Artocarpus Trees Classification using Convolutional Neural Network

Ronjie Mar L. Malinao
Marinduque State College
Tanza, Boac, Marinduque, Philippines
ronjieclear@yahoo.com

Alexander A. Hernandez
Technological Institute of the Philippines
Manila, Philippines
alexander.hernandez@tip.edu.ph

Abstract— Distinguishing tree with almost similar features, from trunk, leaf, and fruit is a complex task which often resulted in inaccuracy. This study focuses on the classification of morphologically identical Artocarpus trees locally called "Rimas" (Artocarpus Altilis) and "Kamansi" (Artocarpus Camansi). Unmanned Aerial Vehicle (UAV) is used to capture Artocarpus images. A Convolutional Neural Network (CNN) is utilized to classify two Artocarpus species. The results are models that classify breadfruit and breadnut at the rate of 85.6% and 90% accuracy, respectively. Thus, the CNN classification can effectively identify almost similar trees with an acceptable result.

Keywords—Artocarpus Tree Classification, Convolutional Neural Network, Unmanned Aerial Vehicle

I. Introduction

Artocarpus are flowering trees of Moraceae family. It is popular in the southeast Asian countries and serves as a staple food in some regions. With its approximately sixty (60) species [1] two (2) among them are almost identical. Breadfruit and breadnut are locally known in the Philippines as "Kamansi" and "Rimas." Breadfruit tree (Artocarpus altilis) or popularly known as "Rimas" tastes like a freshly baked potato bread when cooked which got its name breadfruit [1][3]. On the other hand, breadnut (Artocarpus camansi) or popularly known as "Kamansi" serves as a vegetable dish. Additionally, the seed tastes like a chestnut when boiled. [4] At present, there is an increasing utilization of these trees in local food production and processing. Vitamins and minerals could be found in both fruits and could be the answer to food security. [3] [5] Despite their contribution to the food industry and security, they are still considered as neglected and underutilized crop [6]. Recently, the Bureau of Agricultural Research (BAR) funded the development of breadfruit for food and other products. The department believed that breadfruit could be a significant crop for the Philippines and a source of food for many Filipino especially in poverty regions [7]. A "P36 million Rimas Roadmap" project [8] was initiated by the Bureau of Agricultural Research (BAR) that aims to make breadfruit belong the main crop of the country. The fruit can be an alternative food, a source of flour, medicine and it has high economic value. Additionally, the tree can be used as construction materials, fabric, glue, insect repellent, animal feed according to [9]. There are efforts to tally breadfruit tree to estimate local production. However, there are still no reliable figures to answer the queries due to different problems. One of the most basic and frequent problems is the identification or classification of both breadfruit and breadnut tree. [8] Technological advancement like Remote Sensing empowered agriculturist to map, profile and investigate a tree and even the whole forest. That is why Haskell Native American University together with the National Aeronautics and Space Administration initiated a project that focused on breadfruit identification and counting using a satellite [10]. However, this state-of-the-art equipment is not accessible to the ordinary citizen. Fortunately, there is consumers level Unmanned Aerial Vehicle (UAV) that can perform remote sensing tasks. [11] Artificial Neural Network (ANN) have proven its reliability and accuracy since 2015 when Microsoft researchers [12] have developed an ANN that outperformed human performance in image identification.

Thus, this study aims to accurately classify breadfruit and breadnut tree from a vegetation area using images captured by unmanned aerial vehicle and processed by an artificial neural network. The following part is structured as follows: Section II, review of related literature about breadfruit and breadnut, image processing and artificial intelligence. Section III describes the architecture and method of the experiment. In section IV, the tabular results of the classification is presented. Finally, section V presents the conclusion and further improvement of the work.

II. Review of Related Literature

A. Artocarpus Trees

Artocarpus tree is prevalent in the southeast Asian countries and serves as a staple food in some regions. With its approximately sixty (60) species [6] two (2) among them are almost identical. These are locally known in the Philippines as "Kamansi" and "Rimas."

The Breadnut *(Artocarpus camansi)* or locally known as "*Kamansi*" is a tree native in New Guinea, Indonesia, and the Philippines. It became a backyard tree, and there is currently a cultivation area in the Philippines. A mature tree could produce as many as 600-800 fruits per year. [13] The tree grows to 10-15 meters high and its canopy stretch almost half of its height. The branching structures of breadnut are more open. Its leaves are 40-60 cm. long, consist of four to six pair of lobes and sinuses cut halfway to midrib. Fig. 1 shows a sample leaf and fruit of breadnut.

978-1-5386-7767-4/18/$31.00 ©2018 IEEE

E-Vote
Electronic Voting System for Marinduque State College

Ronjie Mar L. Malinao, MIT[1], Eunice G. de Luna, MA.Ed. [2], Jerold Lantoria, MEP [3]
School of Information and Computing Sciences,
Marinduque State College, Tanza, Boac, Marinduque
[1]ronjieclear@yahoo.com
[2]eunicedeluna_23@yahoo.com
[3]sics_2012@yahoo.com.ph

ABSTRACT

Marinduque State College value the right of every student, and one of these is to vote. Participation in election activity is an opportunity for them to help the student organization in selecting student leaders. The researchers developed an electronic voting application called **E-Vote** to improve the process through fast, timely and reliable election. It was designed for both polling and also for the administrator to view and check the result in convenient way. The application enables the students to vote using a computer during the voting period. The principles of E-Vote are uniform and secret, voter's eligibility, secure, reliable and accountable votes. Rapid Application Development was adopted to craft a unique electronic voting application based on the requirements gathered. It was found much faster and more cost effective compared to the manual paper voting system. It lessen the class interruption because the student can vote during their vacant period and lecture rooms were not used as voting precincts during the election. It is recommended for continuous survey made about the feedback of the user towards the application for further improvement.

Keywords:
Electronic Voting, Rapid Application Development, Web-Driven Application, Marinduque State College

E-Vote
Electronic Voting System for Marinduque State College

Ronjie Mar L. Malinao, MIT[1], Eunice G. de Luna, MA.Ed. [2], Jerold Lantoria, MEP [3]
School of Information and Computing Sciences,
Marinduque State College, Tanza, Boac, Marinduque
[1]ronjieclear@yahoo.com
[2]eunicedeluna_23@yahoo.com
[3]sics_2012@yahoo.com.ph

ABSTRACT

Marinduque State College value the right of every student, and one of these is to vote. Participation in election activity is an opportunity for them to help the student organization in selecting student leaders. The researchers developed an electronic voting application called **E-Vote** to improve the process through fast, timely and reliable election. It was designed for both polling and also for the administrator to view and check the result in convenient way. The application enables the students to vote using a computer during the voting period. The principles of E-Vote are uniform and secret, voter's eligibility, secure, reliable and accountable votes. Rapid Application Development was adopted to craft a unique electronic voting application based on the requirements gathered. It was found much faster and more cost effective compared to the manual paper voting system. It lessen the class interruption because the student can vote during their vacant period and lecture rooms were not used as voting precincts during the election. It is recommended for continuous survey made about the feedback of the user towards the application for further improvement.

Keywords:
Electronic Voting, Rapid Application Development, Web-Driven Application, Marinduque State College

Artocarpus Trees Classification using Convolutional Neural Network

Conference Paper · November 2018

DOI: 10.1109/HNICEM.2018.8666322

2 authors:

Ronjie Malinao
Marinduque State College
2 PUBLICATIONS 4 CITATIONS
SEE PROFILE

Alexander A Hernandez
Technological Institute of the Philippines
79 PUBLICATIONS 86 CITATIONS
SEE PROFILE

Artocarpus Trees Classification using Convolutional Neural Network

Ronjie Mar L. Malinao
Marinduque State College
Tanza, Boac, Marinduque, Philippines
ronjieclear@yahoo.com

Alexander A. Hernandez
Technological Institute of the Philippines
Manila, Philippines
alexander.hernandez@tip.edu.ph

Abstract— Distinguishing tree with almost similar features, from trunk, leaf, and fruit is a complex task which often resulted in inaccuracy. This study focuses on the classification of morphologically identical Artocarpus trees locally called “Rimas” (Artocarpus Altilis) and “Kamansi” (Artocarpus Camansi). Unmanned Aerial Vehicle (UAV) is used to capture Artocarpus images. A Convolutional Neural Network (CNN) is utilized to classify two Artocarpus species. The results are models that classify breadfruit and breadnut at the rate of 85.6% and 90% accuracy, respectively. Thus, the CNN classification can effectively identify almost similar trees with an acceptable result.

Keywords—Artocarpus Tree Classification, Convolutional Neural Network, Unmanned Aerial Vehicle

I. Introduction

Artocarpus are flowering trees of Moraceae family. It is popular in the southeast Asian countries and serves as a staple food in some regions. With its approximately sixty (60) species [1] two (2) among them are almost identical. Breadfruit and breadnut are locally known in the Philippines as “Kamansi” and “Rimas.” Breadfruit tree (Artocarpus altilis) or popularly known as “Rimas” tastes like a freshly baked potato bread when cooked which got its name breadfruit [1][3]. On the other hand, breadnut (Artocarpus camansi) or popularly known as “Kamansi” serves as a vegetable dish. Additionally, the seed tastes like a chestnut when boiled. [4] At present, there is an increasing utilization of these trees in local food production and processing. Vitamins and minerals could be found in both fruits and could be the answer to food security. [3] [5] Despite their contribution to the food industry and security, they are still considered as neglected and underutilized crop [6]. Recently, the Bureau of Agricultural Research (BAR) funded the development of breadfruit for food and other products. The department believed that breadfruit could be a significant crop for the Philippines and a source of food for many Filipino especially in poverty regions [7]. A “P36 million Rimas Roadmap” project [8] was initiated by the Bureau of Agricultural Research (BAR) that aims to make breadfruit belong the main crop of the country. The fruit can be an alternative food, a source of flour, medicine and it has high economic value. Additionally, the tree can be used as construction materials, fabric, glue, insect repellent, animal feed according to [9]. There are efforts to tally breadfruit tree to estimate local production. However, there are still no reliable figures to answer the queries due to different problems. One of the most basic and frequent problems is the identification or classification of both breadfruit and breadnut tree. [8] Technological advancement like Remote Sensing empowered agriculturist to map, profile and investigate a tree and even the whole forest. That is why Haskell Native American University together with the National Aeronautics and Space Administration initiated a project that focused on breadfruit identification and counting using a satellite [10]. However, this state-of-the-art equipment is not accessible to the ordinary citizen. Fortunately, there is consumers level Unmanned Aerial Vehicle (UAV) that can perform remote sensing tasks. [11] Artificial Neural Network (ANN) have proven its reliability and accuracy since 2015 when Microsoft researchers [12] have developed an ANN that outperformed human performance in image identification.

