Journal of Engineering, Management and Information Technology

ASPUR PUBLISHING

Vol. 04, No. 01 (2026) 25-36, doi: doi:10.61552/JEMIT.2026.01.003 - http://jemit.aspur.rs

PERFORMANCE OF VISION TRANSFORMER ON GARBAGE IMAGE CLASSIFICATION

Nam Tran Quy [1]

Received 21.08.2024.
Revised 22.09.2024.
Accepted 21.10.2024.

Keywords:

Garbage Image, Vision Transformer, Transfer Learning, Classification, Performance.

Original research

ABSTRACT

This study makes an evaluation of the performance on Vision Transformer model with size of 16x16 words (ViT 16x16) for classifying of garbage images. There are some convolutional neural network (CNN) with technique of transfer learning, namely VGG16, ResNet50, InceptionV3, EfficientNetB7, which are employed for comparison. In each implementation of respective model, the same techniques for image augmentation and hyper-parameters such as, optimizer, activation function and learning rate are employed as the same values among all models. The same dataset of garbage was also applied into those models with the similar splitting on dataset of training, validation and testing. The dataset with 12 different image labels with various kinds of garbage are employed. The experimental results on performance of all models brings the fact that the ViT 16x16 gave the best results at 92%, which is higher the second best model namely VGG16 at 86% and much higher than most of other pre-train models in evaluating garbage images classification.

1. INTRODUCTION

There have been increasing large amount of waste or garbage in every country in the world. The lack of garbage collection leads to harm our life environment. According to United Nation, the total generation of municipal solid garbage would be increased from 2.1 billion tonnes in 2023 up to 3.8 billion tonnes in 2050. In addition, the total cost of management on garbage would also be grown up from smaller number of USD 252 billion in 2020 and this cost could grow up to USD 640 billion by year of 2050 (United Nations Environment Program, 2024). In this situation, AI (artificial intelligence) technology can help human to automatically recognize the garbage which can make our life getting to sustain the environment for further development. In this expectation, the good algorithm to identify garbage images which will be collected from UAV, drone or fixed camera play key role for garbage management, such as garbage collection, forecasting garbage exposure, warning the polluted environment. The highly believable methodology on garbage classification is an important stage in recycling of garbage disposal. The stage of monitor and analyse collect garbage images automatically is an effective way to warn the citizen and governmental or environmental agencies toward the garbage threat on environment and kick-off the right actions or at least attract public attention.

In this paper, the study implements the transfer learning with pre-trained models to test the performance of traditional CNN and Vision Transformer (ViT) on a garbage image classification. The study implemented 4 traditional CNN architectures, namely VGG16, ResNet50, InceptionV3, EfficientNetB7 and another model in addition of ViT 16x16 for garbage image

[1] Corresponding author: Nam Tran Quy
Email: namtq@dainam.edu.vn

classification. The following paragraph will discuss all details of this study.

2. LITERATURE REVIEW

In this part, the study reviews and searches for other similar works which were using machine learning to identify garbage or waste images. There is one of the typical papers from Li and Liu (2023) who proposed a classification model in which the density function was changed a little. The final model was tested for image classification. The methodology tried to solve the problems of classified models, such as over-fitting, low convergence, and low recall and accuracy of traditional related algorithms. The authors employed some techniques such as neural dropout to overcome over-fitting, employ the optimizer of “adagrad” and Relu activation function for the gradient dispersion. Their experiment outcomes depicted that the author’s algorithm achieved high rate of convergence, recall and also accuracy. In another research, Yulita et al. (2024) tested garbage classification based on features of respective images. They used a CNN transfer learning model of Inception V3 to extract the visual information. And then they employed some machine learning algorithm, namely Extreme Gradient Boosting (XGBoost), Naive Bayes, AdaBoost, and Random Forest to classify visual distribution by image classes. The results came out that the combination of the Inception V3 and XGBoost algorithms produced higher efficiency than other respective combination.

Hossen et al. (2024) classified a lot of different waste categories via a multi-stage machine learning. The author set name of model as Garbage Classifier Deep Neural Network (GCDN-Net) which can classify both single-label and multi-label image labels. The model employed a technique of cross-validation for better accuracy generalization to categorize some types of garbage, including 4 labels of “bulky waste”, “garbage bags”, “cardboard”, and “litter”. The images are both kind of separate images or mixed images with combination. The performance of GCDN-Net resulted the achievement of accuracy at 95.77%, precision rate of 95.78%, 95.77% on recall, 95.77% of F1-score to classify the single-label waste images.

Regarding the multi-label classification, GCDN-Net produced the value of mAP (Mean Average Precision) at 0.69 and value of F1-score of 75.01%. In another noticeable paper, Zhang et al. (2021) implemented a DenseNet169 model to classify waste image. They used a famous dataset, namely NWNU-TRASH, which was employed and their outcomes showed the accuracy of DenseNet169 via transfer learning on image classification was 82% and they asserted it was more accurate than other algorithms in their research. Regarding application of Vision Transformer model, Huang et al. (2021) implemented the model of Vision Transformer. They also implemented on TrashNet dataset. Their results came out with achievement of 96.98% value on accuracy. Alrayes et al. (2023) implemented a model which also proposed method to employ Vision Transformer with Multilayer Hybrid Convolution Neural Network (stand by a short name as VT-MLH-CNN). The implementation of the author’s proposed method proved that the VT-MLH-CNN model improved the accuracy and reduced the time on problem of waste classification. The testing implementation on the TrashNet dataset depicted that the achievement of classification accuracy was up to 95.8%, which was 5.28% and 4.6% greater than other techniques, according to the authors in their paper’s comparison. Liu et al. (2023) carried out the waste separation by deep learning on garbage image of paper, glass, fruit, scrap metal equipment, and other materials. After some experiments, the authors claimed that the accuracy of ResNet-50 was lower than ViT, most likely due to its less extensive set of parameters. And the tested results showed that the more the parameters of the model would lead to the higher the precision of training.

3. METHODOLOGY

This study tries to implement totally 5 advanced image recognition models on garbage image classification.

