single contaminated item can render an entire truckload of recyclables unusable, forcing it to be sent to a landfill. This makes the recycling process expensive and inefficient

Real-World Applications

- **Smart Bins:** For homes and public spaces to automate sorting or guide users to the correct bin.
- **Industrial Recycling Facilities:** To improve the speed and accuracy of large-scale sorting operations.

OUR APPROACH

- **Step 1:** Build a Diverse Data Foundation
  - Aggregate: We combined three different public datasets from Kaggle to create a large and varied collection of waste images.
  - Clean & Balance: We performed data cleaning to remove corrupted files and ensure a balanced distribution between Organic and Recyclable classes, resulting in a robust dataset of over 296,000 images.
- **Step 2:** Intelligent Feature Extraction with ResNet50
  - Transfer Learning: Instead of training from scratch, we used a pre-trained ResNet50 model, a powerful deep neural network, to analyze each image.
  - Generate "Digital Fingerprints": The model converts each image into a 512-point numerical vector. This vector captures the key visual features of the waste item, transforming the complex image into simple, structured data.
- **Step 3:** Optimized Classification
  - Ensemble Models: We fed these "fingerprints" into powerful machine learning classifiers (XGBoost, Random Forest, LightGBM) to make the final prediction.
  - Hyperparameter Tuning: We used advanced optimization techniques (Optuna, RandomizedSearchCV) to fine-tune each classifier, squeezing out the maximum possible accuracy and achieving a top performance of 97.1%.

DATA COLLECTION AND PREPROCESSING

To build a robust model, a large and diverse dataset is essential. We aggregated data from three distinct sources on Kaggle to ensure variability in image quality, lighting, and composition

1. **Waste Classification Data** by techsash
2. **Non and Biodegradable Waste Dataset** by rayhanzamzamy
3. **Recyclable and Household Waste Classification** by alistairking

The raw, combined dataset contained over **296,000 images**. The preprocessing steps were as follows:

- **Data Loading:** Scripts were written to parse the different directory structures of each dataset, labeling images as either organic (True) or non-organic (False).
- **Data Consolidation:** All metadata was consolidated into a single Pandas DataFrame, mapping each image file path to its corresponding label.
- **Data Cleaning:** We performed a sanity check on the dataset to ensure data quality. This involved removing entries with invalid file paths or unsupported extensions (non-JPG/PNG files). We also filtered out non-square images to maintain consistency, which simplifies the resizing process for the model.

After cleaning, the final dataset consisted of 296,493 images, with a reasonably balanced distribution between the two classes (**152,319 Non-Biodegradable** and **144,174 Biodegradable**).

FEATURE EXTRACTION

Instead of training a deep convolutional neural network (CNN) from scratch, which is computationally expensive and requires an enormous amount of data, we employed transfer learning.

- **Model Choice:** We chose the ResNet50 architecture, pre-trained on the ImageNet dataset. ResNet50 is a powerful model that has already learned to recognize a rich hierarchy of features (edges, textures, shapes) from millions of images.
- **Fine-Tuning:** We modified the ResNet50 model for our specific task. The original final classification layer (designed for 1000 ImageNet classes) was replaced with a new sequence of layers culminating in a single output neuron for our binary classification problem.
- **Training the Head:** We "froze" the weights of all the convolutional layers of the ResNet50 model and only trained the newly added classification layers (the "head"). This allowed the model to learn how to use the pre-learned features to distinguish between organic and non-organic waste. The model was trained for several epochs, minimizing the Binary Cross-Entropy loss.
- **Extracting Features:** After fine-tuning, we removed the final classification neuron. The output from the preceding layer 512-dimensional vector was used as the feature embedding for each image. This vector provides a dense, numerical representation of the visual content of the image, suitable for use with classical machine learning algorithms. This process was applied to every image in our dataset, creating a new tabular dataset of feature vectors and their corresponding labels.

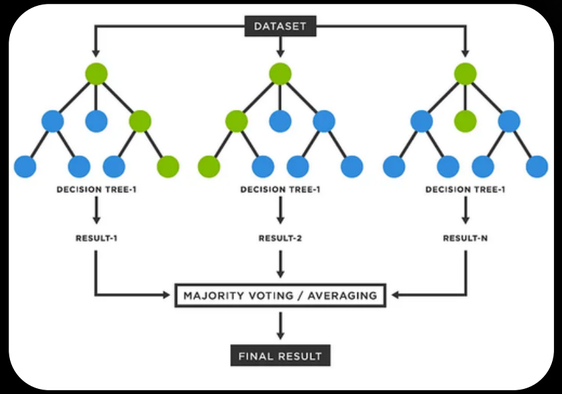
CML FOR CLASSIFICATION

With the 512-dimensional feature vectors extracted, the problem was transformed from an image classification task to a more traditional tabular data classification task. We experimented with three powerful ensemble models.

1. Random Forest: An ensemble of decision trees that operates by constructing multiple trees at training time and outputting the mode of the classes. It is robust to overfitting. We used RandomizedSearchCV for hyperparameter tuning.

2. XGBoost (Extreme Gradient Boosting): A highly efficient and scalable implementation of gradient boosting. It builds models sequentially, with each new model correcting the errors of the previous one.

3. LightGBM: Another gradient boosting framework that uses tree-based learning algorithms. It is known for its high speed and efficiency, especially on large datasets, due to its use of histogram-based algorithms. We used the **Optuna** framework for hyperparameter optimization.



LightGBM

SOLUTION IMPLEMENTATION

Environment and Data Handling

- Language: Python 3
- Core Libraries: We used Pandas for efficient data management and Matplotlib/Seaborn for data visualization and analysis.
- Platform: The entire project was developed and trained in a Jupyter Notebook environment, allowing for rapid experimentation.

The "Eyes": Deep Learning for Feature Extraction

- Framework: PyTorch
- Architecture: We leveraged a pre-trained ResNet50 model. Instead of building a vision model from the ground up, we used this state-of-the-art architecture to analyze the images and convert them into 512-dimensional feature vectors.

The "Brain": Machine Learning for Classification

- Classifiers: We implemented and compared three powerful gradient-boosting and ensemble models:
  - XGBoost
  - LightGBM
  - Scikit-learn's RandomForest
- Model Persistence: The final, optimized models were saved using Joblib for easy deployment and real-world use.

Peak Performance: Automated Optimization

- Tuning: To achieve the highest accuracy, we employed automated hyperparameter tuning libraries:
  - Optuna
  - Scikit-learn's RandomizedSearchCV

RESULT ANALYSIS

Model	Test Accuracy	F1-Score
Random Forest	0.9917	0.99
XGBoost	0.9961	1.00
LightGBM	0.9963	1.00

While all models achieved over 99% accuracy, **LightGBM** demonstrated a slight edge with the highest overall accuracy of **99.63%**.

CONCLUSION

Key Achievements

- **High Accuracy:** We successfully developed a machine learning pipeline that classifies waste with 99.63% accuracy, demonstrating the power of combining deep learning features with optimized classical models.
- **Robust & Generalizable:** Our model, built on a diverse dataset of nearly 300,000 images, proved its effectiveness in a real-world test on a previously unseen image.

Future Work & Next Steps

- **Expand Classification:** Increase the number of waste categories beyond "Organic" and "Recyclable" to include specific materials like glass, paper, different plastics, etc.
- **Real-Time Sorting Integration:** Adapt the model for use in automated sorting facilities, using cameras on conveyor belts to improve the speed and accuracy of recycling streams.