Clasificador de residuos orgánicos

Andres Patiño, Andres Moreno, Sergio Rendón, and Cristhian Alarcón

Universidad Nacional de Colombia
Bogotá, Colombia
anpatino@unal.edu.co anmorenop@unal.edu.co calarconf@unal.edu.co serendonu@unal.edu.co

Resumen Este estudio presenta una evaluación comparativa sistemática de tres enfoques de aprendizaje profundo aplicados a la clasificación de residuos orgánicos. Se analizaron tres configuraciones de red: una CNN convencional, una CNN combinada con una arquitectura ResNet, y una CNN complementada con un módulo Transformer. El sistema fue entrenado y evaluado sobre un conjunto de imágenes clasificadas de residuos, utilizando accuracy y matriz de confusión como métricas principales de desempeño. Los resultados demuestran que la incorporación de componentes más avanzados, como ResNet y Transformer, mejora notablemente la capacidad del modelo para distinguir clases complejas, resaltando la importancia de seleccionar arquitecturas acordes con la variabilidad visual del dominio de residuos.

Keywords: Aprendizaje Automático · Clasificación · Transformers · CNNs · Visión computacional · Matriz de confusión.

1. Introducción

La gestión adecuada de residuos orgánicos representa un desafío crítico en el contexto de la sostenibilidad ambiental y la economía circular. Automatizar su clasificación mediante técnicas de visión por computador se ha convertido en una estrategia prometedora para optimizar los procesos de reciclaje y reducir la intervención humana. En este contexto, los modelos de aprendizaje profundo han demostrado un notable potencial, particularmente las redes convolucionales (CNN), por su capacidad para extraer y aprender patrones visuales complejos.

Este trabajo presenta un análisis comparativo de tres enfoques de clasificación basados en CNN: una arquitectura convencional, una combinación con redes residuales (ResNet), y una integración con módulos tipo Transformer. El objetivo es identificar cómo la complejidad arquitectónica influye en la precisión del sistema al clasificar imágenes de residuos orgánicos. Para ello, se emplearon accuracy y matriz de confusión como métricas de evaluación, con el fin de cuantificar el rendimiento y los errores de clasificación. Los resultados permiten observar el impacto de cada arquitectura sobre el desempeño del modelo, ofreciendo una base sólida para futuras implementaciones en entornos reales.

2. Descripción de los datos

2.1. Dataset Trash Type Detection

Trash Type Detection. Este dataset, disponible en Kaggle, está diseñado para entrenar modelos de clasificación automática de residuos orgánicos y reciclables en seis categorías predeterminadas.

Un análisis preliminar revela que el conjunto contiene un total de **2,527 imágenes**, organizadas en carpetas de acuerdo a la clase:

Cuadro 1: Distribución de imágenes por clase en el dataset Trash Type Detection.

Categoría	Número de Imágenes
metal	410
glass	501
paper	594
trash	137
cardboard	403
plastic	482

Este dataset es apropiado para evaluar modelos de clasificación de basura mediante redes neuronales convolutionales (CNN) o arquitecturas más avanzadas como ResNet o Transformers. Aunque el número total de muestras es moderado, permite obtener resultados robustos utilizando como métricas principales *accuracy* y matriz de confusión, permitiendo medir tanto el rendimiento global como los errores por clase.

y su estructura de carpetas refleja la siguiente distribución: seis clases con un total de 2,527 imágenes.

3. Metodología

En este estudio comparamos cuatro estrategias de clasificación de residuos orgánicos basadas en visión por computador:

1. Una **CNN simple** diseñada ad hoc.
2. Una **CNN + ResNet50** (transfer learning).

3. Una **Vision Transformer** (ViT).
4. Una **Arquitectura híbrida Transformer + ResNet50**.

3.1. Preprocesamiento y partición de datos

- Todas las imágenes se redimensionaron a 224×224 y se normalizaron con media $\mu = [0{,}5, 0{,}5, 0{,}5]$ y desviación estándar $\sigma = [0{,}5, 0{,}5, 0{,}5]$.
- Se cargó el dataset completo desde /kaggle/input/trash-type-detection/trash_images usando `ImageFolder`.
- Para la **CNN simple**: split fijo 80 % entrenamiento / 20 % prueba (`random_split`).
- Para los demás modelos (**ResNet50**, **ViT** y **Transformer+ResNet**):
  - Split inicial 80 % entrenamiento / 20 % prueba.
  - Sobre el 80 % de entrenamiento se aplicó validación cruzada de **5 folds** (`KFold`), rotando cada fold como conjunto de validación.