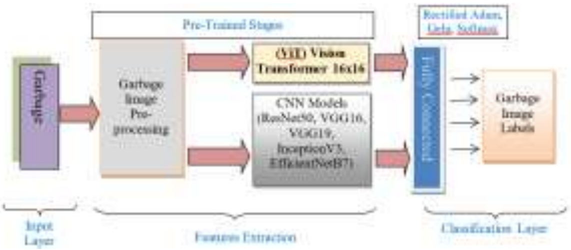
Pre-Trained Stages
Rectified Adam, Gelu, Softmax
Garbage
Garbage Image Pre-processing
(ViT) Vision Transformer 16x16
CNN Models (ResNet50, VGG16, VGG19, InceptionV3, EfficientNetB7)
Fully Connected
Garbage Image Labels
Input Layer
Features Extraction
Classification Layer

Figure 1. Methodology to conduct research on garbage image classification

The methodology is implementation of transfer learning technique. The models chosen for implementing are namely Vision Transformer, VGG16, ResNet50, InceptionV3 and EfficientNetB7.

The study uses the transfer learning via other famous pre-trained models. The study does not make fine-tune to originally test their performance for comparison with ViT 16x16. The models are frozen all parameters in their previous layers except last layers for classification respective to our dataset. The research employed not only identical dataset, but also identical training, validating and testing dataset that are divided by all experimental models. This aims to get similar configuration to evaluate the performance of the model on the same problem.

Figure 1 shows the methodology of experimental implementation. In which, the dataset firstly was put into preprocessing stage. In which, the technique of data augmentation, data cleaning, resizing was employed to make the appropriate dataset for testing in all 5 models for image classification. The continuing paragraphs below will discuss in details all the techniques for implementation, and also will describe the step by step of those 5 testing models. In addition, the respective result will be shown after implementation of each model.

3.1. Preprocessing Dataset

This study employs the dataset published by Mostafa (2020) that contained totally 15,150 images. As description by author namely Mostafa (2020), the author had collected most of the images in this dataset by web scraping, by trying to get images close to garbage images whenever possible, for example in biological garbage category the author searched for rotten vegetables, rotten fruits and food remains, etc. However, for some classes such as clothes or shoes, the author collected images of normal clothes. It means that the author, namely Mostafa put much effort to group garbage images from many sources. The author mentioned some from the Clothing dataset, Garbage Classification dataset and Web Scrapping. The author published dataset on Kaggle (Mostafa, 2020). Figure 2 below depicts a few examples of images from this dataset.

Figure 2. Example of household garbage images dataset

This Mostafa's dataset has 12 labels of image classifications (see Figure 3). The 12 labels comprise of: paper, cardboard, biological, metal, plastic, green-glass, brown-glass, white-glass, clothes, shoes, batteries, and trash.

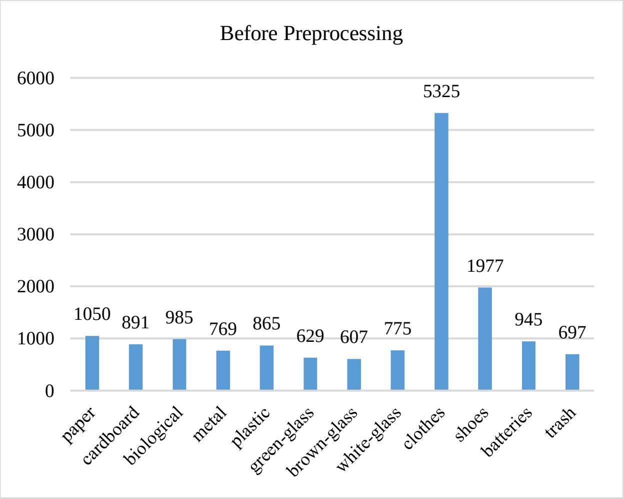

Figure 3. Original dataset by labels

The dataset shows that it the dataset is heavily imbalance (see Figure 3) since the number of labelling images are different. There are some image labels which have much large number but there are many labels with smaller number of its respective images. For example, the number of images belong to clothes and shoes classes, up to 5,325 images of clothes and 1,977 of garbage at shoes, respectively.

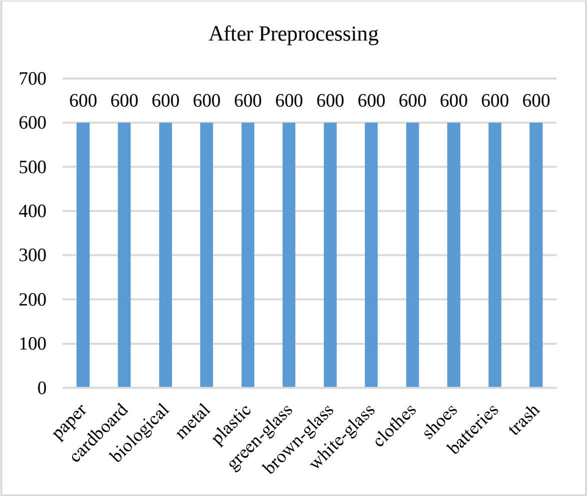

Figure 4. Converted dataset by labels

In order to make this dataset become balance, this study drop out the number of images belong to each label which is over 600 images. As the results, after dropping stage, all the classes of 12 labels have the same number for image are 600 images per class (see Figure 4).

3.2. Data augmentation and hyper-parameters

In order to compare the performance among models, the study divided the preprocessed dataset with 7,200 images of garbage into 3 parts, 80% allocated for training, aallocated 10% for the validating and finally reserve 10% for the testing. All of these 3 parts are randomly division. For each of training dataset, validation dataset and test dataset, the study employs data augmentations to produce more and more images to test the better performance. The first technique is to randomly flip left, right and up, down the direction of random images or transpose the images. The second technique is to randomly rotate 270º, 180º, 90º, and finally make pixel level transforms by adjusting the saturation, contrast, brightness of RGB images with a random factor.

The study carries out the technique of transfer learning to test the accuracy on the above garbage image dataset after preprocessing as mentioned. All the models of Vision Transformer (ViT 16x16), VGG16, ResNet50, InceptionV3 and EfficientNetB7 are implemented with freezing their previous layers except the last layers with fully connected layer and usage of Softmax function for classification of 12 classes on garbage images. In order to appropriate comparison, the models are configured with the same hyper-parameters setup similarly for all tested models.