Thus, this study aims to accurately classify breadfruit and breadnut tree from a vegetation area using images captured by unmanned aerial vehicle and processed by an artificial neural network. The following part is structured as follows: Section II, review of related literature about breadfruit and breadnut, image processing and artificial intelligence. Section III describes the architecture and method of the experiment. In section IV, the tabular results of the classification is presented. Finally, section V presents the conclusion and further improvement of the work.

II. Review of Related Literature

A. Artocarpus Trees

Artocarpus tree is prevalent in the southeast Asian countries and serves as a staple food in some regions. With its approximately sixty (60) species [6] two (2) among them are almost identical. These are locally known in the Philippines as “Kamansi” and “Rimas.”

The Breadnut *(Artocarpus camansi)* or locally known as “*Kamansi*” is a tree native in New Guinea, Indonesia, and the Philippines. It became a backyard tree, and there is currently a cultivation area in the Philippines. A mature tree could produce as many as 600-800 fruits per year. [13] The tree grows to 10-15 meters high and its canopy stretch almost half of its height. The branching structures of breadnut are more open. Its leaves are 40-60 cm. long, consist of four to six pair of lobes and sinuses cut halfway to midrib. Fig. 1 shows a sample leaf and fruit of breadnut.

978-1-5386-7767-4/18/$31.00 ©2018 IEEE

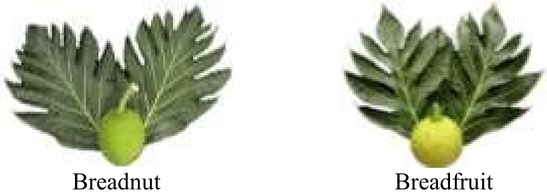


Fig. 1. Sample leaf and fruit of breadnut and breadfruit.

The Breadfruit (Artocarpus altilis) or locally known as “Rimas.” It is incorrectly identified as breadnut due to its characteristics identical to breadnut. [13] This tree grows up to 21 meters but with an average height of 12-15 meters just like the breadnut. The leaves also have 4-6 pairs of lobes but sometimes comes with more deeper sinuses. [14] Fig. 1 shows a sample leaf and fruit of breadfruit.

B. Unmanned Aerial Vehicle

The Unmanned Aerial Vehicle (UAV) or popularly known as “drone,” is flown remotely by a pilot on land capable of capturing birds-eye view images. [15]. Equipped with high definition camera, UAV performs cost-efficient remote sensing device. Several studies prove that UAV help farmers to monitor and manage the agricultural area and vast forest, explicitly monitoring a tree, soil, plant growth in a more effective way [11] [16].

C. Image Pre-processing

Captured images require pre-processing to remove unnecessary noise and convert the images to a form which suitable for machine learning tasks. Thus, numerous image processing methods were utilized by various researchers [15-17]. The image processing includes noise filtering, pseudo-coloring, contrast and edge enhancement, sharpening, slicing, cropping, and image resizing. Matlab is commonly used for image processing. It has image processing toolbox that contains a comprehensive set of algorithms, workflow, analysis, and visualization tool to efficiently classify images.

D. Convolutional Neural Network (CNN)

The CNN focuses on much more rigid machine learning. It is designed to mimic the structure of the brain’s visual cortex. Underneath the cortex is the receptive fields which consist of cells that detect light. The receptive fields become larger as the cell becomes complex in filtering [24]. The CNN has proven its capability on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Researchers in the University of Toronto win [25] the annual ILSVRC using their developed custom CNN, presented in Fig. 2. The CNN is composed of five convolutional layers employing feedforward composed of the following: globally-connected layers, dropout, max-pooling layers, and softmax [26]. Due to its the popularity and capability, a pre-trained CNN was implemented and distributed in machine learning application such as MATLAB. The CNN is called AlexNet and can classify 1000 different categories or classes. [26] Additionally, it can be reconfigured and retrained for a new set of datasets.

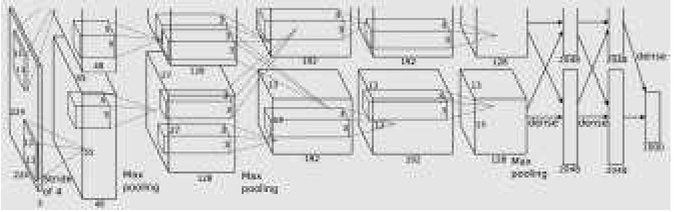


Fig. 2. CNN architecture of A. Krizhevsky et al.

Several studies conducted utilized CNN to classify plants, trees, and crops [27-29]. The results show that CNN yields an acceptable accuracy in classifying plants. Moreover, CNN obtains higher accuracy rate compare to other ANN.

E. F1 Score

F measure of the F1 score is a widely used indicator for evaluating the effectiveness of the machine learning finished model. The F1 score is the mean of recall and precision and assigns both metrics to equal weights. The goal of a good model is to minimize the false negative and false positive [23]. High precision means a low false positive rate. When the model attained 0.78 precision, it is a considerably good model. On the other hand, if the recall is over 0.5, it implies that the model is acceptable.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \qquad (2)$$

III. Methodology

A. System Design

A system architecture is also developed to integrate the classification of breadfruit and breadnut following prior studies [15-17, 29-30], which illustrate essential processes, requirements, algorithms, and theories.

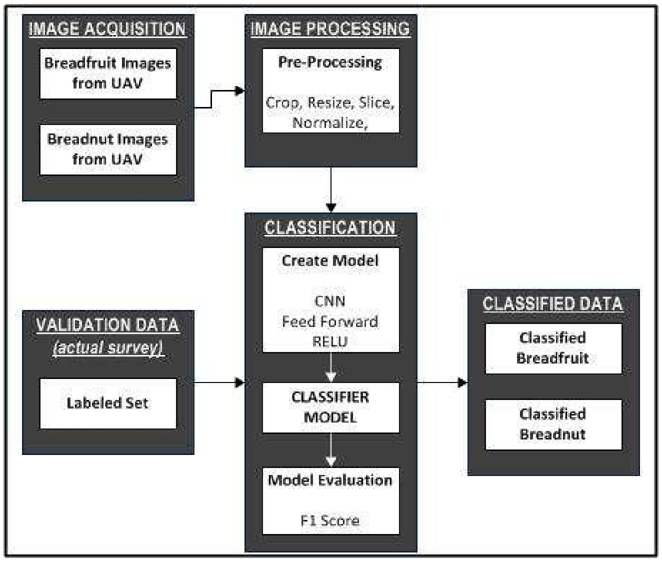


Fig. 3. System Architecture.

Fig. 3 presents the system architecture. Images of the trees were captured by Unmanned Aerial Vehicle (UAV). The tree images were randomly split into a training set, a testing set, and a validation set. The images undergo preprocessing procedure such as resize, crop, and noise reduction. The classification uses the AlexNet [26] convolutional neural network that analyzed the images, employed feature extraction, feedforward, activated by RELU activation function and adjust weights through the recurrent result validation. Another set of validation data was tested in the model for further configuration and improvement of the CNN. The training resulted in a classifier model. The F1 score is utilized in model evaluation, based on the testing set and validation set [23]. As a result, the created and evaluated model identifies a breadfruit and breadnut tree with acceptable accuracy.

B. *Image Acquisition*

Images were acquired using an unmanned aerial vehicle equipped with a 24 mm camera lens and f/2.8 aperture with a resolution of 72 dpi. To ensure image sharpness and prevent from taking blurred images, exposure time set to 1/30 – 1/200 sec. ISO was set to 100 -150 since images were captured during 11:30 AM – 12:30 PM during bright sunny days, these prevent images from shading as recommended in prior studies [15, 17]. Altitude of the UAV was set to 20 m above the ground since the average height of breadfruit, and breadnut tree based on literature [13, 14] was at 15 m.

C. *Dataset*

A total of 2,700 images was used, 900 in each Artocarpus species, 70% of the images were used to train the network, 10% was used to validate the results, and the remaining 20% was used as a testing set (Table I).

TABLE I. DATASET

Species	Training	Testing	Validation	Total
Breadfruit	630	180	90	900
Breadnut	630	180	90	900
Not Artocarpus	630	180	90	900
Total	**1,890**	**540**	**270**	**2,700**

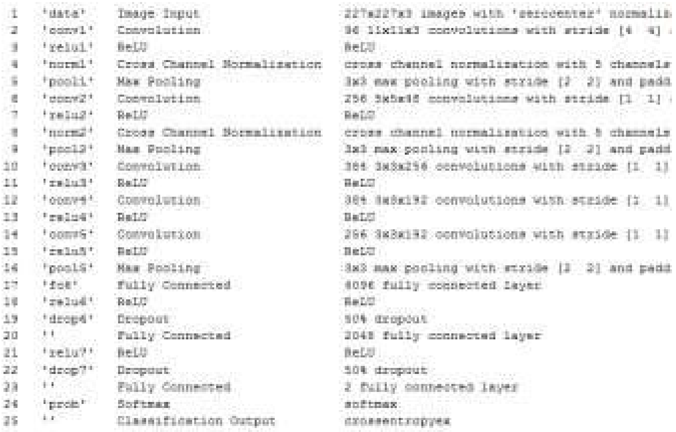
Table I shows the dataset used in this study that consists of 2,700 images of Artocarpus trees. Not Artocarpus images were also included to test the model if it can identify the not Artocarpus. Half of the Not Artocarpus images was also captured by the UAV while the other half came from the publicly available leaf database [31]. The database consists of free natural images and scan-like images from the environmentally controlled area. Additionally, images of grass, soil, roofs and other trees including, narra, mango, coconut, acacia were included as part of the images captured by the UAV.