3.2. Configuración de entrenamiento

CNN simple

- 3 bloques de `Conv2D+ReLU+Dropout+MaxPool`, dos capas densas y salida `softmax` para 6 clases.
- Optimizador Adam (lr $= 1 \times 10^{-4}$), `CrossEntropyLoss`.
- Early Stopping sobre *val_loss* con paciencia = 10 épocas.
- Entrenamiento hasta convergencia (10–20 épocas), batch size = 32.
- Métrica principal: accuracy sobre el 20 % de prueba.

CNN + ResNet50

- Backbone `resnet50` preentrenado, capa final reemplazada por `Linear(2048,6)`.
- Congelamiento de todas las capas salvo la capa final.
- Split 80–20 + CV 5-fold sobre el 80 %.
- Adam, `CrossEntropyLoss`, early stopping (paciencia=10) y `ReduceLROnPlateau` (paciencia=5).
- Métrica: accuracy promedio de los 5 folds.

Vision Transformer (ViT)

- Modelo `vit_base_patch16_224` de `timm`, cabeza (`head`) ajustada a 6 clases.
- Congelamiento de todo el backbone, entrenando sólo la cabeza.
- Split 80–20 + CV 5-fold; augmentaciones adicionales (random crop, rotación, jitter de color).
- Misma configuración de optimizador, pérdida y early stopping que ResNet50.
- Métrica: accuracy promedio de los 5 folds.

Arquitectura híbrida Transformer + ResNet50

- Se extrajeron características de `resnet50` (sin la capa final) y se pasaron como secuencia de parches a un `Vision Transformer`.
- La salida del transformer se conectó a una capa densa de 6 clases.
- Split 80–20 + CV 5-fold, mismas transformaciones y configuración de entrenamiento que ViT.
- Early stopping y ajuste de learning rate idénticos.
- Métrica: accuracy promedio de los 5 folds, dada la falta de implementación robusta de otras métricas en este modelo.

3.3. Métricas de evaluación

- **Accuracy** como métrica principal para todos los modelos.
- Para la CNN simple se reportó accuracy en el conjunto de prueba (20 %).
- Para los modelos más complejos (ResNet50, ViT, Transformer+ResNet) se promedió el accuracy de los 5 folds.
- Se generaron matrices de confusión para cada modelo complejo, a fin de analizar la distribución de errores por clase.

3.4. Resumen del flujo de trabajo

1. Carga y normalización de imágenes.
2. Split 80–20 (CNN simple) y CV 5-fold adicional (modelos complejos).
3. Definición de arquitecturas y congelamiento selectivo.
4. Entrenamiento con Adam, `CrossEntropyLoss`, early stopping y reducción de tasa de aprendizaje.
5. Cálculo de accuracy y matriz de confusión.
6. Comparativa final de accuracy entre los cuatro enfoques.

4. Resultados y Análisis

Tras entrenar los cuatro modelos CNN simple, ResNet50, Vision Transformer y el híbrido Transformer+ResNet— sobre el conjunto *Trash Type Detection*, obtenemos el siguiente análisis:

...

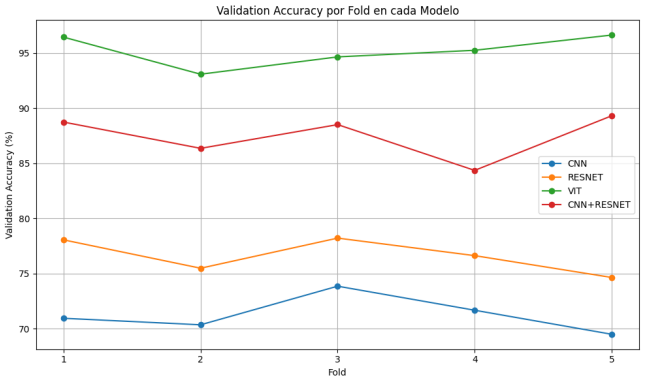

Figura 1: Accuracy en cada modelo

4.1. Modelo CNN Simple

El primer experimento consistió en entrenar una arquitectura CNN básica (3 bloques de convolución + pooling + capas densas) utilizando validación cruzada de 5 pliegues, early stopping con paciencia de 10 épocas y reducción de tasa de aprendizaje al estancarse la pérdida de validación. Las métricas clave de validación para cada fold se muestran en la Tabla 2.