Table-1. Same hyper-parameters settings

Variable	Values
Size of garbage images	224x224
Training Epoches	50
Learning rate (up to 15 epoch)	0.001
Learning rate (15 to 30 down 0.05%)	lr * 0.95
Learning rate (30 to 50 down 0.1%)	lr * 0.9
Optimizer	RectifiedAdam
Batch size	64
Classes	12
Monitor	Validation loss
Patience to stop training	5

The study set the loss function as Cross Entropy Loss, the optimizer is Rectified Adam, the learning rate equal 0.001 value and learning rate itself would decrease 3 times as described in the Table 1.

4. IMPLEMENTATION AND RESULTS

The folowing paragraph will describe shortly the architectures and the outcomers from all the models of Vision Transformer (ViT 16x16), VGG16, ResNet50, InceptionV3 and EfficientNetB7 on preprocessed dataset of garbage images.

4.1. Experiments with Vision Transformer (VIT)

In the year of 2021, Dosovitskiy (2020) proposed Vision Transformer or shorten name as ViT for image processing. ViT is based on the application of Transformer architecture which is usually implemented in natural language processing. The original ideas came from group of researchers from Google Research. They introduced the vision transformer architecture (see Figure 5) for the image. In fact, this architecture had achieved substantial advantages over other SOTA architecture in terms of many problems.

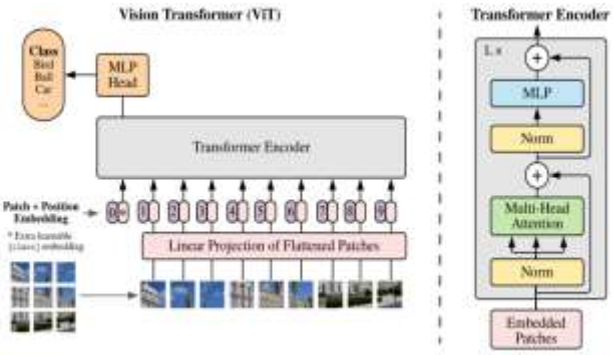

Figure 5. Vision Transformer architecture *(Source: Dosovitskiy 2020)*

As description in the paper by Dosovitskiy (2020), ViT architecture splits an image into fixed-size patches, for example size of 16x16, and put the patches were linearly embedded each of that areas. The authors also added some position scales, and put them into the value of sequence of vectors, then convert them into a standard transformer encoder. The authors employed the approach to add some further learnable parameters as they called name as "classification token" to the sequence for better classification. The special thing is that the transformer architecture employs the mechanisms of the attention mechanism among splitted fixed-size patches, as 16x16 scales this this study. This flow would carefully consider the importance of each area of image by 16x16 scaling.

This study implements the model of Vision Transformer 16x16 on our dataset of garbage image which is mentioned above after preprocessing to 600 images per label. The experiments are also employed the techniques of augmentation of images as mentioned above to make more and more images. After applying the ViT 16x16 on the augmented dataset of garbage images, the outputs pf model on training loss and validation loss are shown in Figure 6.

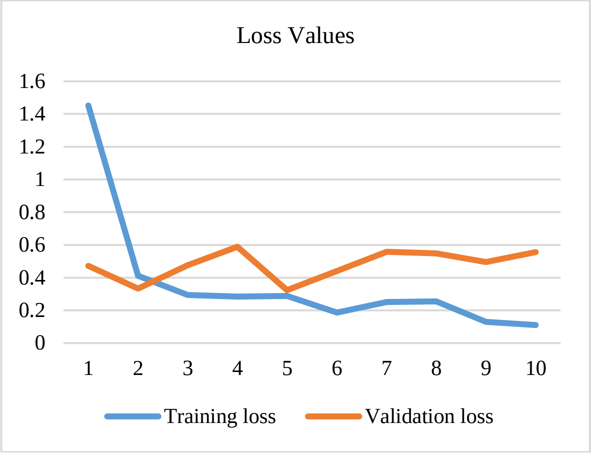

Figure 6. Train and val. loss in ViT 16x16

The results of experiments with Vision Transformer 16x16 on our dataset of garbage image with data augmentation produced the fluctuate of training accuracy and validation accuracy are also depicted in Figure 7.

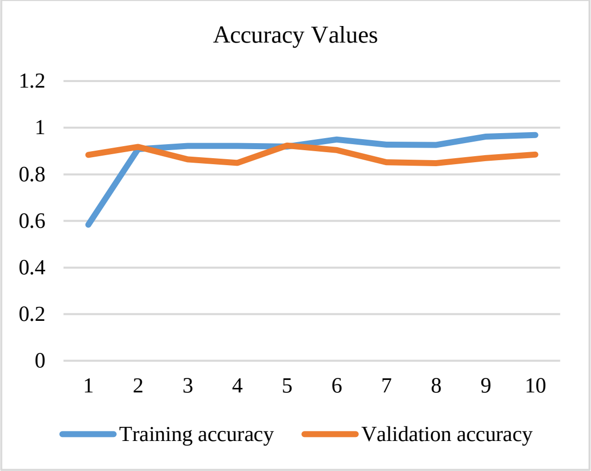

Figure 7. Train and val. accuracy in ViT 16x16

The training process of Vision Transformer 16x16 started with validation loss at 0.4717 and while the study monitors the loss of validation, the training process reaches minimum at epoch 5 with value of 0.3235. After next 5 epochs until epoch 10th, the validation loss did not improve, they still going up, meanwhile our patience was set value at 5 continuous epochs. Therefore, the study stops training and get back the best weight of ViT model at epoch 5th for the best weights of our training process (see Table 2).

