Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Explainable Deep Transfer Learning for Automated Waste Classification: A Comparative Study Using Grad-CAM

Muhammad Naeem Akhtar (Corresponding Author)
Faculty of Computer Science & Information Technology, The Superior University, Lahore, Pakistan Email: mnapucit@gmail.com

Dr. Jawad Ahmad
Faculty of Computer Science & Information Technology, The Superior University, Lahore, Pakistan Email: jawad.ahmad@superior.edu.pk

Muhammad Rizwan
Faculty of Computer Science & Information Technology, The Superior University, Lahore, Pakistan Email: mu.rizwan815@gmail.com

ABSTRACT

The study discussed the automated classification of waste with deep learning focusing on model interpretability. A comparison study of four image classifiers is performed. A custom *CNN* and three pretrained networks *ResNet50, MobileNetV2, and EfficientNetB0* on the publicly available *TrashNet** dataset containing 2527 images, and 6 waste categories. All models are preprocessed using data augmentation and evaluated using accuracy, precision, recall, and F1 score. The results indicate that transfer learning significantly improves the efficiency. EfficientNetB0 performed best among all with accuracy of approximately 84%, closely followed by ResNet50 (~83%). Whereas MobileNetV2 (~76%), while the custom CNN achieved nearly 51% accuracy. Our results are consistent with previous studies conducted on Trash-net showing performance improvement using transfer learning architectures such as *ResNet* and *DenseNet* models [7], [11], [14]. We have also used *Gradient-weighted* Class Activation Mapping *(Grad-CAM)* to explain the predictions of each model. The resulting heat maps show that the pre-trained models pay a lot of attention to useful regions of the waste object while the baseline CNN's attention is spread over a larger area, which is the reason for its being less successful. In the class-wise analysis, it is seen that the majority class *paper* is the easiest to classify, whereas the minority class *trash* is difficult, which highlights the issue of imbalance in the data sets. In summary, our work shows the effectiveness of transfer learning for waste recognition and the importance of explainable AI using Grad-CAM to interpret the behavior of deep models.

Keywords: Waste classification, Explainable AI, TrashNet. Transfer learning, CNN, ResNet50, MobileNetV2, EfficientNetB0, Grad-CAM

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Introduction

This study developed a deep transfer learning framework for automated waste classification using computer vision and Explainable Artificial Intelligence (XAI) interpretability analysis. A comparative evaluation was conducted using one custom CNN architecture and three pretrained transfer learning models, namely ResNet50, MobileNetV2, and EfficientNetB0. The publicly available TrashNet dataset consisting of approximately 2,500 waste images distributed across six categories. All models were trained and evaluated under identical preprocessing, augmentation, and optimization settings to ensure fair analysis. The study also integrated XAI using Grade-CAM to analyze model prediction interpretability.

Background and Motivation

Improper disposal and inefficient recycling of solid waste are among the most serious environmental, economic and public health issues as solid waste is significantly produced due to rapid urbanization, industrialization and population growth globally [7], [17]. The efficient separation of waste is thus critical for waste sustainability but manual sorting is time consuming, labour intensive, inaccurate and inconsistent [7, 8]. Thus, waste classification based on Artificial Intelligence (AI) using computer vision has become a promising solution [8]. Convolutional Neural Networks (CNNs) are one of the most successful techniques of Artificial Intelligence used for image classification[15], learning discriminative features from raw images automatically [1] [2]. However, creating CNNs from scratch takes large data sets and a lot of computational power, while public data sets like TrashNet have relatively limited data sets [30]. To overcome this, transfer learning has been proposed to adapt models pre-trained on large-scale datasets like ImageNet to tasks related to waste classification. Recent studies have shown that state of the art CNNs such as ResNet50, MobileNetV2 and EfficientNetB0 [2]-[4] have consistently achieved better classification accuracy and computational efficiency than traditional CNN architectures, and that transfer learning methods improve classification performance when applied to waste image classification [7], [9], [14].

Problem Statement

Although deep learning-based waste classification technologies have made tremendous progress, there are still several challenges to be solved. For many waste categories, there are similarities in the visual features of the waste, such as similarities in texture between cardboard and paper, or reflective surfaces between glass and metal, resulting in inter-class confusion and low classification reliability. Moreover, public datasets, like TrashNet, are suffering from the problem of class imbalance, thus, the minority classes have fewer training instances resulting in poor recall and generalization performance [30, 35]. Despite achieving high classification accuracy, deep learning models are hard to interpret due to the CNN's “black-box” nature and the lack of an understanding of the regions of an image that contribute to a classification [17] and [19]. This lack of transparency lowers the trustworthiness, especially in smart environmental applications. Furthermore, most of the previous studies focus on classification accuracy and present only partial analysis of class-wise performances, attention localization, explainability, and the effect of preprocessing

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

consistency. Thus, the necessity arises for a waste classification framework which can be understood and compared, has high predictive power, and is explainable.

Research Gap

Based on the literature review, a number of important gaps in the field of waste classification automation research were identified. The majority of the current works have investigated only one deep learning architecture instead of comparing various transfer learning models in similar experimental settings [7], [11], [14]. Besides, they mainly focus on the overall accuracy and neglect to report class-wise accuracy, which leads to the minority classes of trash and metal being difficult [11], [35]. Popular techniques such as Explainable Artificial Intelligence (XAI) including Grad-CAM, have been used in waste classification but they are not widely incorporated in comparative transfer learning frameworks [10]. Moreover, there have been relatively few studies on the consistency of preprocessing on the reliability of Grad-CAM and on the confidence of prediction.

Research Questions

Based on the identified research gaps, the present study addressed the following research questions.

What is the effect of transfer learning for waste classification compared to the use of a conventional CNN?

What is the best transfer learning architecture for classifying the TrashNet dataset: ResNet50, MobileNetV2 or EfficientNetB0?