D. *Image Processing*

Individual images were downloaded from the UAV to a PC and selected to prevent redundancy. The images are cropped to 2951 x 2951 pixels to eliminate unwanted features. Images were sliced into 13 x 13 resulting to 227 x 227 pixel per images. The size of the image was a requirement in the AlexNet CNN [26] and further followed by prior studies [32-34]. Two set of images was then lifted, 30 images from the center were considered as breadfruit or breadnut dataset, while 30 images were selected that compose of the Not Artocarpus dataset. The MATLAB image bach processor module is used to automate the task. The images were compressed using JPEG format. However, images were not transformed into binary images due to the CNN requirement.

E. *Convolutional Neural Network*

After image processing, all the tree images from the dataset were feed into the CNN. The AlexNet pre-trained convolutional neural network was utilized comparable with the studies of [32-34]. An instance of the said network was created and configured in the following 25 layers as shown in Fig. 4.



```
 1   'data'    Image Input                   227x227x3 images with 'zerocenter' normaliz
 2   'conv1'   Convolution                   96 11x11x3 convolutions with stride [4  4]
 3   'relu1'   ReLU                          ReLU
 4   'norm1'   Cross Channel Normalization   cross channel normalization with 5 channels
 5   'pool1'   Max Pooling                   3x3 max pooling with stride [2  2] and padd
 6   'conv2'   Convolution                   256 5x5x48 convolutions with stride [1  1]
 7   'relu2'   ReLU                          ReLU
 8   'norm2'   Cross Channel Normalization   cross channel normalization with 5 channels
 9   'pool2'   Max Pooling                   3x3 max pooling with stride [2  2] and padd
10   'conv3'   Convolution                   384 3x3x256 convolutions with stride [1  1]
11   'relu3'   ReLU                          ReLU
12   'conv4'   Convolution                   384 3x3x192 convolutions with stride [1  1]
13   'relu4'   ReLU                          ReLU
14   'conv5'   Convolution                   256 3x3x192 convolutions with stride [1  1]
15   'relu5'   ReLU                          ReLU
16   'pool5'   Max Pooling                   3x3 max pooling with stride [2  2] and padd
17   'fc6'     Fully Connected               4096 fully connected layer
18   'relu6'   ReLU                          ReLU
19   'drop6'   Dropout                       50% dropout
20   ''        Fully Connected               2048 fully connected layer
21   'relu7'   ReLU                          ReLU
22   'drop7'   Dropout                       50% dropout
23   ''        Fully Connected               2 fully connected layer
24   'prob'    Softmax                       softmax
25   ''        Classification Output         crossentropyex
```

Fig. 4. Network Layers

Fig. 4 shows the composition of the network layers which includes 1 input layer, 1 output layer, 5 convolution layers, 8 activation layers, 5 normalization layers, 3 fully connected layers and 3 layers for max-pooling. Additionally, the training used Stochastic Gradient Descent with Momentum (SGDM) optimizer including options such as momentum 0.9, initial learning rate 0.01, the max epoch of 30, mini-batch size of 128, and validation frequency of 50.

F. *Result Validation*

The performance of the developed model was evaluated using the F1 score similar to the study of [35]. Positive correct results divided by all the positive results is called "precision." While positive correct is divided by all relevant sample is called "recall." The following terms were used to classify the results of the model evaluation result. True Positives (TP) - these were cases in which it predicted breadfruit or breadnut correctly. True Negatives (TN) cases that were correctly predicted not breadfruit or breadnut. False Positive (FP) (Type I Error) cases predicted to be breadfruit or breadnut, but it was not breadfruit or breadnut. False Negative (FN) (Type II error) cases predicted not breadfruit or breadnut, but it was breadfruit or breadnut.

IV. Results and Discussion

A. Image Processing

This study classifies Artocarpus tree images taken by a UAV and processed by the convolutional neural network. Fig. 5 shows the images of Artocarpus and non-Artocarpus tree taken by the UAV. Large images were processed using MATLAB image batch processor. The dimension of each image was 227 x 227 pixel, a resolution of 96 dpi complete with RGB channels and with 24-bit depth. Moreover, redundant images were removed to maintain the 900 images per class.

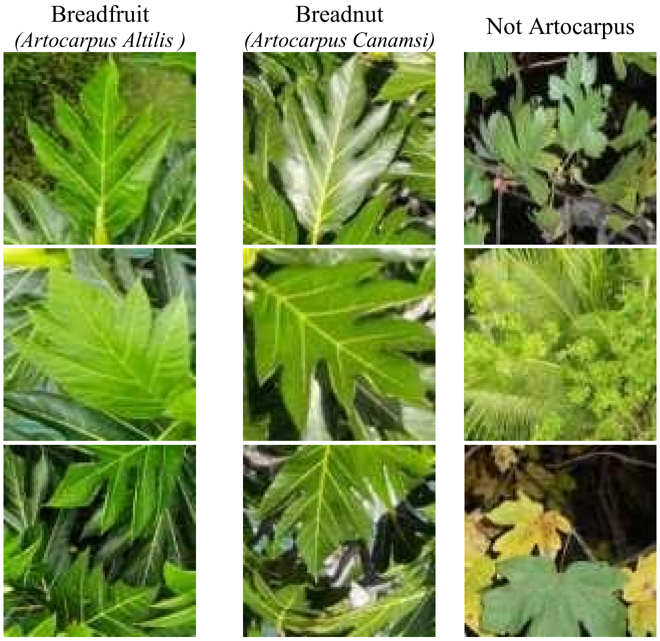


Fig. 5. Sample image dataset: a) breadfruit, b) breadnut and c) not artocarpus

B. Classification

The classification models were created utilizing the AlexNet convolutional neural network. Fig. 6 shows the model training process of Breadnut and Not Artocarpus dataset. It consists of three lines, the blue as the smoothed training line, light blue as the actual training, and the dotted black line denoting the validation accuracy. The validation was conducted every 50 iterations which make the training accuracy dropped to the point where the model learns and reaches the acceptable accuracy. The training lasts for 300 iterations and 30 epoch comparable to [34], and adequate to prevent overfitting.

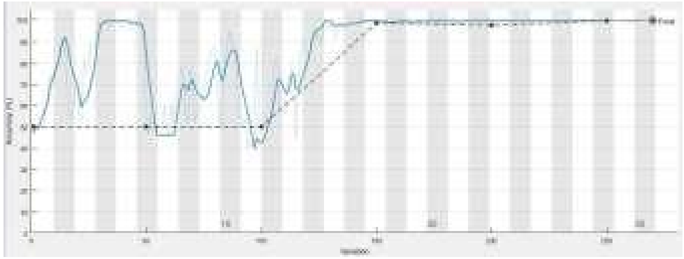


Fig. 6. Model training process (Rx)

There are three models created, listed in Table II. After several modifications and testing, the highest accuracy for each model is saved and further utilized.

TABLE II. Artocarpus Classifiers

Model Name	Dataset	Accuracy
Kx	Breadnut, Not Artocarpus	90%
Rx	Breadfruit, Not Artocarpus	85.6%
RKx	Breadnut, Breadfruit, Not Artocarpus	72.2%

Table 2 shows the three classifiers named, Kx, Rx and RKx. Kx which includes Breadnut and not Artocarpus images, which obtained the highest accuracy of 90%. Rx followed with 85.6%, while the combination of all the dataset yields the lowest accuracy of 72.2%. The naming convention utilizes the local name of the trees which are R for Rimas or Breadfruit, K for Kamansi or Breadnut and x for the Not Artocarpus images.

C. Model Evaluation

After model training and model creation, the model analysis was performed using the F1 score. Table 3 shows the confusion matrix of all models.

TABLE III. Confusion Matrix

Model	Species	True Positive	%	False Positive	%	False Negative	%
Kx	breadnut	319	89%	20	6%	21	6%
Rx	breadfruit	304	84%	42	12%	14	4%
RKx	breadnut, breadfruit	380	70%	91	17%	69	13%

Table 3 shows that the model Kx obtain the highest number of predicted species. Conversely, the model RKx obtained the lowest number of correctly classified images. Additionally, it was noteworthy that the model Rx obtain the least number of False Negative.

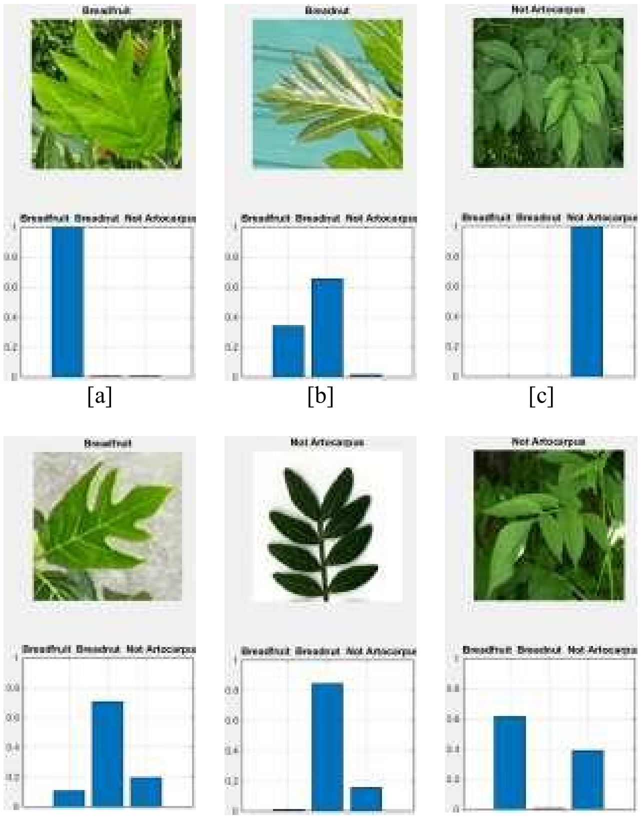
Further validation was conducted (Table 4) with the model Kx tops in the F1 score. Other models also had acceptable F1 scores, which proved that the models were reliable.

TABLE IV. F1 SCORE TABLE

Model	Precision	Recall	F1
Kx	94.10%	93.82%	93.96%
Rx	87.86%	95.60%	91.57%
RKx	80.68%	84.63%	82.61%

D. Classified Data

Test images that were not included in the training set were fed to the models, presented in Fig. 7.