Table-2. Validation loss in ViT 16x16

Epoch	Train loss	Train acc.	Val. loss	Val. Acc.
1	1.4521	0.5835	0.4717	0.8835
2	0.4106	0.908	0.3335	0.9176
3	0.2948	0.9214	0.4765	0.8636
4	0.2844	0.9217	0.5886	0.8494
5	0.2876	0.9193	**0.3235**	0.9233
6	0.1856	0.9486	0.4409	0.9034
7	0.2519	0.9278	0.5583	0.8509
8	0.2543	0.9259	0.5491	0.8476
9	0.1298	0.962	0.4948	0.8693
10	0.1103	0.9684	0.5553	0.8849

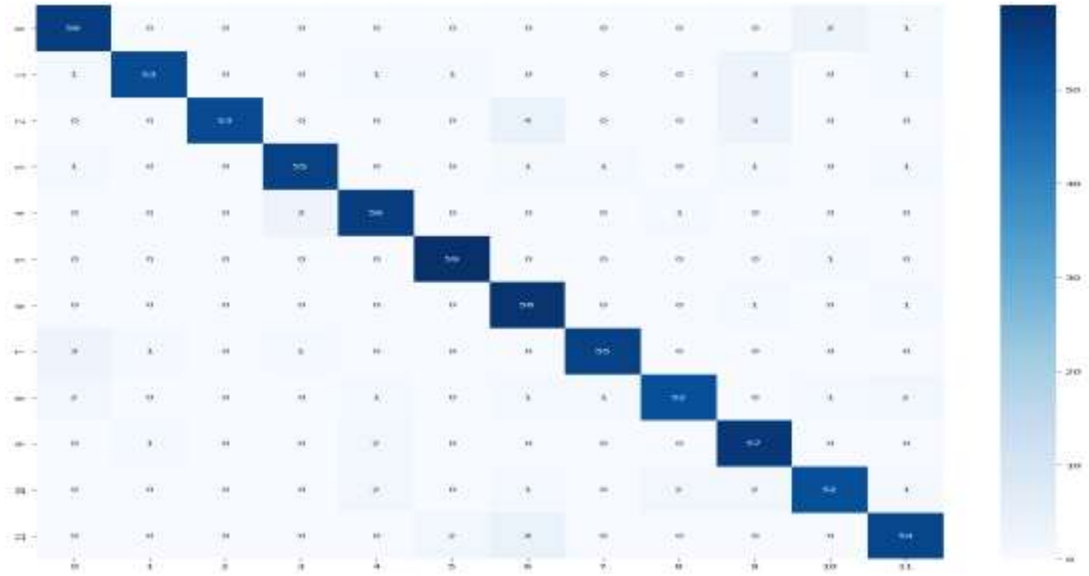

Figure-8. Confusion Matrix of ViT 16x16 on test dataset

In the next step, the study tested the best weights at epoch 5th of Vision Transformer 16x16 which was started on the test dataset.

Table-3. Testing ViT 16x16 on test dataset

Class		Precision	Recall	F1-score	Support
0	green-glass	0.89	0.93	0.91	60
1	metal	0.96	0.88	0.92	60
2	brown-glass	1.00	0.88	0.94	60
3	paper	0.93	0.92	0.92	60
4	clothes	0.90	0.93	0.92	60
5	battery	0.95	0.98	0.97	60
6	biological	0.84	0.97	0.90	60
7	cardboard	0.96	0.92	0.94	60
8	shoes	0.95	0.87	0.90	60
9	white-glass	0.85	0.95	0.90	60
10	trash	0.91	0.87	0.89	60
11	plastic	0.89	0.90	0.89	60
accuracy				**0.92**	720
macro avg.		0.92	0.92	0.92	720
weighted avg.		0.92	0.92	0.92	720

The test dataset is also augmented by the same way to increase the number of images as augmentation of training dataset and validation dataset. We can see in the Table 3 that the accuracy of ViT on classification of 12 classes of garbage image dataset show that value of 92%. In the Figure 8, the experiment on ViT 16x16 shows that the highest number of labels should be mixed is value only at 4 as maximum. There are also some other confusions of labels are valued at 3 which means there are still confusions among these couple of garbage images. In the experimental, the Figure 8 can show that there are some couple which are easily confused. They are cardboard and green-glass, paper and clothes, white-glass and metal, white-glass and brown-glass that are 4 probabilities for confusion at values of 3 labels. The only easy mixture of labels at 4 is between biological and white-glass. Those labels are easy to confuse due their similar characteristics among garbage images.

4.2. Experiments with VGG16

In 2014, namely Simonyan and Zisserman (2014) introduced the VGG network. The architecture of this network model has many different variations: 11 layers, 13 layers, 16 layers and 19 classes. VGG16 refers to a network architecture with 16 layers, while VGG19 refers to a network architecture with 19 layers. The design principle of VGG networks generally includes 2 or 3 convolutional layers followed by a 2D max pooling layer, and then through the last layer which is a flattening layer. This layer aims to convert a 4-dimensional matrix into a 2-dimensional matrix. The next layers are Fully-Connected Layer and a Softmax layer. Since VGG was trained on ImageNet's 1000-class dataset, the final fully connected layer will have a size of 1000. The author Simonyan and Zisserman (2014) in their work had studied the influence of convolutional network depth on results with network accuracy in large-scale image recognition problems. The authors thoroughly evaluated networks of increasing depth using architectures with very small (3x3) convolutional filters. This led to a highly significant improvement which can be achieved over the configurations advanced while making the depth of network to 16-19 layers with weight settings.

When applying the VGG16 (architecture with 16 layers) on our dataset of garbage image with data augmentation produced the fluctuate of training loss and validation loss, validation accuracy as shown in Figure 9.

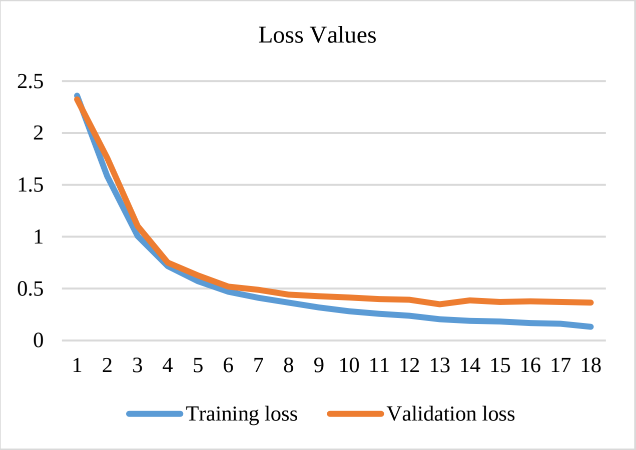

Figure 9. Train and val. loss in VGG16

Meanwhile, the training accuracy and validation accuracy of experiments with the VGG16 on our dataset of garbage image with data augmentation are shown in Figure 10 below.

