Classifying Wastes Using Random Forests, Gaussian Naïve Bayes, Support Vector Machine and Multilayer Perceptron

Ibrahim F. Hanbal¹, Jeffrey S. Ingosan², Neal Arden A. Oyam³, Yafeng Hu⁴

College of Information Technology and Computer Science, University of the Cordilleras, Baguio City 2600, Philippines.

Email: ¹ibrahimhanbal@yahoo.com, ²jsingosan@uc-bcf.edu.ph, ³oyamheal123@gmail.com, ⁴kenhuyafeng@yahoo.com.ph,

Abstract. As the population of the Philippines increases in size, so does the amount of garbage that it generates. A lot of this garbage is not properly segregated when dumped and many citizens are not aware of the effects of unsegregated garbage which leads to them simply disposing mixed waste in dump sites. A method to help classify garbage can help address this issue. This paper presents a comparison of the efficiency of four commonly used machine learning models, Random Forests, Gaussian Naïve Bayes, Support Vector Machines and Multilayer Perceptron in classifying biodegradable and non-biodegradable waste. The data used for training and testing was collected from various sources and then augmented to increase the size of the dataset resulting in 15,000 images of biodegradable and non-biodegradable trash used for training. The four models were trained using the K-Fold cross validation technique with the dataset being split ten times. The results indicate that the four models were able to achieve a relatively high accuracy in classifying images of biodegradable and non-biodegradable trash with the Random Forests, Gaussian Naïve Bayes, Support Vector Machines and Multilayer Perceptron models achieving an accuracy of 97.49%, 81.46%, 89.51%, and 96.44% respectively.

1. Introduction

Solid waste mismanagement is one of the most serious issues around the world. It can cause damage when not addressed. One way to manage solid waste is segregation. It is done by sorting different types of solid wastes. It is therefore important to identify clearly the type of wastes so that they can be disposed properly. In most developing nations, such as the Philippines, a large portion of the waste generated is composed of organic substances which means the use of incineration to dispose waste is less effective[1, 2]. Properly segregating and managing garbage therefore is needed to ensure that waste is properly organized and handled according to the type of garbage. Many studies have been conducted on the growing amount of waste both in Asia and around the world. A study in India shows that even developing countries like China, India and Brazil, accommodating about 87% of the world population, do not follow segmentation of waste that results into large piles of mixed waste in their dump yard and landfills. The researchers conducted several experimental studies to validate the developed approach and obtained a classification success rate in the range of 85-96%[3].

Another study focused on classifying waste presented a mobile manipulator based system for automated solid waste sorting for recycling. The system is capable of automatically recognizing the recyclable materials by applying bag-of-words and K-means clustering to build a Support Vector
Machine classifier model on images obtained from an on-board thermal imaging camera. The experiment obtained a classification accuracy of 94.3% [4].

Another study presented a hybrid system where a convolutional network is trained to detect and recognize generic objects, and a Gaussian-kernel Support Vector Machine is trained from the features learned by the convolutional network. On the test set, which contains different object instances than the training set, the SVM alone yields a 43.3% error rate, the convolutional net alone yields 7.2% and an SVM on top of features produced by the convolutional net yields 5.9%[5].

One study that used machine learning to classify images of garbage used 5000 images in total in the JPG format. Each waste image is labelled as either recyclable or not for training/testing purpose, [6] and a similar study used k-means clustering and linear regression for classifying garbage. The paper focused on the issue of municipal waste prediction and evaluated a dataset containing 63,000 records [7].

This paper is focused on Baguio City since poor solid waste management is still a big issue within the city especially with the growing population and rising number of tourists coming to the city every year[8]. The study focused on classifying biodegradable and non-biodegradable solid waste using multiple machine learning models and comparing the effectiveness and accuracy of these models with the use of the Scikit-learn module.

The researchers focused on four commonly used machine learning models that are available within Scikit-learn. The first model is the Random Forest Classifier. This model is created by taking a large amount of different samples of data and growing “trees” out of them. After which, if input is given to the “forest”, all the trees vote on what they think the output is and the majority of the vote is considered as the possible output [9]. The second model is the Gaussian Naive Bayes Classifier which is based on Bayes’ theorem with an assumption of independence among predictors where this models assess that the presence of a particular feature in a class is unrelated to the presence of any other feature [10]. The third model is the Multilayer Perceptron Classifier which is a type of neural network. It has the input layer which introduces input values into the network, hidden layers which performs classification of features but hides the result and output layers which functions like the hidden layer but displays the results [11]. The fourth model is the Support Vector Machine (SVM) which is used mainly for binary classification. It finds a hyperplane that can classify new data points by seeing which side of this hyperplane it lands on even if small amount of data is used [12].