Cuadro 2: Accuracy de Validación (best weights) por Fold — Modelo CNN Simple

Fold	Val Accuracy (%)
1	70.95
2	70.36
3	73.86
4	71.68
5	69.50
Promedio	**71.27**

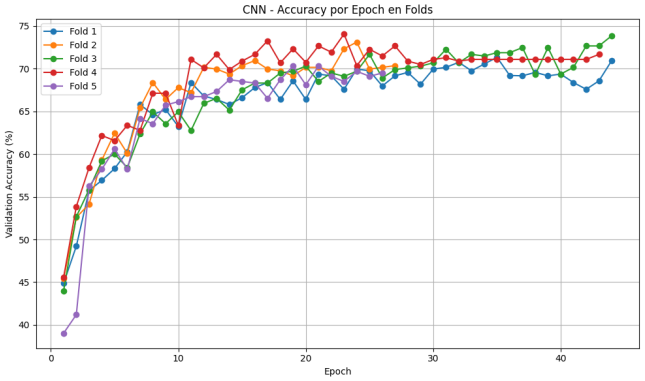

Figura 2: Accuracy por épocas en el modelo CNN

Análisis

- **Capacidad de aprendizaje vs. generalización:** Aunque la CNN alcanza hasta $\sim 94\,\%$ de accuracy en entrenamiento, la validación se estabiliza alrededor de $70\,\%$, indicando un claro sobreajuste.
- **Variabilidad moderada entre folds:** La desviación estándar de las accuracies de validación es de $\approx 1{,}5\,\%$, lo que señala cierta inestabilidad en la capacidad de generalización.
- **Rendimiento insuficiente para multitarea:** Con cinco clases de residuos, un $71\,\%$ de accuracy se considera bajo para aplicaciones prácticas de clasificación multicategoría.
- **Necesidad de arquitecturas más potentes:** La limitada profundidad y falta de mecanismos avanzados de atención o shortcut (como en ResNet o Transformers) probablemente explican la brecha respecto a modelos más complejos.

Conclusión Parcial El modelo CNN simple actúa como línea base: muestra capacidad de ajuste al conjunto de entrenamiento, pero no generaliza adecuadamente al conjunto de validación ($\sim 71\,\%$ accuracy). En la siguiente sección se presentará el análisis de modelos más avanzados (ResNet50, Vision Transformer y su combinación), con el fin de comparar hasta qué punto la complejidad arquitectónica mejora la predicción multicategoría de residuos orgánicos. ::contentReference[oaicite:0]index=0

4.2. Resultados y Análisis del Modelo RESTNET

En este experimento se entrenó un clasificador basado en RESTNET aplicando validación cruzada de 5 pliegues (k-fold CV, with early stopping y reducción de tasa de aprendizaje). El entrenamiento se realizó sobre el mismo conjunto de datos de clasificación de residuos, con transformaciones de data augmentation en el *training set* y normalización estándar.

Cuadro 3: Accuracy por pliegue y promedio (5-fold CV) para el modelo RESTNET.