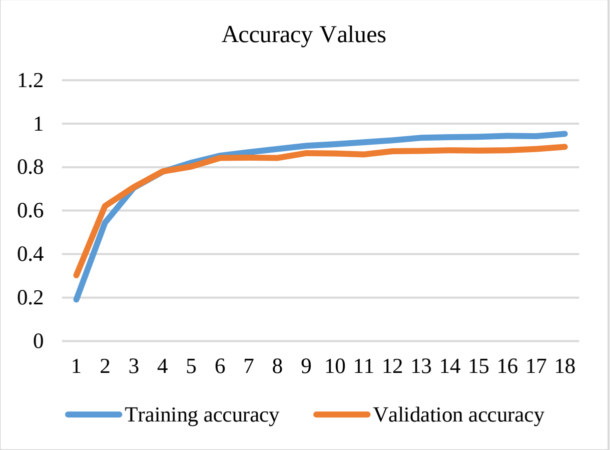

Figure 10. Train and val. accuracy in VGG16

The training process of VGG16 started with validation loss at 2.3240 and reach minimum at epoch 13th with value of 0.3491. After next 5 epochs until epoch 10th, the validation loss did not improve, they still going up, meanwhile our patience set at 5 value. Therefore, the study stops training and get the weight at epoch 13th for the best weights of our training process (see Table 4).

Table 4. Validation loss in VGG16

Epoch	Train loss	Train acc.	Val. loss	Val. Acc.
1	2.3598	0.191	2.324	0.3026
2	1.5833	0.545	1.7603	0.6207
3	1.0073	0.7047	1.1012	0.7088
4	0.7161	0.7785	0.7518	0.7798
5	0.5711	0.8207	0.6261	0.8026
6	0.4705	0.8524	0.5186	0.8423
7	0.4129	0.8687	0.4872	0.8438
8	0.365	0.8832	0.4418	0.8423
9	0.3191	0.8991	0.4285	0.8651
10	0.2831	0.9064	0.4151	0.8636
11	0.2585	0.9144	0.398	0.858
12	0.2393	0.9234	0.394	0.8736
13	0.2038	0.9359	**0.3491**	0.875
14	0.1883	0.9378	0.3874	0.8778
15	0.184	0.9401	0.3701	0.8764
16	0.1681	0.9446	0.3774	0.8778
17	0.1619	0.9431	0.3721	0.8835
18	0.1323	0.9533	0.3649	0.8935

After taking back the weights of VGG16 model at epoch 13th, the study tested the best weights at epoch 13th of VGG16 on the test dataset. The test dataset is also augmented by the same way to increase the number of images as augmentation of training dataset and validation dataset. We can see in the Table 5 that the accuracy of VGG16 on classification of 12 classes of garbage image dataset show that value of 86%.

Table 5. Testing VGG16 on test dataset

Class		Precision	Recall	F1-score	Support
0	green-glass	0.93	0.92	0.92	60
1	metal	0.93	0.90	0.92	60
2	brown-glass	0.86	0.92	0.89	60
3	paper	0.90	0.92	0.91	60
4	clothes	0.83	0.95	0.88	60
5	battery	0.93	0.90	0.92	60
6	biological	0.71	0.70	0.71	60
7	cardboard	0.86	0.92	0.89	60
8	shoes	0.74	0.75	0.74	60
9	white-glass	0.82	0.90	0.86	60
10	trash	0.98	0.82	0.89	60
11	plastic	0.88	0.75	0.81	60
accuracy				**0.86**	720
macro avg.		0.86	0.86	0.86	720
weighted avg.		0.86	0.86	0.86	720

4.3. Experiments with ResNet50

ResNet is a CNN network architecture that was proposed in 2015 by Zheng et al. (2007) from Microsoft research department. ResNet adopted the batch normalization approach on image. The architecture of ResNet was a very deep with up to 152 layers, but thanks to the application of some special techniques, the size of the ResNet50 network only requires about 26 million of parameters, but is still highly effective. The original theory of authors is that, in the previous models, CNN network architectures before ResNet typically improved accuracy by increasing the depth of the CNN network. But the experiments show that at a certain depth threshold, the model's accuracy will saturate and even become counterproductive and make the model less accurate. As going through too many highly depth layers can cause the original information to be lost, the authors from Microsoft researchers have solved this problem on the ResNet network by using a shortcut connection. Accordingly, the network will skip some connections but still keep some information from one layer connecting over next layer, it means the short connections make skipping some middle layers.

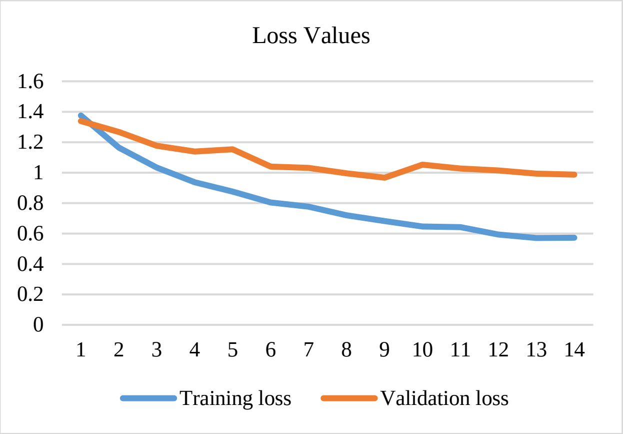

Figure 11. Train and val. loss in ResNet50

The results of experiments with ResNet50 applied on our dataset of garbage image with the same technique of data augmentation and cleaning solutions with previous implementation of ViT 16x16 and VGG16 produced the fluctuation of training loss and validation loss as shown in Figure 11.

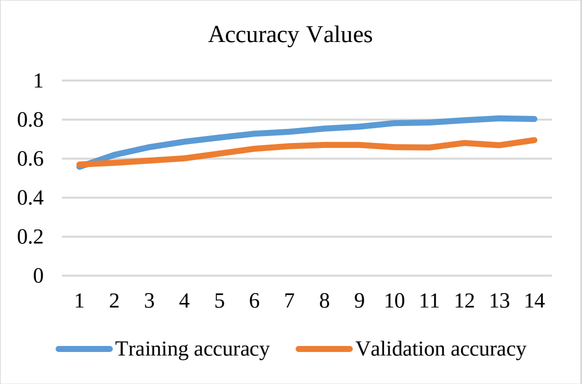

Figure-12. Train and val. accuracy in ResNet50

Meanwhile, the results of experiments with ResNet50 applied on our dataset of garbage image with the same technique of data augmentation and cleaning gave us the training accuracy and validation accuracy as fluctuation in Figure 12.