The main objectives of the study were to collect a sufficiently large dataset containing images of biodegradable waste and non-biodegradable waste. After collecting the images for the dataset, the next objective was to build and train the models with the use of Scikit-learn. After training concluded, the last objective was to evaluate the models with the use of K-Fold cross-validation.

2. Methods
The study was conducted with the use of the Scikit-learn Python module in order to preprocess, train and test multiple machine learning models onto one set of data. Scikit-learn was chosen due to its state-of-the-art implementations of many machine learning algorithms and ease of use. The module is also distributed under the simplified BSD license, encouraging its use in both academic and commercial settings [13, 14].

2.1. Data Collection Process
The dataset collected for the study is originally composed of 1500 images of garbage. The researchers divided the images into two classes: “biodegradable” and “non-biodegradable” with each having 750 images. These were the chosen classes as these were also used to divide the dataset in previous studies. The images were collected from several sources. First, a select number of the images were taken from the public GitHub dataset of Thung and Yang’s Trashnet, which they used for garbage image classification. The dataset contained 2527 images with six classes: glass, paper, cardboard, plastic, metal, and trash [15]. Around 400 of these images were collected from the six categories and used as part of the non-biodegradable class for this study. Another 300 images were created by the
researchers for the biodegradable class by taking pictures of common food wastes in the local area. The remaining 800 images were taken from online sources such as Google and Bing by using a web crawler. Web crawlers are tools that navigate web pages and download content and metadata from these webpages using an algorithm as opposed to manual browsing [16]. A total of 1,282 images were taken from online sources but a large number of these images were not usable for training as they contained watermarks, text, duplicates and non-garbage subjects. Because of this, they were removed from the dataset which meant only 62.40% of the images downloaded online were usable. In order to ensure that the models were not biased when trained, both classes contained an even number of 750 images.

Previous studies show that having a larger dataset to train a machine learning model generally improves the accuracy of that model. The researchers then used three parameters for the analysis of three features of the images, namely: (1) Color Histogram for colors, (2) Hu Moments for shapes and (3) Haralick Texture for textures. These parameters were used for all four machine learning models and are available within Scikit-learn.

2.2. Data Augmentation Process
The more data an algorithm has access to, the better the performance of that algorithm. As the original dataset was only composed of a relatively small number of images, it needed to be increased in order to properly train the machine learning models. This solution to help expand the dataset and improve the accuracy of the model is data augmentation[17]. Data augmentation enhances data by deriving new data from the base data. Some simple data augmentation techniques include cropping, rotating, zooming in and flipping input images to enhance the dataset [18].

The initial 1,500 images where then augmented nine times each which resulted in 13,500 new images to be added to the dataset, with 7,500 images for biodegradable and 7,500 images for non-biodegradable. The techniques used to modify the images are: (1) increasing brightness by 100%, (2) decreasing brightness by 100%, (3) resizing the image by 50%, (4) using a grayscale filter, (5) using a sepia filter, (6) inverting the colors of the image, (7) flipping the image horizontally, (8) flipping the image vertically, and (9) rotating the image 90 degrees clockwise. The modified images were then combined with the original dataset to create the training and testing dataset used for the models.

![Figure 1. Sample of an original image of “biodegradable” trash and the nine modified versions of the image](image)

2.3. Training and Testing
The four models were trained and tested using the 15,000 images of biodegradable and non-biodegradable trash. The method for the training and testing is the K-Folds cross-validator which splits the dataset into k consecutive folds [19]. For this study, the data was split 10 times for the four models. The data was also split with 90% being part of the training set and 10% being part of the testing set. From the total images in the dataset, 13,500 images were used for training and 1,500 images were used
for testing in each split. The mean results of the training and testing were then displayed for evaluation.