[a] [b] [c]

[d] [e] [f]

Fig. 7. Classified Images

Fig. 7 above shows the classified images, images [a], [b] and [c] were correctly classified. Sample [b] graph denotes that the image was more than 60% breadnut but also around 30% breadfruit and less than 10% classified as Not Artocarpus. Images [d], [e], [f] were samples of misclassified. Breadfruit [d] and Not Artocarpus [e] were misclassified as breadnut. While there were also Not Artocarpus [f] that was misclassified as Breadfruit in the data set.

V. CONCLUSIONS

This paper presents that Convolutional Neural Network is capable of classifying trees with almost identical morphological features. Using images from the uncontrolled environment taken by a UAV in a specified altitude is more than enough to serve as an input to the CNN. Utilizing Convolutional Neural Network fasten the work of the classification by incorporating feature extraction in its layers. Multiple layers of the network incorporate different activation function, fully connected layers, normalization, and pooling. The AlexNet pre-trained CNN does not recognize both trees right away, but with firm training and correct configuration, AlexNet successfully predicts a new set of images. The developed models, Kx (93.96%), Rx (91.57%), and RKx (82.61%) is an excellent contribution to the agricultural sector and the AI community.

However, there are limitations presented in this study. First, all of the accuracy and even the F1 scores are below 95%. Thus, further work could focus on increasing the accuracy. Second, another pre-trained CNN could be considered or developing a new CNN is necessary in future studies. Third, images are still handpicked to eliminate redundancy. Although each area of the image is unique, there is a large area of soil, grass, a roof that is almost identical to each other and could negatively impact the training process. Thus, automatic image reduction process is necessary to be developed. Lastly, a large number of images could translate into a much accurate model.

ACKNOWLEDGMENTS

The following institutions make this initial project possible: Commission on Higher Education, Technological Institute of the Philippines – Manila, Marinduque State College, and Marinduque Provincial Agriculture Office. The authors are also grateful to the anonymous reviewers for the constructive feedback and suggestions to improve the paper.

REFERENCES

[1] Zerega, N. J., Ragone, D., Motley, T. J., "Systematics and species limits of breadfruit (Artocarpus, Moraceae)," In *Systematic Botany*, vol. 30, no. 3, pp. 603-615, 2005.

[2] F. R. Fosberg. "Introgression in Artocarpus (Moraceae) in Micronesia", In *Brittonia*, vol. 2, no. 12, pp. 101-113. 1960. Springer.

[3] Sikarwar, M. S., Hui, B. J., Subramaniam, K., Valeisamy, B. D., Yean, L. K., & Kaveti, B. A Review on Artocarpusaltilis (Parkinson) Fosberg (breadfruit)., In *J App Pharm Sci*, vol. 4, p. 8. 2014.

[4] Lim, T. K., "Artocarpus camansi," In *Edible Medicinal And Non Medicinal Plants*, pp. 304-308, 2012. Springer.

[5] Ong, H. G., & Kim, Y. D., "The role of wild edible plants in household food security among transitioning hunter-gatherers: evidence from the Philippines," In *Food Security*, vol. 9, no. 1, pp. 11-24, 2017.

[6] S. &. S. Deivanai. 2010. "Breadfruit (ArtocarpusaltilisFosb.)–An underutilized & neglected fruit plant species", In *Middle-East Journal of Scientific Research*, vol. 6, no. 5, pp. 418-428.

[7] M. C. O. Fresco. "Growing Breadfruit", In *BAR Chronicle*, Retrieved June 15, 2018, from http://businessdiary.com.ph/11775/growing-breadfruit/ #ixzz4tNRrn4x2. 2002.

[8] The Philippine Star. "DA readies P36 million roadmap for 'rimas", Retrieved June 15, 2018, from http://www.philstar.com/agriculture/2013/03/10/917738/da-readies-p36-million-roadmap-rimas. 2013.

[9] Englberger, L., Aalbersberg, W., Ravi, P., Bonnin, E., Marks, G. C., Fitzgerald, M. H., & Elymore, J., "Further analyses on Micronesian banana, taro, breadfruit & other foods for provitamin A carotenoids & minerals", In *Journal of Food Composition and Analysis*, vol. 16, no. 2, pp. 219-236. 2003.

[10] P. B. C. "Counting ulu from space, Pacific Business Center Program". Retrieved June 15, 2018, from http://www.samoanews.com/linking-samoans/ counting-ulu-space.

[11] Wallace, L., Lucieer, A., Malenovský, Z., Turner, D., & Vopěnka, P., "Assessment of forest structure using two UAV techniques: A

comparison of airborne laser scanning and structure from motion (SfM) point clouds", In *Forests*, vol. 7, no. 3, p. 62, 2016.

[12] He, K., Zhang, X., Ren, S., & Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," In *IEEE International Conference on Computer Vision*, 2015. IEEE.

[13] Ragone, D., "Artocarpus camansi (breadnut)," The Breadfruit Institute, National Tropical Botanical Garden, Hawaii, pp. 1-11, 2006.

[14] Ragone, D., "Artocarpus camansi (breadnut)," The Breadfruit Institute, National Tropical Botanical Garden, Hawaii, 2006.

[15] T. P. Marthinus Reinecke, "The influence of drone monitoring on crop health and harvest size", In NextComp IEEE Mauritius, South Africa. 2017. IEEE.

[16] Nevalainen, O., Honkavaara, E., Tuominen, S., Viljanen, N., Hakala, T., Yu, X., & Tommaselli, A. M., "Individual tree detection and classification with UAV-based photogrammetric point clouds and hyperspectral imaging", In *Remote Sensing*, 9(3), 185. 2017.

[17] Vaughn, N. R., Moskal, L. M., & Turnblom, E. C, "Tree species detection accuracies using discrete point lidar and airborne waveform lidar", In *Remote Sensing*, 4(2), 377-403. 2012.

[18] Wilson, B., "The Machine Learning Dictionary,"University of New South Wales, 2012. Retrieved June 18, 2018, from http://www.cse.unsw.edu.au/~billw/mldict.html#activnfn.

[19] Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y. X., Chang, Y. F., & Xiang, Q. L., "A leaf recognition algorithm for plant classification using probabilistic neural network. In Signal Processing & Information Technology", In *IEEE International Symposium*. 2007. IEEE

[20] Satnam Singh, Manjit Singh Bhamrah, "Leaf Identification Using Feature Extraction and Neural Network", In *IOSR Journal of Electronics and Communication Engineering*, vol. 10, no. 5, pp. 134-140, 2015.

[21] R. Janani, A. Gopal, "Identification of selected medicinal plant leaves using image features and ANN", In *Advanced Electronic Systems (ICAES)*, 2013 International Conference, Page(s): 238 – 242, 2013.

[22] Belavagi, M. C., & Muniyal, B., "Performance evaluation of supervised machine learning algorithms for intrusion detection," In *Procedia Computer Science*, vol. 89, pp. 117-123, 2016.

[23] Powers, D., "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," In *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.

[24] Hijazi, S., Kumar R., Rowen, C., "Using Convolutional Neural Networks," In *IP Group*, Cadence, 2015.

[25] Stanford Vision Lab , "Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)," Retrieved June 18, 2018, from http://image-net.org/challenges/LSVRC/2012/results.html.

[26] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "Imagenet classification with deep convolutional neural networks," In *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[27] Elhariri, E., El-Bendary, N., & Hassanien, A. E., "Plant Classification System Based on Leaf Features", In *Computer Engineering & Systems (ICCES)*, 9th International Conference on pp. 271-276. 2014. IEEE.

[28] Reyes, A. K., Caicedo, J. C., & Camargo, J. E, "Fine-tuning Deep Convolutional Networks for Plant Recognition". In *CLEF (Working Notes*). 2015.

[29] Lee, S. H., Chan, C. S., Wilkin, P., & Remagnino, P, "Deep-plant: Plant identification with convolutional neural networks", In *Image Processing (ICIP)*, IEEE International Conference on pp. 452-456. 2015. IEEE.

[30] Lobell, D. B., Asner, G. P., Law, B. E., & Treuhaft, R. N., "View angle effects on canopy reflectance and spectral mixture analysis of coniferous forests using AVIRIS", In *International Journal of Remote Sensing*, vol. 23 no. 11, pp. 2247-2262. 2002.

[31] ImageCLEF, " Plant identification task 2011", Retrieved June 20, 2018, from https://www.imageclef.org/2011/plants.

[32] Toma, A., Stefan, L. D., & Ionescu, B., "UPB HES SO@ PlantCLEF 2017: Automatic Plant Image Identification using Transfer Learning via Convolutional Neural Networks", In *CLEF (Working Notes)*. 2017.

[33] Ghazi, M. M., Yanikoglu, B., & Aptoula, E., "Plant identification using deep neural networks via optimization of transfer learning parameters", In *Neurocomputing*, vol. 235, pp. 228-235. 2017.

[34] Tan, J. W., Chang, S. W., Kareem, S. B. A., Yap, H. J., & Yong, K. T., "Deep Learning for Plant Species Classification using Leaf Vein Morphometric", In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2018. IEEE.

[35] Yalcin, H, "Plant phenology recognition using deep learning: Deep-Pheno", In *Agro-Geoinformatics*, 6th International Conference on pp. 1-5. 2017. IEEE.

Classifying Breadfruit Tree using Artificial Neural Networks

Ronjie Mar L. Malinao
Technological Institute of the Philippines, Manila
Philippines
ronjieclear@gmail.com

Alexander A. Hernandez
Technological Institute of the Philippines, Manila,
Philippines
alexander.hernandez@tip.edu.ph

ABSTRACT

This is a research-in-progress of designing an intelligent morphological analysis for Artocarpus Altilis or commonly called "breadfruit." This research applied image processing, artificial intelligence (AI) and system design. Using Unmanned Aerial Vehicle (UAV), images are captured, processed and fed to the artificial intelligence for classification. The initial result yields a 75% accuracy using the initial dataset. This study proves that using UAV combined with AI could substantially contribute to the agricultural industry in efficiently classifying breadfruit. This paper recommends further enhancement of the system.