In the Table 6, the study shows that the training process of ResNet50 started with validation loss at 1.3382 and reach minimum at 9th epoch with value of 0.9666. After next 5 epochs until epoch 14th, the validation loss did not improve, they still going up, meanwhile our patience set at 5 value. Therefore, the study stops training and get the weight at epoch 9th for the best weights of our training process.

Table 6. Validation loss by ResNet50

Epoch	Train loss	Train acc.	Val. loss	Val. Acc.
1	1.3751	0.5576	1.3382	0.5696
2	1.1630	0.6189	1.2671	0.5781
3	1.0341	0.6585	1.1755	0.5895
4	0.9370	0.6858	1.1384	0.6009
5	0.8752	0.7076	1.1540	0.6250
6	0.8042	0.7269	1.0396	0.6506
7	0.7766	0.7366	1.0316	0.6634
8	0.7203	0.7538	0.9946	0.6705
9	0.6815	0.7630	**0.9666**	0.6705
10	0.6460	0.7809	1.051	0.6577
11	0.6424	0.7852	1.0272	0.6562
12	0.5937	0.7965	1.0131	0.6804
13	0.5703	0.8059	0.9935	0.6690
14	0.5721	0.8030	0.9861	0.6946

In the next step, the study tested the best weights of ResNet50 model at epoch 9th of the test dataset. The test dataset is also augmented by the same way to increase the number of images as augmentation of training dataset and validation dataset. We can see in the Table 7 that the accuracy of RestNet50 on classification of 12 classes of garbage image dataset show that value of 66% rate of accuracy.

Table 7. Testing ResNet50 on test dataset

Class		Precision	Recall	F1-score	Support
0	green-glass	0.63	0.67	0.65	60
1	metal	0.91	0.82	0.86	60
2	brown-glass	0.49	0.53	0.51	60
3	paper	0.77	0.82	0.79	60
4	clothes	0.85	0.85	0.85	60
5	battery	0.48	0.50	0.49	60
6	biological	0.52	0.47	0.49	60
7	cardboard	0.81	0.80	0.81	60
8	shoes	0.64	0.48	0.55	60
9	white-glass	0.55	0.68	0.61	60
10	trash	0.61	0.63	0.62	60
11	plastic	0.67	0.65	0.66	60
accuracy				**0.66**	720
macro avg.		0.66	0.66	0.66	720
weighted avg.		0.66	0.66	0.66	720

4.4. Experiments with InceptionV3

In 2016, the authors of Szegedy et al. (2016) proposed a CNN model which scaled up networks in order to utilize the added computation with highly efficient by the technique of suitably factorized convolutions and aggressive regularization. The authors proposed a new architecture with improved performance with the layout of network which is depicted in Figure 1. They used the mechanism of factorization n×n. The Inception architecture transform the input size of each module to become output size of the next module. The authors employed variations of reduction technique to reduce the sizes of grid among the Inception blocks (Szegedy et al. 2016).

The results of experiments with InceptionV3 on our dataset of garbage image with the same technique of data augmentation and cleaning with previous implementation of ViT 16x16, VGG16 and ResNet50 would produce the fluctuate of training loss and validation loss as shown in Figure 13 below.

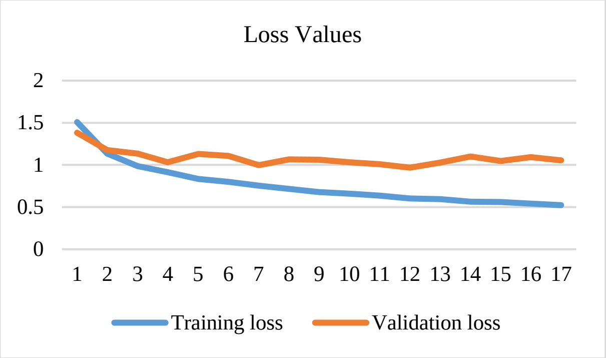

Figure 13. Train and val. loss in InceptionV3

At the same time, the outcome of experiments with InceptionV3 on our dataset of garbage image produced the fluctuate of training accuracy and validation accuracy as shown in Figure 14 below.

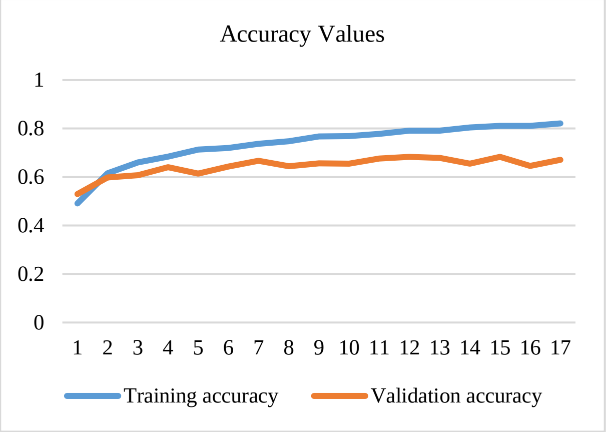

Figure 14. Train and val. accuracy in InceptionV3

The training process of InceptionV3 started with validation loss at 1.3818 in the first epoch and reach minimum at epoch 12th with value of 0.9674. After next 5 epochs until epoch 17th, the validation loss did not improve, they still going up, meanwhile our patience set at value of 5 (see Table 8). Therefore, the study stops training and get back the weight of InceptionV3 at epoch 12th for the best weights of our training process with InceptionV3.

Table 8. Validation loss in InceptionV3

Epoch	Train loss	Train acc.	Val. loss	Val. Acc.
1	1.5075	0.4911	1.3818	0.5298
2	1.1337	0.6153	1.1747	0.5980
3	0.9867	0.6604	1.1323	0.6080
4	0.9132	0.6847	1.0300	0.6406
5	0.8344	0.7135	1.1303	0.6136
6	0.7991	0.7194	1.1092	0.6435
7	0.7542	0.7373	0.9981	0.6676
8	0.7177	0.7479	1.0660	0.6449
9	0.6798	0.7670	1.0600	0.6562
10	0.6607	0.7684	1.0305	0.6548
11	0.6371	0.7786	1.0077	0.6761
12	0.6004	0.7913	**0.9674**	0.6832
13	0.5949	0.7906	1.0276	0.6790
14	0.5629	0.8045	1.1002	0.6548
15	0.5615	0.8111	1.046	0.6832
16	0.5423	0.8113	1.0924	0.6463
17	0.5229	0.8212	1.0553	0.6705

In the following step, the study tested the best weights of InceptionV3 model at epoch 12th on the test dataset. The test dataset is also augmented by the same way to increase the number of images as augmentation of training dataset and validation dataset. We can see in the Table 9 that the accuracy of InceptionV3 on classification of 12 classes of garbage image dataset show that value of 69% of accuracy.