What is the impact of data imbalance on (class-wise) classification behaviour and performance of minority classes?

Is it possible to enhance the interpretability and explain the model's attention behaviour in waste classification systems using Grad-CAM?

What effect does the preprocessing consistency have on the reliability and prediction confidence of Grad-CAM?

The research contribution of the study is a comparative study of CNN, transfer learning models, and an analysis of the influence of data imbalance and preprocessing consistency through the use of explainability provided by Grad-CAM.

The rest of this paper is organized as follows. The literature review is conducted in Section 2, the dataset, methodology and experimental set up is presented in Section 3–5, the obtained results and Grad-CAM analysis is discussed in Section 6-8 and the study is concluded in Section 9 with future research direction.

Literature Review

Waste Classification Using Deep Learning

In recent years, the amount of municipal solid waste has been rapidly growing, which has led to the need for an intelligent waste management system, so automated waste classification is an essential area of research. In image recognition, deep learning, specifically CNNs, has been proven to be very successful in learning discriminative features from raw images [1]. CNN-based and transfer learning-based methods have been proven to obtain high classification accuracy on TrashNet and other benchmark datasets, where the pretrained models have been superior to conventional CNNs in

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

feature extraction, generalization and classification reliability [7]-[9], [11]. In particular, Lilhore et al. [7] and Islam et al. [8] presented the use of deep learning algorithms in waste management for sustainability, while Meng et al. [11] and Shukla et al. [9] claimed that the transfer learning models like ResNet50, MobileNet and EfficientNet outperform models trained from scratch. These improvements have not yet made classification of waste easy, as categories are still visually similar, the number of data is still small, class imbalance exists, and model interpretability has not been achieved, which still makes waste classification difficult, especially for the minority classes, and impractically used in practice [11] [21].

Transfer Learning Architectures for Waste Classification

The use of a model that has been pre-trained on a large-scale dataset, like ImageNet, has become a popular strategy for image classification across various domain-specific tasks [6]. It works well on the classification of waste, where the datasets are not too large and it is not feasible to train deep networks from scratch. However, it has been demonstrated that transfer learning can also decrease training difficulty, increase accuracy in feature extraction, in the classification, and in the generalization of the model with a limited training set [24]. Pretrained architectures have been shown to consistently outperform traditional CNN-based architectures in terms of faster convergence, reuse of features and reduced overfitting, as found by Kunwar et al. [14] and Nguyen et al. [13]. The results were similar to those found in other studies of computer vision and it is well established that transfer learning between tasks and domains is effective for small and imbalanced datasets [24, 27, 28]. The observations are used to support the type of transfer learning models used in the present study which has used the TrashNet dataset made up of around 2500 images in six waste categories.

Comparative Analysis of CNN, ResNet50, MobileNetV2, and EfficientNet

Custom CNN

Custom CNNs are widely adopted as their baseline models for testing more sophisticated architectures [1]. They are effective learners of spatial features, but have limited performance due to small data sets and limited generalization capability, and their performance is often inferior to transfer learning models in waste classification tasks [7, 8, 15].

ResNet50

The ResNet50 network uses residual learning to facilitate efficient deep neural network training, and has shown to effectively extract features and converge stably on image classification problems [2]. The classification accuracy of ResNet50 has been found to consistently be high for waste classification, especially for visually similar categories, in previous studies [7], [11], [24].

MobileNetV2

MobileNetV2 is an efficient network architecture based on depthwise separable convolutions and inverted residual blocks (IRNs) [3] for mobile and embedded applications. Compared to larger models, it provides a good classification accuracy

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

while maintaining a high computational efficiency, which is important for real-time waste classification applications [9], [26].

EfficientNet

EfficientNet introduces compound scaling to optimize network depth, width and image resolution at the same time without compromising the accuracy, with relatively few parameters [4]. EfficientNet is a recent CNN model that has been reported to outperform several conventional CNN models and transfer learning models in waste classification using efficient feature representation and computational efficiency [9, 12, 25] and hence, it is included in the present study.

Explainable Artificial Intelligence and Grad-CAM

Although deep learning models have high classification accuracy, they are often regarded as "black-box" systems because it is hard to interpret their predictions. To tackle this issue, Explainable Artificial Intelligence (XAI) is employed to enhance the transparency of the models and to make it easier for users to understand why the model has made a given prediction [17] and [19]. Gradient-weighted Class Activation Mapping (Grad-CAM) is one popular approach for visualizing areas of the image that most significantly influence a model's prediction among the many techniques used in XAI [5]. Grad-CAM has proven to be effective in various image classification applications, such as medical imaging and environmental monitoring, where it is used to extract meaningful features, uncover model biases, and boost the trustworthiness of AI systems [18, 20]. Recently, it has been used for waste classification to check if models are concerned with relevant waste objects instead of background information [10]. Comparative explainability analysis across transfer learning architectures and the impact of preprocessing consistency on reliability of Grad-CAM, however, is largely underexplored, and this is the motivation for the present study.

Table 2.1 gives an overview of the recent research in waste classification using deep learning techniques. It demonstrates the general advantage of transfer learning over the traditional CNN models and limited research has been conducted on comparative XAI analysis and preprocessing consistency. These gaps are motivation for the present study.