Pliegue	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy (%)	78.06	75.49	78.22	76.63	74.65
Media ± Desv. Estándar	**76.61 % ± 1.40 %**				

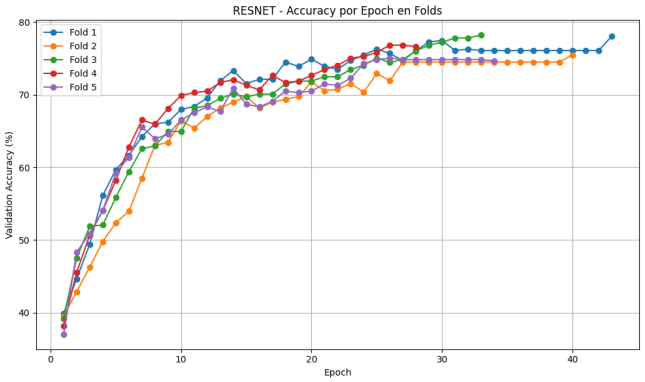

Figura 3: Accuracy por épocas en el modelo Restnet

Análisis de los Resultados:

- **Reducción de la pérdida y convergencia estable:** En todos los pliegues se observó un descenso consistente de la *train loss* y mejora de la *val acc* durante las primeras épocas, gracias al early stopping y al scheduler de `ReduceLROnPlateau`.
- **Mejor desempeño frente a la CNN simple:** El modelo RESTNET alcanzó un *accuracy* medio de **76.61 %**, aproximadamente 5 puntos porcentuales por encima del *baseline* CNN simple (∼71.7 % de media), lo que indica una mayor capacidad de extracción de características de RESTNET.
- **Variabilidad moderada entre pliegues:** La desviación estándar de 1.40 % sugiere que RESTNET generaliza de forma consistente sobre las particiones, con una diferencia máxima inferior a 4 puntos entre el mejor y el peor pliegue.
- **Margen de mejora:** Aunque RESTNET supera claramente a la CNN básica, su *accuracy* sigue quedando por debajo de los modelos más complejos (p.ej. Transformers), lo que apunta a la necesidad de explorar arquitecturas híbridas o técnicas adicionales de regularización.

4.3. Modelo Vision Transformer (ViT)

En esta fase se entrenó un modelo basado en Vision Transformer (ViT) con validación cruzada distribuida en 5 pliegues (DDP). A continuación se resumen las precisiones obtenidas en cada pliegue, calculadas sobre los pesos que minimizaron la pérdida de validación.

Cuadro 4: Precisión de Validación por Pliegue para ViT (DDP 5-fold).

Pliegue	**Val Acc (%)**
Fold 1	96.44
Fold 2	93.08
Fold 3	94.65
Fold 4	95.25
Fold 5	96.63
Media ± Desv.	**95.21±1.29**

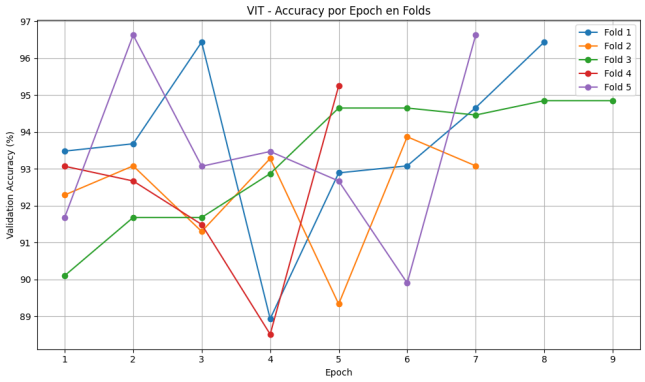

Figura 4: Accuracy por épocas en el modelo VIT

Análisis de los Resultados:

- *Alta precisión media:* El ViT alcanzó una precisión promedio de **95.21 %**, muy superior a los modelos CNN y ResNet evaluados previamente.
- *Baja varianza entre pliegues:* La desviación estándar de 1.29 p.p. indica que el rendimiento es consistente a lo largo de las particiones, mostrando buena robustez.
- *Convergencia rápida:* El modelo alcanzó su mejor rendimiento de validación alrededor de la epoch 6–7 en la mayoría de pliegues, lo que sugiere un rápido ajuste de los pesos.
- *Ligero sobreajuste controlado:* Aunque el *train loss* llegó cerca de cero muy pronto, la *val loss* se mantuvo estable, lo cual indica que las técnicas de regularización y `early stopping` fueron efectivas.