Table 9. Testing InceptionV3 on test dataset

Class		Precision	Recall	F1-score	Support
0	green-glass	0.88	0.72	0.79	60
1	metal	0.85	0.87	0.86	60
2	brown-glass	0.59	0.57	0.58	60
3	paper	0.80	0.75	0.78	60
4	clothes	0.90	0.92	0.91	60
5	battery	0.53	0.50	0.51	60
6	biological	0.56	0.65	0.60	60
7	cardboard	0.76	0.75	0.76	60
8	shoes	0.44	0.53	0.48	60
9	white-glass	0.77	0.85	0.81	60
10	trash	0.68	0.68	0.68	60
11	plastic	0.64	0.53	0.58	60
accuracy				**0.69**	720
macro avg.		0.70	0.69	0.69	720
weighted avg.		0.70	0.69	0.69	720

4.5. Experiments with EfficientNetB7

In 2019, a study published by authors of Tan and Le (2019) which proposed a model of CNN-based EfficientNet. The main idea was based on the observation that CNN models could usually achieve a good result with a certain amount of computational resources. Therefore, in order to increase accuracy, the models often had to make one of the following three directions: increasing model depth; widening each layer; or increasing the quality of image. Instead of that approach, the group of authors of Tan and Le (2019) had presented a new, more balanced approach to extend the CNN model for better accuracy with fewer parameters, increasing the ability to calculate FLOPS (number of number of floating-point operations per second). They proposed their scaling method that concurrently scale up three of them (depth, width and quality) at the same time, still employ a simple and giving combination coefficient of model. In addition, the authors introduced the series of EfficientNet chain from B0 to B7. In which, the configuration of B0 was the basic model and B1-B7 were the expanded model of B0.

This study chooses the EfficientNetB7 to apply on our dataset of garbage image. The dataset of garbage images is employed with the same technique of data augmentation and cleaning with previous implementation of ViT 16x16, VGG16, ResNet50 and InceptionV3. The final results would produce the fluctuate of training loss and validation loss as shown in Figure 15 below.

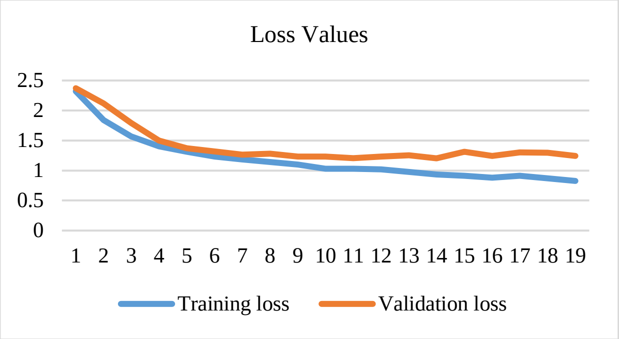

Figure 15. Train and val. loss in EfficientNetB7

Regarding the accuracy of training process, the model of EfficientNetB7 on our dataset of garbage image brought the values of training accuracy and validation accuracy as shown in Figure 16 below.

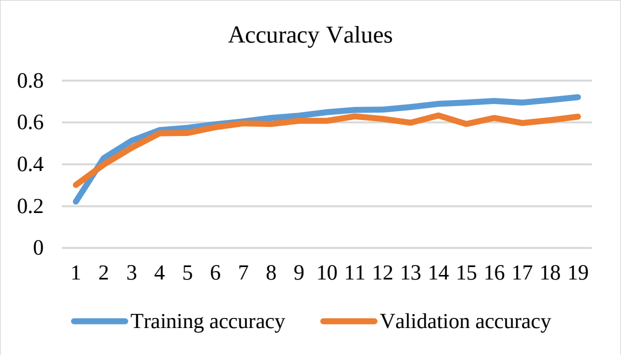

Figure 16. Train and val. accuracy in EfficientNetB7

Looking into Table 10, the training process of EfficientNetB7 started the first epoch with validation loss at 2.3714 and reach minimum at epoch 14th with value of 1.2032. After next 5 epochs until epoch 19th, the validation loss did not improve, they still going up, meanwhile our patience set at value of 5. Therefore, the study stops training and get back the weight of EfficientNetB7 at epoch 14th for the best weights of our training process with EfficientNetB7.

Table 10. Validation loss in EfficientNetB7

Epoch	Train loss	Train acc.	Val. loss	Val. Acc.
1	2.3207	0.2215	2.3714	0.3011
2	1.8413	0.4290	2.1170	0.3977
3	1.5693	0.5137	1.7924	0.4787
4	1.4050	0.5632	1.4987	0.5483
5	1.3145	0.5740	1.3688	0.5497
6	1.2349	0.5915	1.3190	0.5767
7	1.1848	0.6054	1.2656	0.5952
8	1.1434	0.6220	1.2825	0.5923
9	1.0993	0.6330	1.2325	0.6080
10	1.0296	0.6499	1.2335	0.6075
11	1.0318	0.6608	1.2073	0.6293
12	1.0171	0.6618	1.2319	0.6165
13	0.9763	0.6734	1.2555	0.5994
14	0.9346	0.6891	**1.2032**	0.6335
15	0.9143	0.6957	1.3110	0.5923
16	0.8791	0.7023	1.2456	0.6222
17	0.9148	0.6960	1.3004	0.5966
18	0.8722	0.7080	1.2991	0.6108
19	0.8268	0.7210	1.2443	0.6278

In the next step, this study continues to test the model of EfficientNetB7 with the best weights at epoch 14th. The test dataset is also augmented by the same way to increase the number of images as augmentation of training dataset and validation dataset of EfficientNetB7.