Table 1.1: Comparative Summary of Related Studies

Ref. #	Study	Models Used	Dataset	XAI Used	Key Findings	Limitation
[7]	Lilhore et al.	VGG, ResNet, Hybrid TL	TrashNet	No	High classification accuracy	Limited explainability
[8]	Islam et al.	DenseNet201-based CNN	Multi-dataset	Grad-CAM	Strong feature extraction	Complex architecture

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

[9]	Shukla et al.	ResNet50, EfficientNet, MobileNet	Garbage dataset	No	EfficientNet performed strongly	Limited interpretability
[10]	Manik et al.	CNN	Household Waste	Grad-CAM	Improved interpretability	Limited model comparison
[11]	Chen et al.	ResNet50	TrashNet	No	High accuracy using TL	Small dataset concerns
[12]	Qiu et al.	EfficientNetV2	Garbage dataset	No	Improved classification accuracy	Increased complexity
[13]	Nguyen et al.	Multiple TL models	Waste dataset	No	Better generalization	Limited XAI analysis
[14]	Kunwar et al.	Multiple TL architectures	Waste dataset	No	TL outperformed CNN	Computational tradeoffs

Dataset and Preprocessing

TrashNet Dataset Description

The experiments were conducted using one of the most popular benchmark datasets for image-based waste classification called TrashNet [30]. Has approximately 2500 labelled images of 6 different categories: cardboard, glass, metal, paper, plastic, trash. They were captured under varying lighting conditions, backgrounds and orientations to provide a realistic and varied image to evaluate deep learning models. As a plus, this dataset also has the issue of classes that are visually similar, background variations, intra-class variations and the small sample sizes found in minority classes, making it a good test bed for the robustness of transfer learning architectures.

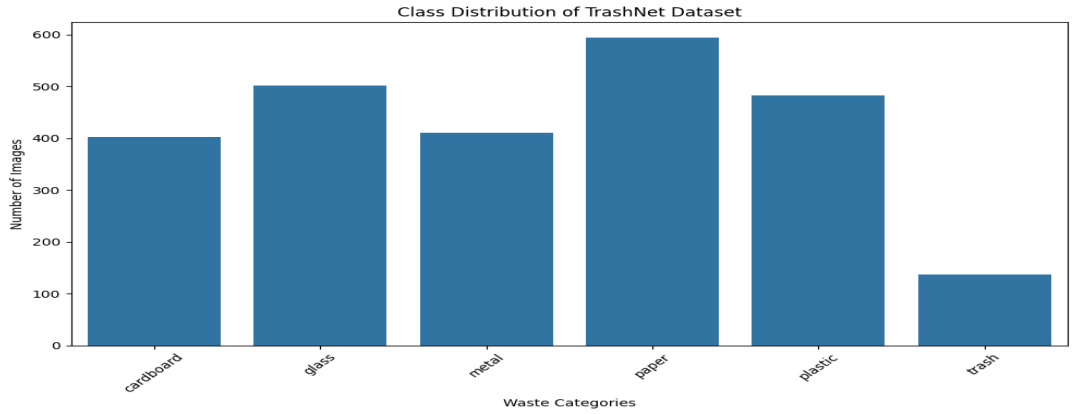

Figure 3.1: Class Distribution of TrashNet Dataset

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Dataset Distribution Analysis

The TrashNet is a data set which is widely respected, however, the distribution of the classes is not balanced with more samples in the paper and plastic classes than in the trash class. This imbalance did affect the results of the experiment, as the most frequent classes tended to have higher recall and more stable classification rates, and the less frequent classes, especially the trash classes, had lower rates of recall and F1-score. Secondly, there was a high degree of intra-class variability for classes with similar visual appearance (cardboard, paper and plastic) and this is a disadvantage of this data.

Data Augmentation

The augmentation was limited to the training set only, and implemented using TensorFlow/Keras image preprocessing utilities [31] to improve model generalization and reduce overfitting. Random rotation, width and height shifting, zooming and horizontal flipping were added as augmentations, ensuring that the models could learn robust features in different orientations, positions, and scales. No changes were made to the validation images, so as to assess performance in an unbiased manner.

Image Preprocessing

To ensure the images can be processed by the pre-trained models ResNet50, MobileNetV2, and EfficientNetB0 [33] all images were resized to 224 × 224 pixels. The image was resized, normalized, converted to a tensor, and batch generated in the pre-processing pipeline. To ensure transfer learning models are compatible with the pre-trained ImageNet models, the architecture-specific preprocessing functions were applied. It was found that prediction confidence was lowered and heatmaps were less accurate due to inconsistent preprocessing during Grad-CAM analysis. Both prediction confidence and attention localization have improved, with a model-specific preprocessing step – indicating the need for consistency in preprocessing when using explainable AI applications.

Train-Validation Split

The data set was split into training and validation datasets in an 80:20 ratio. 80% of the images were used for training the model and 20% for validation. The split was done with the TensorFlow image data generator with a fixed random seed to ensure it is reproducible. The accuracy, precision, recall, F1 score and confusion matrix analysis were used to measure model performance.

Proposed Methodology

Overview of Proposed Framework

The proposed framework was designed for comparison of performance and interpretability of deep learning models for automated waste classification. Data preprocessing, data augmentation, splitting the data into train and validation sets, training models, evaluating their performance, and explaining why models make the predictions they do (Grad-CAM). The following models were tested with the same experimental setup: a Custom CNN, ResNet50, MobileNetV2, and EfficientNetB0. The general approach used is shown in Figure 4.1.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Model Architectures

The baseline model was a Custom CNN comprising convolutional, max-pooling, dropout, fully connected and Softmax layers for six-class classification. The benefits of transfer learning are assessed using three pre-trained architectures (ResNet50, MobileNetV2 and EfficientNetB0) trained on the ImageNet benchmark dataset. Global Average Pooling, dense classification layers, and an output layer Softmax were used to replace the original classification layers. The selected models were ResNet50, which excels in feature extraction; MobileNetV2, which is lightweight and compact; and EfficientNetB0, which uses a balanced scaling approach and is efficient in terms of parameters.

Model Training Strategy

The same experimental conditions were used for all models for a fair comparison. In all the experiments, the Adam optimizer (learning rate = 0.0001, batch size = 32, 15 epochs) and the categorical cross-entropy loss function were employed. To assess convergence and model generalization, the validation accuracy and validation loss were monitored.