CCS CONCEPTS

•**Machine Learning**→**Learning Paradigms**→Supervised Learning by classification;

KEYWORDS

Breadfruit, Artocarpus, Image Processing, Artificial Neural Networks

1 INTRODUCTION

Breadfruit tree (Artocarpus altilis) or popularly known as "Rimas" in the Philippines, a type of flowering tree of Moraceae family and belongs to Artocarpus genus [1] It tastes like a freshly baked potato bread which got its name breadfruit [2].

Currently, there is expanding the use of breadfruit in food processing and food production. Breadfruit can help to improve health and nutrition and alleviate world hunger [3]. Some usages and significant potential in the food industry, breadfruit is a neglected and underutilized crop [4].

On the other hand, Bureau of Agricultural Research (BAR) is funding the development of breadfruit for food and other products. The department believed that breadfruit could be a major crop for the Philippines and source of nutrient for many Filipino especially in poverty regions. [5]. A "P36 million Rimas Roadmap" project was initiated by Bureau of Agricultural Research (BAR) that aims to make breadfruit belong the main crop of the country. The fruit can be a source of alternative food (alternative for rice), can be a source of flour, medicine and it has high economic value. Additionally, the tree is used in construction materials, fabric, glue, insect repellent, animal feed according to [5].

There are attempts to count the breadfruit tree to estimate local production; there are still no reliable statistics of a number of breadfruits. The following reasons for not having an accurate number of breadfruits are: the trees are scattered in the forest, few data available and poor market research[6], and people are not familiar with the tree characteristics[7] that resulted in errors and inaccurate estimation of production.Since breadfruit is essential for the industry development and food manufacturing, Haskell Native American University together with National Aeronautics and Space Administration initiated a project that focused on breadfruit identification and counting [8].

The Philippine Statistics Authority has recorded, but a total number of trees is still unclear [9]. As stated above, it is necessary to know the number of breadfruit for planning and estimation of food production using breadfruit.

Thus, tree manual classification is complex and consumes time [10]. This study aims to design a solution to the following problem: 1.) difficulty in breadfruit identification 2.) no variety classification of breadfruit tree conducted 3.) insufficient records of existing breadfruits. The study aims to develop a system that identifies breadfruit tree from the location and its variety using an intelligent leaf morphological analysis.

2 RELATED WORK

2.1 BREADFRUIT TREE

There are estimated less than three hundred (300) breadfruit tree found in various provinces of Mindanao while one thousand (1,000) suckers are currently in growing stage. Breadfruit is planted on an expansive one thousand two hundred hectares (1,200 ha) in Bicol Region through the Biodiversity Research, Conservation, and Propagation in Bicol (RBR-CPB) that is previously funded by the Bureau of Agricultural Research (BAR) [11]. Philippine Statistics Authority listed breadfruit as a minor fruit crop however the total number of trees is still unclear [9]. In the upcoming years, thousands of breadfruit tree is subject to profiling and individual evaluation. Thus, a room for digital profiling for ease of assessment will be an essential tool.

2.2 Simple and Morphological Shape Descriptors (SMSD)

Researchers [12] discussed SMSD is referring to leaf geometric shape properties which are centroid, perimeter, area, major axis length, minor axis length, and diameter. Additionally, morphological descriptors from the SMSD was computed and utilized.

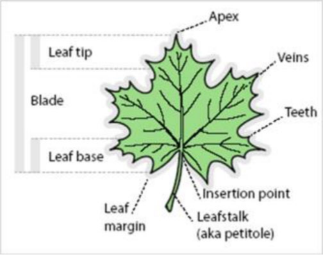


Figure 1: Leaf parts.

ACIT2018, June 2018, Kunming, China

ACM ISBN 978-1-4503-6574-1/18/06 $15.00.
DOI: 10.1145/3265007.3265014

In Fig. 1 shows the parts of a leaf. There are two classifications of shape descriptors; Region-based and Contour based. Features obtain based from the edges of a leaf is called Region-based. On the other hand, features obtained from the lines of a lead is called Contour-based.

2.3 Image Processing Approach

Image processing methods is an advantage for tree identification [14-16]. Using a mobile phone could enhance the recognition rate of a user. A user could enhance plant recognition rate using a mobile phone equipped with a built-in camera and image recognition application. The picture of the plant is uploaded to the application to analyze and identify possible variety and name of the plant. By these non-professionals could easily use the system without having to undergo intensive plant identification training. Researchers urge to dedicate their time to develop plant identification process further.

The process involves the following; 1.) Image Acquisition – retrieving an image of a tree or a leaf for further analysis and classification. 2.) Preprocessing – enhance image data by removing irrelevant part of the image and retaining images that are necessary for further processing. Procedures include; image resizing, image segmentation, image correction and image content enhancement. 3.) Feature extraction & description – measuring an important part of the image. Important parts are labeled by several sets of numbers to characterize the plant property.

2.4 Machine Learning for Leaf Morphological Analysis

From the study [17], machine learning was applied in studying the leaf morphological. Machine learning algorithm performs a task to learn from the examples of leaves in the database to correctly identify and discriminate the leaves. To train the system, 7,597 leaves were manually prepared for morphological analysis. Images were randomly divided into two, the first half is labeled training set, and the other half is labeled test set. The first data set is used to develop a classifier while the latter is used to validate the result which was repeated ten (10) times. The accuracy is based on the averages of those ten repetitions.

2.5 Unmanned Aerial Vehicle

Unmanned aerial vehicle (UAV) or popularly known as "drone,"is the unmanned aerial vehicle which is remotely controlled or autonomously flown by a pilot on land. [21].

A study [22] claims that farmers have benefited using drones by efficiently managing their land. Drone help to determine the area that needed irrigation, monitor plant growth, determine spoil problem, spot fire before it spread, and much more.

3 METHODOLOGY

3.1 System Design

The study aims to develop a system that can identify breadfruit tree from a certain location. Furthermore, it sought to detect breadfruit variety using an intelligent leaf morphological analysis.

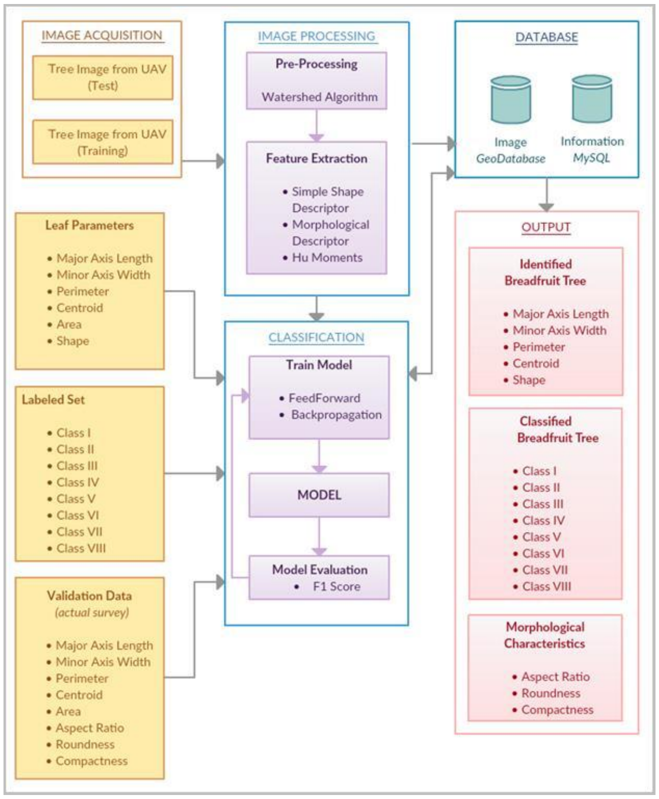


Figure 2: System Design.

Fig. 2 depicts the architecture of the system. Images of the trees are captured by Unmanned Aerial Vehicle (UAV). The collection of tree images was randomly split into a test set and training set. The images undergo preprocessing procedure and then extract using Watershed algorithm. The watershed algorithm used to remove the unnecessary part and extract best features of the image. Morphological and geometrical features of the leaf are extracted using Simple Shape Descriptor, Morphological Descriptors and Hu Moments. Important features such as minor axis width, major axis length, perimeter, centroid, area, shape, aspect ratio, roundness, compactness were extracted and send to the Artificial Intelligence Neural Network (ANN) for classification. Feedforward will learn the descriptors, and consequently, Backpropagation will validate result by calculating errors. Leaf parameters (minor axis width, major axis length, perimeter, centroid, area, and shape) and Labeled set (breadfruit varieties from Class I to IV) based on literature are inserted to train the model. To further increase the accuracy of the model, validation of classification result is conducted based on Validation Data that is observed by a human expert. The F1 score is utilized in model evaluation. The result of the classification is stored in the information database and later displayed to the user. As a result, the system could identify a breadfruit tree in a middle of a forest, determine its exact location and variety. This system architecture is based on the study of the following [23–25].

3.2 Image Acquisition

Aerial image acquisition is the process of photographing the ground from a flying unmanned aerial vehicle at a certain altitude. Capturing an image start at the requirement such as area or object to be captured, this serves as an input in the flight planning. Flight planning set the required elevation, speed, starting point, end point, flight pattern, vertical overlap and horizontal overlap. The settings will be used the following: speed – 20 kph, vertical overlap-70%, horizontal overlap-80% and elevation of 20, 30,40 and 50 meters to ensure that the leaf is captured in higher resolution. Field deployment ensures that the area especially the weather is in good condition before conducting the flight, the UAV in this study will be flown in the clear sky during noon time. Flight plan verification step is conducted to double check the flight plan. Most of the UAV these days has an autonomous capability and will landing once the flight plan is completed or if the unnecessary event happens.

Images are captured at the top position of the tree in different altitude which are 20, 30, 40 & 50 meters above the ground. Moreover, images are captured during noon time to minimize shading as recommended by a study [26].

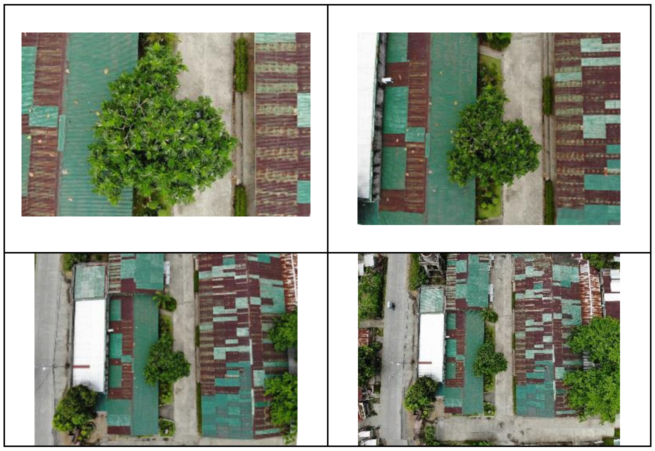


Figure 3: Four different altitude.