3. Findings
The images where label according to the type of garbage and uploaded for training for the four models. The images were scanned and features of the images were extracted in the process to be used for training and testing the models. After training the 15,000 images of biodegradable and non-biodegradable trash using the K-Folds cross-validator of Scikit-learn with 10 splits, the results of the training are displayed below:

| Machine Learning Models | K-Folds cross-validator accuracy |
|--------------------------|---------------------------------|
| Model                    | Accuracy (Mean) | Accuracy (Percentage) |
| Random Forest Classifier | 0.974949        | 97.49%                |
| Gaussian Naïve Bayes    | 0.814628        | 81.46%                |
| Support Vector Machine  | 0.895151        | 89.51%                |
| Multi-Layer Perceptron  | 0.964429        | 96.44%                |

The results of the training and testing using the K-Folds cross-validator show high accuracy values for all of the four models with the Random Forest Classifier having the highest accuracy and the Gaussian Naïve Bayes having the lowest accuracy for classification.

![Boxplot showing the distribution of data for the four models.](image)

4. Discussions
The results of the K-Folds cross-validator show that the Random Forest Classifier was able to achieve the highest accuracy at 97.49% with the Multi-Layer Perceptron (MLP) neural network achieving a similarly high accuracy with 96.44%. There are several possible reasons for the high accuracy of the Random Forest Classifier. The number of trees used for the Random Forest Classifier was 100 trees, this allowed the classifier to better take into account the features of the different types of trash for both the biodegradable and non-biodegradable classes and spread these features out among the many trees in the classifier. Another factor is the images themselves with most of them being edited or modified versions of the original 1,500 images. This creates a more diverse set of tree structures known as bootstrap aggregating [9]. The MLP neural network with a similarly high accuracy also benefitted from the augmented data set. As neural networks perform better with more data, the MLP was able to achieve a high accuracy with the large dataset of images[20]. To further improve the accuracy of these
models, more data can be collected and augmented to better cover a wider variety of different types of trash.

The Naïve Bayes classifier and Support Vector Machine (SVM) had lower results for their accuracy at 81.46% and 89.51% respectively. The SVM may have had some difficulty creating an optimal hyperplane for the data as many of the images contain a lot of noise in the background. Additionally, by augmenting the dataset to increase the total number of images, this has affected the accuracy of the SVM model as the complexity of the model scales with the size of the dataset [21]. One possible reason for the Naïve Bayes classifier having the lowest accuracy may be that classifying biodegradable and non-biodegradable trash relies on the features of the trash as both these classes have very distinct differences in the images of the datasets [10].

**Table 2.** Comparison of the accuracy of the Four Models with previous studies

| Model                  | Potential Distribution Modelling– Average over 35 datasets [23] | Waste Classification |
|------------------------|----------------------------------------------------------------|----------------------|
| Random Forest Classifier | 91.43%                                                          | 97.49%               |
| Gaussian Naïve Bayes   | 84.14%                                                          | 81.46%               |
| Support Vector Machine | 87.23%                                                          | 89.51%               |
| Multi-Layer Perceptron | N/A                                                             | 96.44%               |

As seen in Table 2, the accuracy of the models in classifying wastes is higher, in the case of the Random Forest Classifier and MLP or quite close, in the case of the Naïve Bayes classifier and SVM classifier.

5. **Conclusions**

The results of the cross-validator show that the Random Forest Classifier and MLP neural network are able to classify images of trash into the biodegradable and non-biodegradable categories with a high accuracy. While the Naïve Bayes classifier and SVM classifier did not reach the same level of accuracy, the two models still showcased a relatively good accuracy. Despite the images of trash having different features, shapes and textures, along with the edited images having different sizes and colours, the augmented dataset was sufficient to train the four models to correctly classify biodegradable and non-biodegradable with a high degree of reliability.

Although the initial set of images only numbered at 750 per class, augmenting the data by editing the images manually and adding them to the original dataset proved to be fruitful as they allowed the four models to achieve a high degree of accuracy after being trained by the expanded dataset containing 15,000 objects of unedited and modified pictures of biodegradable and non-biodegradable trash.