Performance Evaluation

The proposed models are evaluated by the use of accuracy, precision, recall, f1 score and confusion matrices. Accuracy evaluated the overall classification performance, precision, recall and F1-score were used for class-wise evaluation, especially for an imbalanced dataset, TrashNet. The confusion matrices were further used to select misclassification pattern between similar visually similar waste types.

Explainability Using Grad-CAM

Grad-CAM was incorporated to enhance the interpretability of the model by producing saliency maps that show the regions of the image that are important for each prediction. The visualizations were employed to evaluate feature localization, prediction accuracy, and model attention to confirm that the models paid attention to the appropriate waste objects and not background information.

Methodological Workflow

The overall procedure of the proposed framework is shown in figure 4.1 starting from the data preprocessing till the comparative performance analysis and Grad-CAM analysis.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

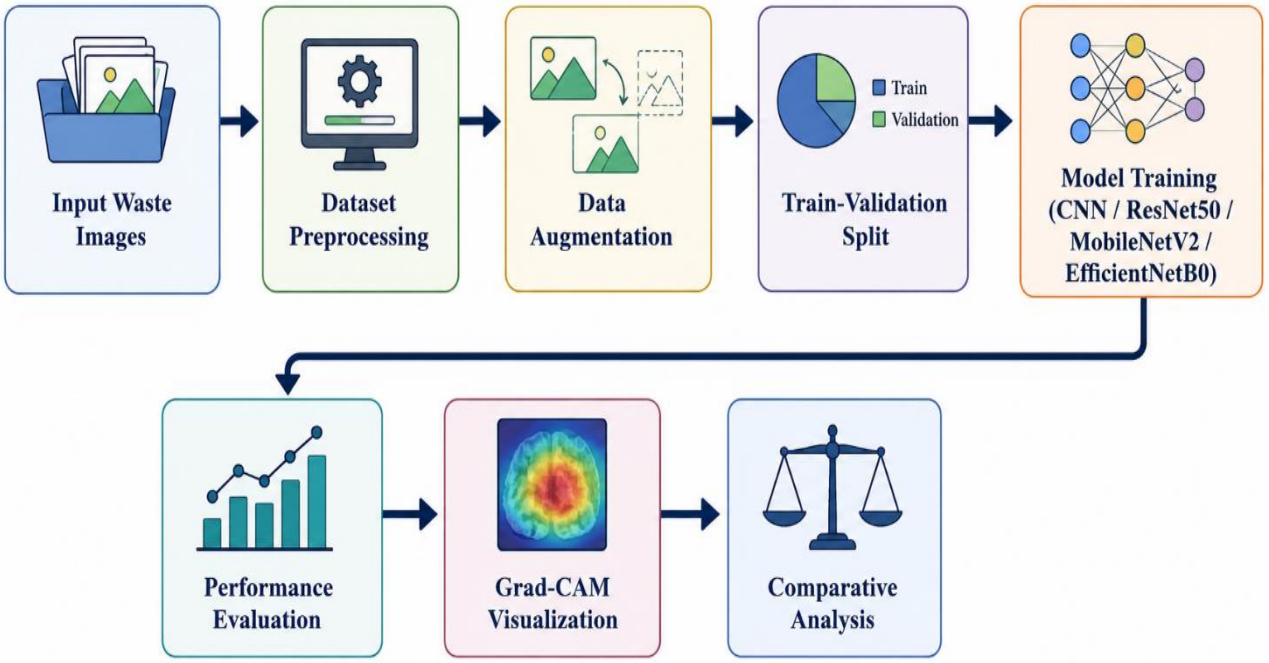

Figure 4.1: Methodology Workflow

Experimental Setup

Development Environment

The experiments were carried out in Python 3.11 with TensorFlow and Keras in Visual Studio Code (VS Code) on a Windows operating system. Transformers TensorFlow and Keras were chosen because of their wide support for transfer learning, image processing, model evaluation, and implementing Grad-CAM [29]. In none of the experiments can you find any support of the native GPU, so all the calculations were done on CPU. The experimental conditions were the same for all models to be correctly compared.

Experimental Configuration

This study uses the following software frameworks, libraries and training configuration, which are summarized in Tables 5.1 and 5.2. The deep learning models [2]-[4] have been implemented using TensorFlow and Keras and the evaluation metrics have been calculated using Scikit-learn. The packages NumPy, OpenCV, PIL, Matplotlib, and Seaborn were used for numerical computation, image processing, and visualization. All of the transfer learning models used pre-trained weights on ImageNet with the frozen convolutional layers, and the classification layers were to be trained. The feature extraction method made the system cost-effective and prevented overfitting in the small-size TrashNet dataset.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 5.1: Software Frameworks and Libraries

Tool / Library	Purpose
Python 3.11	Core programming language
TensorFlow 2.21.0	Deep learning framework
Keras	High-level neural network API
NumPy	Numerical computations
Matplotlib	Graph plotting and visualization
Seaborn	Confusion matrix visualization
Scikit-learn	Evaluation metrics
OpenCV	Image processing operations
PIL (Python Imaging Library)	Image loading and manipulation

Table 5.2:

Parameter	Value
Image Size	224 × 224
Batch Size	32
Epochs	15
Learning Rate	0.0001
Optimizer	Adam
Validation Split	20%
Loss Function	Categorical Cross-Entropy
Activation Function	ReLU / Softmax

Hyperparameter Configuration

Experimental Consistency

The same preprocessing pipeline, train-validation split, data augmentation and hyperparameters, and the same evaluation metrics were used for all models. To enhance reproducibility a fixed random seed was used. When implementing the Grad-CAM, the heatmaps generated by the model-specific preprocessing functions were more reliable, and the predictions with higher confidence values, showing that consistency in preprocessing is crucial to the success of transfer learning and explainable AI [35].