In Fig. 3 shows four different altitudes used to capture images. Which are 20, 30, 40 & 50 meters above the ground.

3.3 Pre-Processing

The image undergoes preprocessing and extraction which includes image composition, background segmentation, background subtraction and thresholding. Watershed algorithm segments the background and the vegetation. Using dilation and erosion operation markers are created to represent the background and the foreground of the image. Once the background and foreground (leaf) were segmented, the background will then be removed. Binary leaf images are extracted using thresholding.

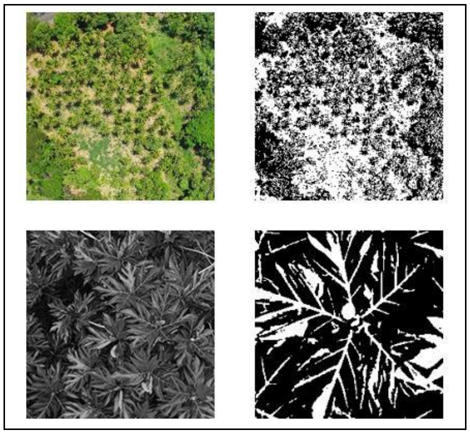


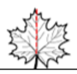
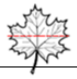
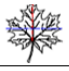
Figure 4: Image Processing.

Fig. 4 depicts the step by step process of extracting leaf from the images using the watershed algorithm. The extracted image then converted and compared to the image descriptor to identify what particular tree it is. Once an image is detected, a breadfruit tree classification using the neural network follows.

3.4 Feature Extraction

The process of lifting important measurable details in an image. The contour and region-based shape descriptor Simple Morphological Shape Descriptor (SMSD) extract leaf, length, area, perimeter, centroid, roundness, compactness and aspect ratio measures. Morphological features are extracted based on central moments [27].

Table 1: Descriptor for leaf morphological characters

Descriptor	Explanation	Pictogram
Major Axis Length (L)	Line segment connecting the base and the tip of the leaf.	
Minor Axis Width (W)	Maximum width that is perpendicular to the major axis	
Aspect Ratio (AR)	Ratio of major axis length to minor axis length— explains narrow or wide leaf or flower characteristics	
Roundness	Illustrate the difference between organ and a circle	
Perimeter (P)	Summation of the distance between each adjoining pair of pixels around the border of the organ	
Centroid	Represents the coordinates of the organ's geometric center	
Compactness	Ratio of the perimeter over the object's area; provides information about the general complexity and the form factor, it is closely related to roundness	
Area (A)	Number of pixels in the region of the organ	

Leaf descriptors were presented in Table 1. Several studies set standard features and sizes to easily identify the breadfruit tree using its leaf characteristics [28].

Table 2. Breadfruit Tree Variety

Variety	Shape	Estimated Measurement	Aspect Ratio
Variety I	oblong	(11.5 cm) long – (7.5 cm) wide	1.53
Variety II	round	(9 cm) long, (9cm) wide	1.00
Variety III	round	(12.5 cm) long, (11cm) wide	1.14
Variety IV	oblong	(20 cm) long, (15 cm) wide	1.33

Breadfruit tree variety was presented in Table 2 to identify the breadfruit tree according to measurement and shapes.

3.5 Classification

Artificial Neural Network (ANN) classifies the trees morphological data. Eight features serve as the input which is connected to two layers, each with eight hidden nodes. Each input is weighted and activated using the Sigmoid Function Eq. (1). The output of each node serves as an input to the succeeding nodes. The output nodes produce 1 and 0 denoting breadfruit and not breadfruit. Additional Forward Feed and Backpropagation ensure a robust classifier is created.

$$sigmoid = \frac{1}{1 + f^{-\sum(v_i w_{ij})+b}} \tag{1}$$

4 RESULTS AND DISCUSSION

This study sought to identify breadfruit tree and its classification. There are four categories of breadfruit tree present in the research area. Each set of images is fed to the neural network using feed forward and back propagation and sigmoid activation function. A model is first created using a pair of images in each category, and the result showed in the table below.

Table 3: Breadfruit Tree Variety Identification

Variety	Test Set 1	Accuracy	Test Set 2	Accuracy
I	25/32	78%	26/32	81%
II	22/32	69%	21/32	66%
III	24/32	75%	23/32	72%
IV	23/32	72%	22/32	69%

Table 3 shows the accuracy of classifying each variety of breadfruit. Based on the result, it shows that the Variety I have the highest accuracy among others. The average of Test Set 1 is 73.4% while the average of Test Set 2 is 71.9%, overall the accuracy is 72.66%. Although the accuracy is below 80%, it assures that the model is not overfitted. Several studies [31][32] confirms that a neural network accuracy to be marked as passed must have at least above 60% accuracy. Another study [33] that uses UAV has also acquired less than 80% accuracy.

5 CONCLUSIONS

This paper presents a research-in-progress of designing an intelligent morphological analysis of ArtocarpusAtilis (Breadfruit). It is found that several factors affect the image recognition rate of a neural network. Among those are image angle, nearby trees or structures and altitude of the UAV. It is concluded that using a UAV with AI to identify and classify breadfruit tree is effective. The accuracy of 72.66% is reasonable enough to help the local government unit to identify and classify breadfruit tree and its variety in the province. Thus, this study shares the development of the breadfruit tree roadmap initiated by the department of agriculture.

However, this study is not yet complete and needs further development. First, by capturing additional tree images in a different location. Through hundreds of images, the classification could be increased, and the accuracy likewise is increased. Second, performing in-depth image processing and analysis to increase the efficiency of the classifiers. Lastly, perform further validation using F1 score, and actual comparison of identification rate of an agriculturist and the system.

ACKNOWLEDGMENTS

The following institutions make this initial project possible: Commission on Higher Education, Technological Institute of the Philippines – Manila & Marinduque State College.

REFERENCES

[1.] F. R. Fosberg. 1960. Introgression in Artocarpus (Moraceae) in Micronesia, Brittonia, vol. 2, no. 12, pp. 101-113.

[2.] Sikarwar, M. S., Hui, B. J., Subramaniam, K., Valeisamy, B. D., Yean, L. K., & Kaveti, B. 2014. A Review on Artocarpusaltilis (Parkinson) Fosberg (breadfruit)., J App Pharm Sci, vol. 4, p. 8.

[3.] D. BRAUN. 2014. National Geographic, National Geographic, 2014. [Online]. Available: https://blog.nationalgeographic.org/tag/david-braun/.

[4.] S. &. S. Deivanai. 2010. Breadfruit (ArtocarpusaltilisFosb.)–An underutilized & neglected fruit plant species, Middle-East Journal of Scientific Research, vol. 6, no. 5, pp. 418-428.

[5.] M. C. O. Fresco. 2002 Growing Breadfruit, BAR Chronicle, August 2002. [Online]. Available: http://businessdiary.com.ph/11775/growing-breadfruit/#ixzz4tNRrn4x2.

[6.] D. F. Avegalio. 2017. Breadfruit Production Guide - Hawaii Department of Agriculture,

[7.] Englberger, L., Aalbersberg, W., Ravi, P., Bonnin, E., Marks, G. C., Fitzgerald, M. H., & Elymore, J. 2003. Further analyses on Micronesian banana, taro, breadfruit & other foods for provitamin A carotenoids & minerals., Journal of Food Composition and Analysis, vol. 16, no. 2, pp. 219-236.

[8.] P. B. C. Program. 2017. Counting ulu from space, Pacific Business Center Program. [Online]. Available: http://www.samoanews.com/linking-samoans/counting-ulu-space.

[9.] Philippine Statistics Authority, Fruitcrops Parameters And Fruiting Season. 2010. [Online]. Available: https://psa.gov.ph/gsearch?%2F=Fruitcrops+Parameters+and+Fruiting+Season.

[10.] Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y. X., Chang, Y. F., & Xiang, Q. L. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. In Signal Processing & Information Technology, IEEE International Symposium.

[11.] The Philippine Star. 2013. DA readies P36 million roadmap for 'rimas,. [Online]. Available: http://www.philstar.com/agriculture/2013/03/10/917738/da-readies-p36-million-roadmap-rimas.

[12.] D. &. L. G. Zhang. 2004. Review of shape representation & description techniques, Pattern Recognition, vol. 37, no. 1, pp. 1-19.

[13.] Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. I. 2011. A comparative experiment of several shape methods in Recognizing plants, arXiv preprint arXiv:1110.1509, 2011.

[14.] K. J. &. O. M. A. Gaston. 2004. Automated species identification: why not? Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 359, pp. 655-667

[15.] Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., ... & Planquè, R. 2014. LifeCLEF: Multimedia life species identification., EMR@ ICMR, pp. 7-13.

[16.] Wäldchen, J. A. N. A., Thuille, A., Seeland, M., Rzanny, M., Schulze, E. D., Boho, D., ... & Mäder, P. 2016. Flora Incognita–Halbautomatische Bestimmung der Pflanzenarten Thüringens mit dem Smartphone. Landschaftspflege und Naturschutz in Thüringen, 53(3), 121-125.

[17.] P. &. E. I. H. Wilf. 2015. Green Web or megabiased clock? Plant fossils from Gondwanan Patagonia speak on evolutionary radiations., New Phytologist, vol. 207, no. 2, pp. 283-290.

[18.] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[19.] H. H. Martens. 1959. Two notes on machine learning., Information Control, vol. 2, pp. 364-379.

[20.] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[21.] L. Carey, International Civil Aviation Organization UAS Study Group, ICAO UAS. 2010. [Online]. Available: http://www.dcabr.org.br/download/eventos/eventos-realizados/2010/seminario-vant-27-10-2010/cd-uvs-yearbook/pdf/P051-053_ICAO_Lesly%20Carey.pdf.