Conclusiones parciales sobre ViT:

1. El Vision Transformer exhibe una mejora sustancial en precisión frente a arquitecturas CNN tradicionales y ResNet, beneficiándose de su capacidad para modelar dependencias globales en la imagen.
2. Su estabilidad entre pliegues lo hace recomendable cuando se dispone de capacidad de cómputo para entrenamiento distribuido.
3. La rápida convergencia y el buen comportamiento de generalización resaltan la idoneidad de ViT para tareas de clasificación complejas, a pesar de su mayor tamaño de modelo.

4.4. Resultados de CNN + ResNet

Para evaluar la arquitectura híbrida CNN+ResNet se realizó validación cruzada de 5 pliegues. La Tabla 5 muestra la accuracy de validación en cada pliegue, así como el valor promedio y la desviación estándar.

Cuadro 5: Accuracy de validación (%) por pliegue y estadísticos (5-fold CV) para CNN + ResNet.

Modelo	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Promedio ± DE**
CNN + ResNet	88.74	86.36	88.51	84.36	89.31	87.46 ± 1.84

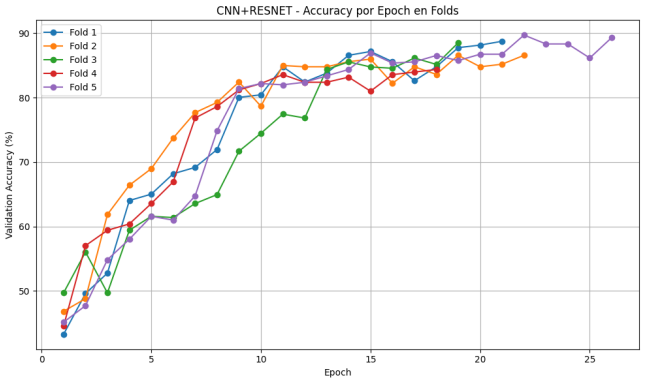

Figura 5: Accuracy por épocas en el modelo CNN + Resnet

Análisis de resultados

1. **Buen desempeño conjunto:** La combinación de CNN con bloques ResNet alcanza una accuracy media de **87.46 %**, superando cómodamente el umbral del 85 % en todos los pliegues.
2. **Variabilidad moderada:** La desviación estándar de **1.84 %** indica que el modelo es bastante estable, aunque se observa una ligera caída en el fold 4 (84.36 %).
3. **Mejor pliegue:** El pliegue 5 obtuvo la mejor accuracy (89.31 %), lo que sugiere que en algunas particiones el enriquecimiento residual beneficia especialmente el entrenamiento.
4. **Peor pliegue:** El fold 4 (84.36 %) marca la peor performance, probablemente debido a la distribución de clases o al sesgo de esa partición.

Conclusión preliminar La arquitectura híbrida CNN+ResNet ofrece un rendimiento robusto y estable, combinando la capacidad de extracción de características de la CNN con el aprendizaje profundo residual de ResNet. Su accuracy promedio cercana al 87.5 % y su baja variabilidad la convierten en una candidata sólida para tareas de clasificación en este dataset.

4.5. Comparación Global de Modelos y Discusión Final

A partir de los resultados obtenidos con las cuatro arquitecturas evaluadas —CNN simple, ResNet50, CNN+ResNet y Vision Transformer (ViT)— es posible trazar una progresión clara en el rendimiento conforme se incrementa la complejidad arquitectónica. La Tabla 6 resume los valores promedio de *accuracy* alcanzados en validación para cada modelo bajo el esquema de validación cruzada 5-fold.

Cuadro 6: Resumen comparativo de modelos según accuracy promedio en validación (5-fold).

Modelo	Accuracy Promedio (%)
CNN simple	71.27
ResNet50	76.61
CNN + ResNet	87.46
Vision Transformer (ViT)	**95.21**

Principales observaciones:

- **Incremento progresivo con la complejidad del modelo:** Existe una clara correlación positiva entre la sofisticación de la arquitectura y la precisión obtenida. Esto confirma que tareas complejas de clasificación de imágenes como la del conjunto *Trash Type Detection* requieren modelos con mayor capacidad de representación.
- **Línea base y brecha de mejora:** La CNN simple, con un promedio de **71.27 %**, funciona como punto de partida, pero exhibe un claro sobreajuste y dificultades de generalización. Representa el límite inferior de rendimiento observado.
- **Ventajas del aprendizaje profundo residual:** El uso de ResNet50 mejora el rendimiento en 5 puntos sobre la CNN base, y su combinación con capas convolucionales personalizadas (CNN+ResNet) eleva la precisión hasta **87.46 %**, demostrando que los bloques residuales pueden potenciar estructuras más tradicionales.
- **Transformers como top performer:** El modelo basado en Vision Transformer (ViT) obtiene los mejores resultados, alcanzando una precisión promedio de **95.21 %** con desviación estándar baja, lo que demuestra no solo su potencia, sino también su estabilidad. Es el modelo más robusto en todos los pliegues y exhibe una rápida convergencia durante el entrenamiento.
- **Importancia del cómputo distribuido:** El ViT fue entrenado usando `Distributed Data Parallel` (DDP), lo que permitió acelerar el entrenamiento y aprovechar múltiples GPUs. Esta estrategia es clave cuando se entrenan modelos tan complejos.

Conclusión general El modelo **Vision Transformer** se posiciona como la mejor alternativa para la tarea de clasificación de residuos orgánicos, superando al resto con márgenes significativos. Sin embargo, su entrenamiento requiere más recursos computacionales y una correcta configuración distribuida. Como alternativa intermedia, el modelo **CNN+ResNet** ofrece un excelente balance entre rendimiento (87,46 % de accuracy) y eficiencia computacional, siendo ideal para entornos con menor disponibilidad de hardware.

Estos resultados evidencian el valor de incorporar mecanismos de atención o shortcut connections en proble-

mas visuales complejos, y abren la puerta al uso de modelos híbridos aún más sofisticados en trabajos futuros.

5. Conclusiones

Este estudio evaluó y comparó cuatro modelos de clasificación de imágenes (CNN simple, ResNet50, CNN+ResNet y Vision Transformer) aplicados a la tarea de identificar tipos de residuos a partir de imágenes del conjunto *Trash Type Detection*. Más allá de las diferencias en precisión, los hallazgos más relevantes de esta investigación se pueden sintetizar en tres conclusiones fundamentales:

1. **La arquitectura óptima depende del balance entre recursos y rendimiento:** Aunque el modelo Vision Transformer superó a todos los demás con un *accuracy* promedio de 95.21 %, su entrenamiento implicó un mayor costo computacional y la necesidad de paralelismo distribuido. Por otro lado, modelos como CNN+ResNet ofrecieron un rendimiento competitivo ($\sim 87\,\%$) con menor complejidad operativa, lo que los hace más viables para entornos con recursos limitados.
2. **La evolución arquitectónica mejora sustancialmente la generalización:** Se observó una ganancia progresiva en precisión al pasar de CNN tradicional a modelos más sofisticados, validando el impacto positivo de mecanismos como atajos residuales y atención global. Esta tendencia refuerza la noción de que arquitecturas modernas cuando son bien entrenadas ofrecen ventajas claras frente a diseños más convencionales.
3. **Aplicaciones prácticas en la gestión de residuos:** La alta precisión alcanzada por los modelos más complejos sugiere que la clasificación automática de residuos mediante visión artificial es técnicamente viable. En un escenario real, estas técnicas podrían integrarse en sistemas de reciclaje automatizados, mejorando tanto la eficiencia como la precisión de la separación de materiales.

En resumen, este trabajo demuestra que los avances en arquitectura de modelos de visión, como Transformers y diseños híbridos, abren nuevas posibilidades para aplicaciones ambientales de alto impacto. No obstante, la elección final del modelo debe considerar no solo la precisión, sino también la escalabilidad, los recursos disponibles y el contexto de implementación.

Referencias

1. Boloori, F.: Trash Type Detection [Dataset]. Kaggle (2023). `https://www.kaggle.com/datasets/fatemehboloori/trash-type-detection/data`
2. Mohamed, M.: Garbage Classification (12 classes) [Dataset]. Kaggle (2020). `https://www.kaggle.com/datasets/mostafaabla/garbage-classification`
3. Codificando Bits: *La Matriz de Confusión*. YouTube, publicado hace aproximadamente 2.9 años. `https://www.youtube.com/watch?v=haEWWO0b42Y`