Experimental Results and Comparative Analysis

Overview of Experimental Findings

This section provides experimental comparison of the proposed waste classification

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

framework with a baseline CNN, and three transfer learning CNN architectures namely MobileNetV2, ResNet50, and EfficientNetB0. The processing, augmentation and hyperparameters settings were the same for the same models, so that a fair comparison could be made. The result is shown that the transfer learning is able to enhance the accuracy of classification, feature extraction and model generalization largely than a traditional CNN. Overall, EfficientNetB0 and ResNet50 were the most successful architectures with EfficientNetB0 closely trailing in performance.

Comparative Performance Analysis

The baseline CNN achieved the lowest classification accuracy (51%), indicating limited feature learning capability from the relatively small TrashNet dataset. Table 6.1 compares the summary of the overall classification performance of all selected models.

Table 6.1: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score
CNN	51%	0.56	.51	.49
MobileNetV2	76%	0.77	0.76	0.76
ResNet50	83%	0.85	0.84	0.84
EfficientNetB0	84%	0.84	0.83	0.83

Taking advantage of the lightweight transfer learning, MobileNetV2 achieved an overall accuracy of 76%. The highest performance was obtained using ResNet50 (around 83%) and EfficientNetB0 (around 84%), respectively, which illustrates the considerable impact of using pretrained feature representations on classification. EfficientNetB0 outperformed the baseline CNN by about 33 percentage points and MobileNetV2 by about 25 percentage points in classification accuracy. The results obtained in this work are in line with earlier research, which has shown that transfer learning architectures have been more successful than usual CNN models in waste classification tasks [7, 9, 11, 14].

Class-wise Performance Analysis

From the class-wise evaluation, it could be observed that there is a significant difference in the level of classification with respect to the waste categories. The number of training samples, especially paper and cardboard, always led to higher precision, recall and F1-scores across all the architectures. The class having the lowest number of training samples and higher visual variation, however, was the trash class, which had the worst performance.

While all the Transfer learning models outperformed the baseline CNN in the recognition of minority-class, EfficientNetB0 and ResNet50 yielded the most even performance across all the six classes. MobileNetV2 also performed well with much less of computational resources, which is suitable for the lightweight deployment.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 6.2: Class-wise F1-score for All Models

Class	CNN	MobileNetV2	ResNet50	EfficientNetB0
Cardboard	0.64	0.85	0.90	0.88
Glass	0.49	0.77	0.84	0.81
Metal	0.48	0.75	0.79	0.79
Paper	0.57	0.84	0.90	0.93
Plastic	0.36	0.67	0.82	0.79
Trash	0.24	0.43	0.65	0.56

Confusion Matrix Analysis

The confusion matrices clearly indicate that, with the transfer learning approach, the classification behaviour is improving. Figure 6.1 compares the confusion metrices of all models. The baseline CNN resulted in an unacceptable high number of errors between visually similar classes, such as paper-plastic and glass-metal, and cardboard-paper. The transfer learning architectures helped significantly reduce these errors due to their excellent feature extraction capability. However, there were some misclassification patterns that persisted in all the models. Occasionally plastic samples were classified as glass and the reflective characteristics of metal and glass were similar. Similarly, the trash category had a relatively low recall due to its low representation and diversity in appearance. Based on these observations, it is concluded that the automated waste classification still has one major challenge which is similarity in the visual aspect of the waste.