**References**

[1] Alvim-Ferraz M C M and Afonso S A V 2005 Incineration of healthcare wastes: management of atmospheric emissions through waste segregation *Waste Management* vol 25 Issue 6 2005 pp 638-48

[2] Gonzales L B 2016 Urban Sprawl: Extent and Environmental Impact in Baguio City Philippines *Spatium* 1 pp 7-14 DOI:10.2298/SPAT1636007G

[3] Gundupalli S P, Hait S and Thakur A 2017 Multi-material classification of dry recyclables from municipal solid waste based on thermal imaging *Waste Management* 70 DOI:10.1016/j.wasman.2017.09.019

[4] Thakur A, Paulraj S G and Hair S 2016 Automated Municipal Solid Waste Sorting for recycling using a mobile manipulator in ASME 2016 Int. Design Engineering Technical Conf. (IDETC)
At Charlotte, North Carolina, USA Volume 5A DOI:10.1115/DETC2016-59842

[5] Huang F and LeCun Y 2006 Large-scale Learning with SVM and Convolutional Nets for Generic Object Categorization in 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR’06) New York, USA 2006 pp 284-91

[6] Chu Y, Huang C, Xie X, Tan B, Kamal S and Xiong X 2018 Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling in Computational Intelligence and Neuroscience vol 2018 Article ID:506087

[7] Livani E, Nguyen R, Denzinger J, Ruhe G and Banack S 2013 A Hybrid Machine Learning Method and Its Application in Municipal Waste Prediction Advances in Data Mining. Applications and Theoretical Aspects ICDM 2013 Lecture Notes in Computer Science vol 7987 ed P Perner (Berlin, Heidelberg: Springer)

[8] Atienza V 2011 Review of the Waste Management System in the Philippines: Initiatives to Promote Waste Segregation and Recycling through Good Governance in Economic Integration and Recycling in Asia: An Interim Report, Chosakenkyu Hokokusho, Institute of Developing Economies 2011

[9] Oshiro T M and Perez P S and Baranauskas J A 2012 How Many Trees in a Random Forest? Machine Learning and Data Mining in Pattern Recognition MLDM 2012 Lecture Notes in Computer Science vol 7376 ed P Perner (Berlin, Heidelberg: Springer)

[10] Zhang H 2004 The Optimality of Naive Bayes Proc. of the Seventeenth Int. Florida Artificial Intelligence Research Society Conf. FLAIRS 2004 2

[11] Ruck D, Rogers S and Kabriksy M 1989 Feature Selection Using a Multilayer Perceptron in the J. of Neural Network Computing 2

[12] Smola A and Scholkopf B 2004 A tutorial on support vector regression Statistics and Computing (2004) 14 199 https://doi.org/10.1023/B:STCO.0000035301.49549.88

[13] Pedregosa et al 2011 Scikit-learn: Machine Learning in Python JMLR 12 pp 2825-30

[14] Buitinck et al 2013 API design for machine learning software: experiences from the scikit-learn project European Conf. on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (2013) arXiv:1309.0238

[15] Thung G and Mingxiang Y 2016 Classification of Trash for Recyclability Status arXiv Preprint Retrieved from: https://github.com/garythung/trashnet

[16] Shrivastava V 2018 A methodical study of web crawler in J. of Engineering Research and Application ISSN: 2248-9622 vol 8 Issue 11 (Part -I) Nov 2018 pp 1-8

[17] Perez L and Wang J 2017 The Effectiveness of Data Augmentation in Image Classification using Deep Learning arXiv:1712.04627 [cs.CV]

[18] Mikołajczyk A and Grochowski M 2018 Data augmentation for improving deep learning in image classification problem in the 2018 Int. Interdisciplinary PhD Workshop (IIPhDW), Swinoujście 2018 pp 117-22

[19] Raschaka S 2018 Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning arXiv:1811.12808 [cs.LG]

[20] Pal S K and Mitra S 1992 Multilayer Perceptron, Fuzzy Sets, and Classification IEEE Transactions on Neural Networks vol 3 No 5 September

[21] Burges C J C 1998 A tutorial on support vector machines for pattern recognition Data Mining and Knowledge Discovery 2 pp 121–67 (1998)

[22] Khan R, Hanbury A, and Stoettinger J 2010 Skin detection: A random forest approach 2010 IEEE Int. Conf. on Image Processing pp 4613-6

[23] Lorena A C, Jacintho L F, Siqueira M F, De Giovanni R, Lohmann L G, De Carvalho A C and Yamamoto M 2011 Comparing machine learning classifiers in potential distribution modelling Expert Systems with Applications 38(5) pp 5268-75