[22.] T. P. Marthinus Reinecke. 2017. The influence of drone monitoring on crop health and harvest size, in NextComp IEEE Mauritius, South Africa.

[23.] Nevalainen, O., Honkavaara, E., Tuominen, S., Viljanen, N., Hakala, T., Yu, X., ... & Tommaselli, A. M. 2017. Individual tree detection and classification with UAV-based photogrammetric point clouds and hyperspectral imaging. Remote Sensing, 9(3), 185.

[24.] Aakif, A., & Khan, M. F. 2015. Automatic classification of plants based on their leaves. Biosystems Engineering, 139, 66-75.

[25.] Hossain, J., & Amin, M. A. 2010. Leaf shape identification based plant biometrics. In Computer and Information Technology (ICCIT), 2010 13th International Conference on (pp. 458-463). IEEE.

[26.] Teague, M. R. 1980. Image analysis via the general theory of moments. JOSA, 70(8), 920-930.

[27.] J. Morton, Breadfruit, [Online]. Available: https://www.hort.purdue.edu/newcrop/morton/breadfruit.html

[28.] S. Grossberg. 1982. How does a brain build a cognitive code?, Studies of mind and brain, pp. 1-52.

[29.] Lobell, D. B., Asner, G. P., Law, B. E., & Treuhaft, R. N. 2002. View angle effects on canopy reflectance and spectral mixture analysis of coniferous forests using AVIRIS. International Journal of Remote Sensing, 23(11), 2247-2262.

[30.] Yu, X., Hyyppä, J., Hyyppä, H., & Maltamo, M. 2004. Effects of flight altitude on tree height estimation using airborne laser scanning. Proceedings of the Laser Scanners for Forest and Landscape Assessment–Instruments, Processing Methods and Applications, 02-06.

[31.] Mishra, T., Kumar, D., & Gupta, S. 2014. Mining students' data for prediction performance. In Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on (pp. 255-262). IEEE.

[32.] Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. 2008. Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology, 9(1), 72-79.

[33.] Vaughn, N. R., Moskal, L. M., & Turnblom, E. C. 2012. Tree species detection accuracies using discrete point lidar and airborne waveform lidar. Remote Sensing, 4(2), 377-403.

Weissman Score of Lossless Compression Algorithms

Ronjie Mar L. Malinao
Technological Institute of the Philippines, Manila
ronjieclear@yahoo.com

Jasmin D. Niguidula
Technological Institute of the Philippines, Manila
jasmin.niguidula@tip.edu.ph

ABSTRACT

In today's technological advancement we generate a lot of data. Data grow much faster than our storage devices which lead us to delete the less important one to occupy new much important data. Can an algorithm rearrange and store those data eighty percent less of its original size within seconds? In this study, several compression algorithms embedded in compression application were scrutinized. Compression ratio, compression speed was gathered and analyzed. Finally, a compression algorithm efficiency metric called Weissman Score was used to define which algorithm is the best. It is found that WinRAR which utilizes proprietary compression algorithm rank first in the said metric.

Keywords: *Data Compression, Weissman Score, Information Theory*

1. INTRODUCTION

Today we are surrounded by computers, systems, mobile phones, close circuit television and other electronic devices that collect, process and store data. Data that are so massive, it needed advance technology to deal with it. This massive amount of data once used was thrown in historical archived [1] or immediately deleted and wasted [2]. Data needs to be deleted to prevent redundancy and due to its lack of importance; like taking a photograph using a smartphone just to realize the photo is blurred. Most of the time, data was deleted to allocate new space for the upcoming data. Thus, deleting one of the several equally important data give us a hard time to choose from.

There are so many data that may not be needed today but could be processed and provide valuable information in future. Rather than deleting, several companies evolved focusing on data reduction. One of them process data through deduplication, thin provisioning, and compression which then shrink data with a ratio more than 10:1[3].

Data compression can be embedded in the way how a processed data transform into a file with its corresponding file format. Compression was divided into two categories, lossless and loosy. Loosy compression focuses on preserving the meaning or perceptible information [4], like a photograph wherein some dot or pixel can be omitted and may not be recognized by a human eye. Lossless compression preserves the data or the exact digital information [4], like a spreadsheet file wherein an omitted single dot in a monetary value could dramatically change its meaning. Thus, the precise information should be compressed and decompress into its original form.

Compression is done using several compression algorithms. This study focuses on the "Lossless" compression algorithms which are embedded and implemented through many compression applications including proprietary and open-source.

Objective

This study sought to evaluate lossless compression algorithms with the following objective:

1. To identify which has the best compression ratio
2. To identify which has the fastest compression process
3. To identify which utilizes fewer system resources
4. To identify Weissman Score of each compression algorithm

2. RELATED WORKS & LITERATURE

2.1 History of compression

Date compression starts from the invention of Morse Code, wherein shorter “code words” was used for common letters [5]. Modern compression starts in the 1940s with the development of Information Theory. Claude Shannon and Robert Fano devised a systematic way to assign the said code words based on probabilities of block [6]. Lately, David Huffman came up a new method called Huffman Coding [7] which reverse the probability tree of Shannon and Fano. In 1977 Abraham Lempel & Jacob Ziv suggested the basic idea of pointer-based encoding [8]. They use a dictionary based algorithm called LZ77 and LZ78 which later become the basis for many other compression algorithms.

2.2 DEFLATE

Phillip W. Katz developed compressed data format called DEFLATE; it combines Lempel-Ziv compression with Huffman coding [9]. It is used in PKZIP archiving tool introduced for Microsoft DOS and later implemented into succeeding Microsoft operating system. The method and device were granted US patent. However, the algorithm producing DEFLATE file is not covered by the patent. Thus, it is implemented in images files like Portable Network Graphics (.png), archive file format (.zip) and compression application like WinZip [10]. Experiment [11] shows that WinZip is the fair choice if aiming for a fast compression application.

2.3 LZMA

This compression algorithm was a combination of LZ77, Markov Chains and Range Coder developed by Igor Pavlov [12]. This algorithm is used in an open source file archiver called 7-Zip which is developed by Pavlov. File compressed using 7z application is converted into .7z file format. Several experiments [13] [14] conducted shows that 7-Zip is much better than ZIP.

2.4 LZSS

This compression algorithm also came from the roots of LZ77, created by James Storer and Thomas Szymanski [15]. Windows operating system utilizes the algorithm in its New Technology File System (NTFS) file compression [16]. In terms is archiving application, a study [17] describe that WinRAR uses the LZSS & prediction by partial matching algorithms. However, the said application is proprietary and did not show its compression algorithm in public. WinRAR produces a .rar file which is common on file sharing sites. An experiment [18] shows that even though it is not the leader in compression ratio and compression speed it is the best.

3. METHODOLOGY

The following procedure, tools, settings, configuration and equations were use in this study.

3.1 File

Files were group depending on their purpose. Even though some of the file formats are naturally compressed, some algorithms could recompress those files in a minimum ratio.

Table 1. File Group & Format

File Group	File Format
Documents	.docx, .xlsx, .pptx, .pdf
Images	.bmp, .gif, .jpg, .png
Movies	.avi, .mp4, .mkv,
Music	.mp3 & .wav

Table 1 shows thirteen (13) common lossy and lossless file format [19] [20] and their corresponding file group [21] used in this study.

Table 2. File Specifications

File Group	File Format	Percentage from Group	Total Size (bytes)	Average Size (bytes)
Documents				
426,274,962 bytes	.pdf	25.0%	106,592,512	3,331,016
	.docx	25.2%	107,557,584	3,361,174
	.xlsx	24.6%	105,010,733	3,281,585
	.pptx	25.1%	107,114,133	3,347,316
Images				
431,376,055 bytes	.bmp	24.8%	106,893,100	1,875,317
	.gif	25.2%	108,729,786	1,647,421
	.jpg	25.1%	108,353,450	1,900,937
	.png	24.9%	107,399,719	1,884,205
Movies				
22,199,411,901 bytes	.avi	33.2%	7,367,789,959	736,778,996
	.mkv	33.6%	7,459,060,616	745,906,062
	.mp4	33.2%	7,372,561,326	737,256,133
Music				
848,128,550 bytes	.mp3	49.9%	423,333,660	7,426,906
	.wav	50.1%	424,794,890	7,452,541

Table 2 shows the size of each file group and file format. Each file group has similar amount file, and average file size is close to each other to prevent bias.

3.2 Compression Ratio

To address the compression ratio of each algorithm the formula below was derived from a study [22] a website [23] and a book [24]. Size is treated in bytes to ensure data compression difference will be shown down to its lowest measurable unit. Compression Ratio is determined by the following:

$$CR \equiv \sum_{n=1}^{FQ} \mathrm{F}_n \left(\frac{\mathrm{US}}{\mathrm{CZ}}\right)$$

Where:

CR – Compression Ratio
CZ – Compressed Size
US – Uncompressed Size
FQ – File Quantity
F – File

3.3 Compression Speed

In evaluating data compression algorithms, speed is always in terms of uncompressed data handled per second [23]. Compression and decompression time is the amount of time in hours, minutes and seconds consume to complete the whole operation. Thus, it is written in a formula stated below. For higher accuracy and better comparison, gathered time is converted in seconds and results speed is transformed into megabytes per second.

$$CS \equiv \sum_{n=1}^{FQ} \mathrm{F}_n \frac{CZ\frac{10^6}{1MB}}{(\mathrm{ET}-\mathrm{ST})}$$

Where:

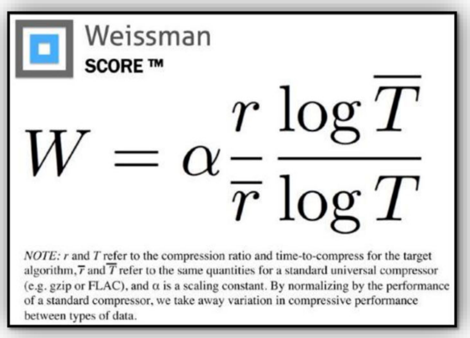
CS – Compression Speed
CZ – Compressed Size
FQ – File Quantity
MB – Mega Bytes
ST – Start Time
ET – End Time
F – File

3.4 Compression Resource Utilization

A compression algorithm could be fast and get higher compression ratio but at the expense of too much CPU and RAM utilization. Thus, it is an equally important task to monitor the system resources utilized by those algorithms. The built-in Microsoft Resource Monitor together with MSI Afterburner application is used to monitor CPU and Memory same application utilized in the field of algorithm analysis [25] and energy efficiency analysis [26].