Table 11. Testing EfficientNetB7 on test dataset

Class		Precision	Recall	F1-score	Support
0	green-glass	0.78	0.70	0.74	60
1	metal	0.76	0.73	0.75	60
2	brown-glass	0.44	0.47	0.45	60
3	paper	0.66	0.55	0.60	60
4	clothes	0.69	0.70	0.69	60
5	battery	0.37	0.32	0.34	60
6	biological	0.49	0.60	0.54	60
7	cardboard	0.66	0.73	0.69	60
8	shoes	0.50	0.33	0.40	60
9	white-glass	0.72	0.88	0.79	60
10	trash	0.56	0.78	0.65	60
11	plastic	0.67	0.48	0.56	60
accuracy				**0.61**	720
macro avg.		0.61	0.61	0.60	720
weighted avg.		0.61	0.61	0.60	720

The outcomes were shown in Table 11. The study can see in the Table 11 that the accuracy of EfficientNetB7 on classification of 12 classes of garbage image dataset show that value of 61% rate for accuracy.

5. COMPARISON AND DISCUSSION

The measurement on accuracy rate of those 5 models which are all employed on the same garbage image dataset. In order to make suitable comparison, the were some identical hyper-parameters were applied. The final results among those 5 models are depicted in Table 12 below. In which the model ViT 16x16 reaches the best weight at epoch 5th with validation loss at 0.3235 and accuracy at 0.9233. The model of VGG16 reaches the best weight at epoch 13th with validation loss at 0.3491 and accuracy at 0.8750. Meanwhile, the model of ResNet50 reaches the best weight at epoch 9th with validation loss at 0.9666 and accuracy at 0.6705. The model of InceptionV3 reaches the best weight at epoch 12th with validation loss at 0.9674 and accuracy at 0.6832. Finally, the model of EfficientNetB7 reaches the best weight at epoch 14th with validation loss at 1.2032 and accuracy at 0.6335.

Table 12. Performance on accuracy of models

Model	Train Loss	Train Acc.	Valid Loss	Valid Acc.	Test Acc.
ViT 16x16	0.2876	0.9193	0.3235	0.9233	**0.92**
VGG16	0.2038	0.9359	0.3491	0.8750	0.86
ResNet50	0.6815	0.7630	0.9666	0.6705	0.66
InceptionV3	0.6004	0.7913	0.9674	0.6832	0.69
EfficientNetB7	0.9346	0.6891	1.2032	0.6335	0.61

In the Table 12 above, the final result shows that the test accuracy of ViT 16x16 is the highest value at 92% of accuracy on the same testing dataset of garbage image classification. The second highest model measured by accuracy is the model of VGG16 with 86%. The other all remaining models are relatively also not too far behind the accuracy, range around 61% to 69% of accuracy of the same test dataset and same technique of data augmentation. This study can claim that the accuracy of ViT 16x16 is the best among those 5 models which were all implemented on the same garbage image dataset.

6. CONCLUSION

In this study, were employed the transfer learning on some pre-train models for implementation on garbage images. They were namely Vision Transformer (ViT 16x16), VGG16, ResNet50, InceptionV3, and EfficientNetB7. Those all models are implemented on the same training, validation and test dataset of garbage images and also the same data augmentation for 3 datasets of training, validation and testing. The study also keeps the same hyper-parameters such as rectified adam optimizers and learning rate at 0.001 with deceasing by epochs, Gelu activation function to make respectively comparison of accuracy of model. The outcomes of the mentioned experiments show that the Vision Transformer (ViT 16x16) produced the best performance at 92% of accuracy among other models in the problem of garbage images classification. The other future researches may continue to employ Vision Transformer model with various kinds of other image dataset to make confirmation of this hypothesis.

References:

Alrayes, F. S., Asiri, M. M., Maashi, M. S., Nour, M. K., Rizwanullah, M., Osman, A. E., ... & Zamani, A. S. (2023). Waste classification using vision transformer based on multilayer hybrid convolution neural network. *Urban Climate*, *49*, 101483. DOI: 10.1016/j.uclim.2023 .101483.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. DOI: 10.48550/arXiv.2010.11929

Hossen, M. M., Ashraf, A., Hasan, M., Majid, M. E., Nashbat, M., Kashem, S. B. A., ... & Chowdhury, M. E. (2024). GCDN-Net: Garbage classifier deep neural network for recyclable urban waste management. *Waste Management*, *174*, 439-450. DOI: 10.1016/j.was man.2023.12.014.

Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling waste classification using vision transformer on portable device. *Sustainability*, *13*(21), 11572. DOI: 10.3390/su13211 1572

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Li, Y., & Liu, W. (2023). Deep learning-based garbage image recognition algorithm. *Applied Nanoscience*, *13*(2), 1415-1424. 10.1007/s13204-021-02068-z

Liu, J., Sun, J., & Zhou, X. (2023, April). Comparison of ResNet-50 and vision transformer models for trash classification. In *Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022)* (Vol. 12610, pp. 486-491). SPIE. DOI: 10.1117/12 .2671208

Mostafa M. (2020). Garbage Classification (12 classes) - Images dataset for classifying household garbage, Software Developer at KUKA Robots, Augsburg, Bavaria, Germany, Kaggle Dataset, https://www.kaggle.com/datasets/mostafaabla/garbage-classification

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826). DOI: 10.48550/arXiv.1512.00567

Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

United Nations Environment Programme (2024). Global Waste Management Outlook 2024. Retrieved on June 08, 2024 from https://www.unep.org/resources/global-waste-management-outlook-2024

Yulita, I. N., Ardiansyah, F., Sholahuddin, A., Rosadi, R., Trisanto, A., & Ramdhani, M. R. (2024, January). Garbage Classification Using Inception V3 as Image Embedding and Extreme Gradient Boosting. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 1394-1398). IEEE. DOI: 10.1109/ICETSIS61505.2024.10459560

Zhang, Q., Yang, Q., Zhang, X., Bao, Q., Su, J., & Liu, X. (2021). Waste image classification based on transfer learning and convolutional neural network. *Waste Management*, *135*, 150-157. DOI: 10.1016 /j.wasman.2021.08.038

Zheng, N., Loizou, G., Jiang, X., Lan, X., & Li, X. (2007). Computer vision and pattern recognition. International Journal of Computer Mathematics. 84(9), 1265–1266.

Nam Tran Quy
Dai Nam University,
Vietnam.
namtq@dainam.edu.vn
ORCID: 0009-0002-5671-2747