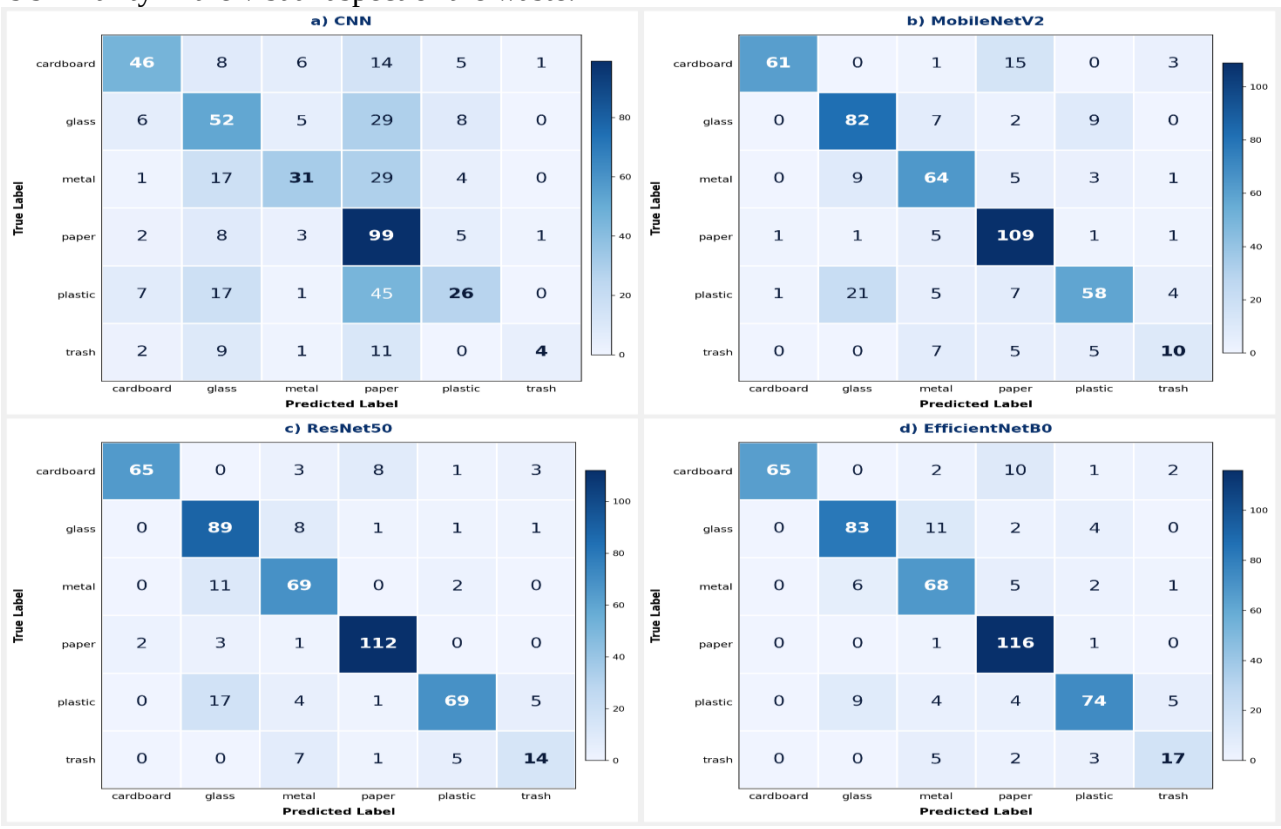

Figure 6.1: Confusion Matrix of All Models

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Error Analysis and Impact of Dataset Imbalance

Results of the experiments show that the imbalance of the data sets clearly had an impact on the model's performance. Minority classes consistently performed worse in terms of recall and stable predictions while majority classes performed better. Especially the class of "trash" had lower recall and F1-scores in all architectures.

In addition to the problem of class imbalance, the presence of overlapping textures in the images and reflective surfaces also caused classification errors. Often, feature representations of similar materials, like cardboard and paper, or glass and metal, were generated and led to inter-class confusion. Transfer learning was able to significantly reduce these errors but not eliminate them. In future, strategies for balanced sampling; weighted loss function; synthetic data augmentation, and larger waste datasets are worth investigating to further increase minority class recognition.

Discussion

The results of the experiments directly answer the research questions posed in Section 1. First, transfer learning yielded significant performance gains in the waste classification task compared to the baseline CNN, as a sign of the effectiveness of the pretrained feature representations. Second, EfficientNetB0 generated the highest classification accuracy and most efficient feature extraction while ResNet50 generated similar results but MobileNetV2 was a practical compromise between accuracy and computational complexity. Thirdly, the class-wise analysis showed that the imbalance of the dataset was mainly for minority class which is trash, even after adopting the transfer learning.

Overall, the experimental results confirmed previous results reported in the literature [7], [9], [11] and [14] and showed that transfer learning is a trustworthy approach for automatic waste classification. In the following chapter, these quantitative results are further substantiated by analysing the discriminative image parts in the model predictions using Grad-CAM.

Explainable AI and Grad-CAM Analysis

Grad-CAM Framework

Transfer learning models had high classification accuracy but it is equally important to understand why it is making the particular prediction, so that the model can be reliably deployed in an intelligent waste management system. To enhance the interpretability of the model, Gradient-weighted Class Activation Mapping (Grad-CAM) was integrated in the proposed model. Grad-CAM is a method that creates visual heatmaps for the parts of image regions that most influence the prediction of a deep learning model, enhancing the transparency of the deep learning models [17-20]. EfficientNetB0 showed to be the most accurate overall classifier, hence it was chosen for analysis of explainability.

We adopted a final convolutional layer of EfficientNetB0 for the implementation of Grad-CAM. Gradients were calculated for each test image associated with the predicted class to create heat maps for the regions that most contributed to the classification decision for each test image. These visualizations were employed to assess the degree of feature localization, a measure of object-centered attention; background influence, a measure of the ability to ignore background information; and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

prediction confidence.

Class-wise Grad-CAM Analysis

Grad-CAM analysis was computed on images from all six classes of waste in order to evaluate the visual attention behaviour of EfficientNetB0.

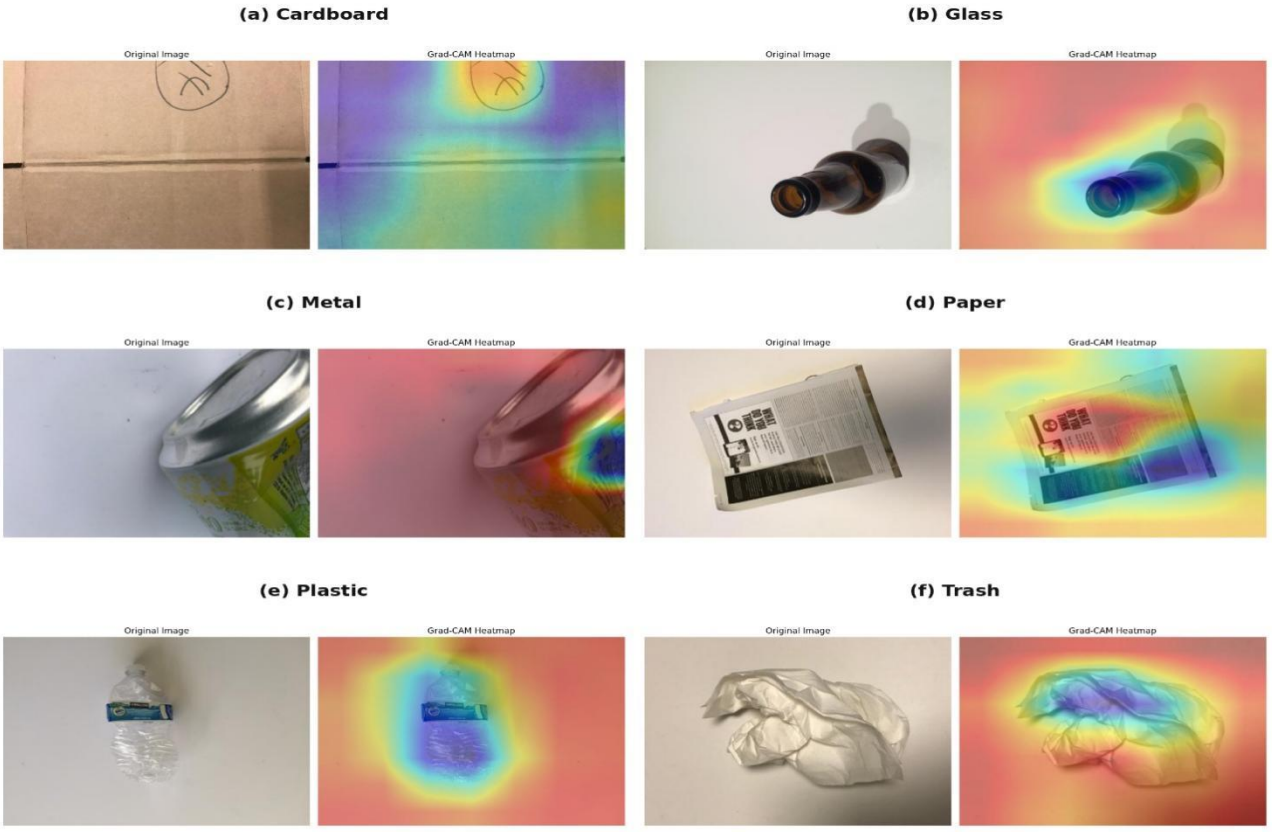

Figure 7.1: Class-wise Grad-CAM Results

Figure 7.1 shows the Grad-CAM visualizations created by EfficientNetB0 for samples of the six TrashNet classes. Reds imply image areas that have the maximum influence on the predicted class; and blue implies image areas that have the least influence on the predicted class. Further, the Grad-CAM visualizations showed that EfficientNetB0 always targeted meaningful regions related to the object being classified, rather than background regions that are irrelevant to the task, which suggests good feature learning and attention localization capabilities. Confidence scores for cardboard, glass and plastic were also high (0.9987, 0.9900 and 0.9848 respectively), indicating that there are identifiable visual features in these categories that can be learnt by the model. Paper (0.7074), trash (0.8410) and especially metal (0.5771) however had comparatively lower confidence as a result of their high intra-class variability, reflective surfaces and high similarity in the visual appearance.

The observations show that the EfficientNetB0 was able to learn semantically meaningful representation from the TrashNet dataset and perform well on localization performance in various waste categories.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 7.1 Class-wise Grad-CAM Confidence Scores

Original Image Class	Predicted Class	Confidence Score
Trash	Trash	0.8410
Glass	Glass	0.9900
Cardboard	Cardboard	0.9987
Metal	Metal	0.5771
Plastic	Plastic	0.9848
Paper	Paper	0.7074

Table 7.1 further substantiates the observations from Grad-CAM. The more confident scores tended to be associated with more focused attention maps, whereas less confident scores yielded more diffuse maps of attention. The correlation between these relationships implies a link between prediction confidence and attention localization and the separability of features within the waste categories.

Effect of Preprocessing on Explainability

Early implementation of Grad-CAM yielded the poorest prediction confidence scores and activation maps were diffuse, with weak object localization, due to the different preprocessing applied. Following the EfficientNetB0 architecture-specific preprocessing function, the prediction confidence has significantly improved, with some classes achieving more than 98% confidence.

The modified preprocessing pipeline produced more focused heatmaps which means stronger object-based attention as well as less interference from the background, and stable prediction behaviour. The results prove that consistency in preprocessing is a crucial factor in the classification accuracy as well as the reliability and interpretability of the Grad-CAM visualizations. Previously, researchers have investigated how preprocessing affects the feature representations of CNNs and CNN explainability [35], [37], [38].

Discussion

The Grad-CAM analysis not only supports the quantitative findings outlined in chapter 6, but also confirms the models' decision-making behaviour visually. The superior classification performance of EfficientNetB0 and reliable feature localization capabilities was achieved with a consistent focus on meaningful object regions. In contrast, attention patterns and confidence were weak for the categories that were more difficult to visually distinguish, for example, metal and trash, due to class imbalance and overlap in visual features.

Overall, the explainability analysis directly tackles RQ4 and RQ5, showing that Grad-CAM produces visual explanations of model predictions and that there is a significant impact on the attention localization and the prediction confidence by applying different type of preprocessing consistency. The results increase transparency and reliability of the proposed transfer learning framework for automated waste classification.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 6 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Conclusion and Future Work

In this study an explainable deep learning approach for automated waste classification was compared to a baseline CNN, using three different transfer learning architectures: MobileNetV2, ResNet50, and EfficientNetB0. The pre-processing, data augmentation, and training procedures were fixed to make a fair comparison between all the models; all the models were trained and evaluated on the TrashNet dataset. The experimental results showed that the transfer learning significantly outperformed the conventional CNN in terms of classification accuracy, feature extraction ability and generalization capabilities. EfficientNetB0 was found to be the best overall model with an accuracy of 84%, with ResNet50 coming in with an accuracy of 83%, and MobileNetV2 offering a good trade-off between accuracy and computational load for resource-limited settings.

The comparative analysis also revealed that class imbalance and visual similarity between waste categories were the two main reasons for misclassification. The categories with low visual diversity and small numbers of training samples (e.g., trash) were always under-performing compared to the well-represented categories (for example paper). However, the class-wise performance of transfer learning models was more stable than the baseline CNN and the inter-class confusion reduced significantly. The incorporation of Grad-CAM to boost the interpretability of automated waste classification was one of the significant contributions of this research. The generated heatmaps validated that the main attention of EfficientNetB0 was not given to meaningless parts of the image, but to relevant parts, visually confirming the classification of the parts of the image. The experiments further showed that constant architecture-specific preprocessing significantly boosted the performance of the Grad-CAM localization method and predicted accuracy, emphasizing the significance of the preprocessing consistency in explainable deep learning processes for architectures.

In general, the results show that applying transfer learning and explainable artificial intelligence is an accurate, transparent and reliable solution for intelligent waste classification. The proposed approach can be implemented in automated waste segregation systems, smart recycling systems, environmental monitoring, and sustainable waste management.

Although these findings were promising, there were some limitations to this study. The experiments have been run on the relatively limited TrashNet dataset, which does not comprehensively represent all variability of real-world waste streams. In addition, even though the transfer learning architectures were very sophisticated, comparable categories like paper/cardboard or glass/metal remained a classification problem.

The proposed framework needs to be assessed on a larger and more diverse set of waste data, and the class-balancing methods and synthetic data generation need to be explored, as well as transformer-based and attention-driven vision models. Furthermore, by expanding the framework to real-time applications at the edge and smart recycling systems, a greater level of practical applicability could be achieved in the context of smart environmental management.

References

[1] Rehman, Muhammad, Imran Kamal Mirza, Fawad Nasim, and M. Arfan Jaffar. "3-Channel Motor Imagery Classification using Conventional Classifiers and

Online ISSN 3007-3197
Print ISSN 3007-3189

Deep Learning Models." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[2] Khan, Shaz Mumtaz, Fawad Nasim, Jawad Ahmad, and Sohail Masood. "Deep learning-based brain tumor detection." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[5] Nasim, Muhammad Fawad, Muhammad Anwar, Almuhannad S. Alorfi, Hamza Awad Ibrahim, Ali Ahmed, Arfan Jaffar, Sheeraz Akram, Abubakar Siddique, and Hafiz Muhammad Zeeshan. "Cognitively inspired sound-based automobile problem detection: A step toward explainable AI (XAI)." International Journal of Advanced and Applied Sciences 12, no. 8 (2025): 1-15.

[6] Aish, Muhammad Abdullah, Amina Abdul Ghafoor, Fawad Nasim, Kiran Irfan Ali, Shamim Akhter, and Sumbul Azeem. "Improving stroke prediction accuracy through machine learning and synthetic minority over-sampling." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[7] Kumar Lilhore, Dr & Simaiya, Sarita & Dalal, Surjeet & Radulescu, Magdalena & Lorente, Daniel. (2024). Intelligent waste sorting for sustainable environment: A hybrid deep learning and transfer learning model. Gondwana Research. 146. 10.1016/j.gr.2024.07.014.

[8] Islam, M. S. B., Sumon, M. S. I., Majid, M. E., Kashem, S. B. A., Nashbat, M., Ashraf, A., ... & Chowdhury, M. E. (2025). ECCDN-Net: A deep learning-based technique for efficient organic and recyclable waste classification. Waste Management, 193, 363-375.

[9] Shukla, Supriya & Sharma, Divyansh & Binjola, Garima & Tyagi, Vanshika. (2026). A Comparative Analysis of Transfer Learning Architectures for Solid Waste Detection. 905-910. 10.1109/IC3ECSBHI67834.2026.11469120.

[10] Manik, F. Y., Nainggolan, P. I., Harumy, T. H. F., Ginting, D. S. B., Maharani, A., Ramadayanti, H., & Harifin, M. P. (2026). Explainable CNN-Based Multiclass Household Waste Classification Using Grad-CAM for Smart Waste Management. International Journal of Advanced Computer Science & Applications, 17(1), 580.

[11] Chen, Y., Sun, J., Bi, S. et al. Multi-objective solid waste classification and identification model based on transfer learning method. J Mater Cycles Waste Manag **23**, 2179–2191 (2021). https://doi.org/10.1007/s10163-021-01283-8

[12] J. Qiu, Y. Zhang, and P. Chen, "An improved EfficientNetV2 model for garbage image classification," Neural Computing and Applications, vol. 37, no. 2, pp. 1123–1137, 2025.

[13] Ngoc-Bao-Quang Nguyen, Tuan-Minh Do, Cong-Tam Phan, Thi-Thu-Hong Phan,
Towards accurate and efficient waste image classification: A hybrid deep learning and

Online ISSN 3007-3197
Print ISSN 3007-3189

machine learning approach, Ain Shams Engineering Journal, Volume 17, Issue 4, 2026,104062, ISSN 2090-4479

[14] S. Kunwar and A. Rai, "Managing household waste through transfer learning and deep learning approaches," Multimedia Tools and Applications, vol. 83, no. 5, pp. 14587–14605, 2024.

[15] Taj, Fatima, Ranya Muqaddas, Adnan Yousaf, Muhammad Fawad Nasim, Maham Naeem, and Eisha Nawaz. "Ensembled Deep Learning (DL) Methods for Detecting Skin Cancer Using CNN Algorithm with Multi Model." In 2025 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1-6. IEEE, 2025.

[16] Y. Yang and J. Thung, "Classification of trash for recyclability status," Stanford University CS229 Project Report, 2016.

[17] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

[18] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," ITU Journal: ICT Discoveries, vol. 1, no. 1, pp. 1–10, 2017.

[19] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.

[20] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.

[21] H. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

[22] M. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale," KI – Künstliche Intelligenz, vol. 34, no. 2, pp. 193–198, 2020.

[23] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLOS ONE, vol. 10, no. 7, pp. 1–46, 2015.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[26] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," Insights into Imaging, vol. 9, no. 4, pp. 611–629, 2018.

[27] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

Online ISSN 3007-3197
Print ISSN 3007-3189

[29] G. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 60, pp. 1–48, 2019.
[30] G. Thung and M. Yang, "TrashNet dataset," Stanford University, 2016. https://github.com/garythung/trashnet
[31] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.
[33] F. Chollet, Deep Learning with Python. Shelter Island, NY, USA: Manning Publications, 2018.
[34] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[35] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[36] C. Molnar, Interpretable Machine Learning. lulu.com , 2022.
[37] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.
[38] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in Proc. Int. Conf. Engineering and Technology (ICET), 2017, pp. 1–6.
[39] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viégas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 2668–2677.