3.5 Weissman Score

Weissman Score is a metric that could be used to score multiple algorithms. Developed in Stanford by Professor Tsachy Weissman and then-PhD student Vinith Misra [27]. Since the inception of the said metric (Fig. 1), several studies [28] [29] [30] have already adopted, and future researchers stated to utilize the same. The Weissman Score served as an evaluation method of each compression algorithm. Implementing the said metric requires a standard universal compressor. Thus built-in Microsoft Zip (MS Zip) serves as the basis for comparison. While alpha (α) or the scaling constant is set one (1) that ensure no negative score is produced.



Weissman

SCORE ™

$$W = \alpha \frac{r}{\overline{r}} \frac{\log \overline{T}}{\log T}$$

NOTE: r and T refer to the compression ratio and time-to-compress for the target algorithm, $\overline{r}$ and $\overline{T}$ refer to the same quantities for a standard universal compressor (e.g. gzip or FLAC), and α is a scaling constant. By normalizing by the performance of a standard compressor, we take away variation in compressive performance between types of data.

Figure 1. Weissman Score

3.6 System Configuration

Table 3. System Configuration

System Component	Description
Processor	Intel Core i5-4690K 3.5 GHz
Memory	G. Skill 4GB DDR3 800 MHz (2 pcs.)
Storage	Samsung SSD 840 530 MBps (read) / 130 MBps (write)
Operating System	Windows 7 Home Premium 64-bit

Table 3 shows the computer configuration used in this study. All component settings were set to default, and no overclocking or performance enhancement application was use.

Default Windows file system could affect the result of compression due to real-time compression performed by the operating system [31]. Since Windows 8 and 8.1 have a different file system, Windows 7 was selected due to the similarity of their NTFS file system and to ensure all compression application is compatible.

3.7 Compression Application

There are several algorithms that can compress data into its minimal form, among them are implemented in Windows operating system applications.

Table 4. Compression Application

Application Name	Algorithm Used	Version	Release Date
MS Zip	Deflate	6.1	July 14, 2009
WinZip	Deflate	21	October 25, 2016
WinRAR	Proprietary	5.40	August 16, 2016
7-Zip	LZMA	16.04	October 4, 2016

Table 4 shows the popular compression application [32] [33] used in this study. There are three different algorithms in each application. While MS Zip and WinZip share the same algorithm MS Zip role is to set the ground basis for other application. Moreover, all compression program was 64 bit and all application is assured latest version aside from MS Zip.

4. RESULTS & DISCUSSIONS

4.1 Compression Ratio

Table 5. Compression Ratio of Document Files

Extension Name	MS Zip	WinZip	7Zip	WinRAR
.pdf	1.2 : 1	1.2 : 1	1.4 : 1	1.3 : 1
.docx	1.1 : 1	1.1 : 1	3.1 : 1	1.3 : 1
.xlsx	1.5 : 1	1.5 : 1	3.2 : 1	1.7 : 1
.pptx	1.1 : 1	1.1 : 1	1.3 : 1	1.1 : 1
Average	1.2 : 1	1.2 : 1	1.9 : 1	1.3 : 1

Table 5 shows document group file formats and their compression ratio. It is remarkable that 7-Zip has the largest compression ratio specifically in the processing of MS Excel files which are the .xlsx. Moreover, the average compression ratio of 1.9:1 indicate that almost two document file could fit in a storage amount of a single file if it compressed using 7-Zip.

Table 6. Compression Ratio of Image Files

Extension Name	MS Zip	WinZip	7Zip	WinRAR
.bmp	2.6 : 1	2.7 : 1	5.3 : 1	3.1 : 1
.gif	1.0 : 1	1.0 : 1	3.1 : 1	1.0 : 1
.jpg	1.0 : 1	1.0 : 1	1.0 : 1	1.0 : 1
.png	1.0 : 1	1.0 : 1	1.1 : 1	1.0 : 1
Average	1.2 : 1	1.2 : 1	1.6 : 1	1.2 : 1

Table 6 shows the file format of common image files. All application could not compress files in .jpg; this is because those files are already compressed in lossy. On the other hand, only 7Zip can compress images in .gif which indicate that LZMA could recompress a file which already compressed using LZW. On the other hand, .bmp images are raw files which can be further compress up to 5.3:1 using the 7Zip.

Table 7. Compression Ratio of Music Files

Extension Name	**MS Zip**	**WinZip**	**7Zip**	**WinRAR**
.mp3	1.0: 1	1.0: 1	1.0: 1	1.0: 1
.wav	1.1: 1	1.1: 1	1.4: 1	1.8: 1
Average	1.1: 1	1.1: 1	1.2: 1	1.3: 1

Table 7 shows two common music files. It can be noticed that .mp3 files cannot be compressed further enough by any compression application due to its high compression ratio from the original CD quality file. However, .wav file which typically a raw uncompressed file can be compressed up to 1.8:1 using WinRAR. In terms of compressing .wav files the proprietary algorithm used by WinRAR dominate among others.

Table 8. Compression Ratio of Movie Files

Extension Name	**MS Zip**	**WinZip**	**7Zip**	**WinRAR**
.avi	1.0: 1	1.0: 1	1.0: 1	1.0: 1
.mkv	1.3: 1	1.3: 1	1.2: 1	1.3: 1
.mp4	1.0: 1	1.0: 1	1.0: 1	1.0: 1
Average	1.1: 1	1.1: 1	1.1: 1	1.1: 1

Table 8 shows three common movie files. Only .mkv files can be compressed by any of the compression application. Most of the application can compress up to 1.3:1 aside from 7Zip which render 1.2:1. The .mp4 file could contain compressed audio and video file thus it cannot be recompressed further. While.avi file could contain uncompressed and compressed file, sample files used in this study may contain more compressed file resulting to 1:1 ratio. Moreover, the average of 1.1:1 across all movie files indicate that movies are well compressed by their respective codec and it could be a waste of time and resources to compress these kinds of files.

Table 9. Compression Ratio of All Files

File Group	**MS Zip**	**WinZip**	**7Zip**	**WinRAR**
Documents	1.18:1	1.18:1	1.93:1	1.30:1
Images	1.19:1	1.19:1	1.62:1	1.21:1
Music	1.08:1	1.09:1	1.18:1	1.32:1
Movie	1.09:1	1.10:1	1.09:1	1.10:1
Average	1.13:1	1.14:1	1.46:1	1.23:1

Table 9 shows the average compression ratio of each application considering all kind of file group. It is remarkable that 7Zip with its LZMA algorithm have the largest ratio among others while the built-in Windows compression got the lowest ratio. This proves that same compression algorithm (deflate) could generate identical or close result even though it is embedded in a different kind of application.

4.2 Compression Speed

Table 10. Compression Speed of Document Files *(MB/seconds)*

	Native	**WinZip**	**7-Zip**	**WinRAR**
.pdf	18.15	18.48	8.33	20.33
.docx	18.00	21.82	11.53	21.37
.xlsx	18.21	14.95	7.15	20.86
.pptx	17.92	20.85	9.04	19.27
Average	18.07	19.03	9.01	20.46

Table 10 shows the compression speed of documents file group. It is found that WinZip is the fastest compression application that process specifically the MS Word document or the .docx file. On the other hand, 7Zip got slower performance among others.

Table 11. Compression Speed of Image Files *(MB/seconds)*

	Native	**WinZip**	**7-Zip**	**WinRAR**
.bmp	19.60	12.40	6.75	23.71
.gif	17.58	14.01	11.03	18.85
.jpg	18.79	17.22	7.60	20.67
.png	19.33	20.90	7.82	19.33
Average	18.82	16.13	8.30	20.64

Table 11 shows that compression speed of Images file (row). WinRAR got the fastest compression speed among other specifically the .bmp files. 7Zip, on the other hand, got the lowest compression speed, with the .jpg format; it is almost a quarter of speed compare to .bmp.

Table 12. Compression Speed of Music Files *(MB/seconds)*

	Native	**WinZip**	**7-Zip**	**WinRAR**
.mp3	20.60	25.55	7.34	19.32
.wav	18.33	25.97	6.87	26.14
Average	19.40	25.76	7.10	22.22

Table 12 shows compression speed of the two popular music files (rows). WinRAR got the fastest compression speed in .wav files, but WinZip got the fastest speed and in .mp3 files. Overall, WinZip got the fastest compression speed.

Table 13. Compression Speed of Movie Files *(MB/seconds)*

	Native	**WinZip**	**7-Zip**	**WinRAR**
.avi	15.05	26.44	7.15	15.16
.mkv	27.56	42.80	9.06	19.28
.mp4	64.43	125.79	24.31	54.83
Average	60.29	105.80	24.07	52.12

Table 13 shows that WinZip dominate all compression application and it almost doubles the speed among others. However, based on the table X{Ratio} only .mkv can be compressed down to much smaller size and those other file formats have no significant change in size.

Table 14. Compression Speed of All Files *(MB/sec)*

	MS Zip	**WinZip**	**7Zip**	**WinRAR**
Documents	18.1	19.0	9.0	20.5
Image	18.8	16.1	8.3	20.6
Music	19.4	25.8	7.1	22.2
Movie	60.3	105.8	24.1	52.1
Average	29.1	41.7	12.1	28.9

Table 14 shows all file group and their average compression speed using specified compression application. In general, WinZip is the fastest compression application with the average speed of 41.7 MB per second followed by MS Zip. Thus, it proves that Deflate algorithm can compress data in far more quickly compare to others.

4.3 Compression Resources

Table 15. Compression Resources Utilized *(maximum)*

File Group	**MS Zip**		**WinZip**		**7Zip**		**WinRAR**	
	CPU	RAM	CPU	RAM	CPU	RAM	CPU	RAM
Documents	20%	344 MB	100%	456 MB	48%	261 MB	60%	295 MB
Images	25%	339 MB	100%	545 MB	48%	282 MB	53%	335 MB
Music	27%	649 MB	100%	703 MB	50%	288 MB	65%	323 MB
Movie	26%	259 MB	100%	1468 MB	57%	310 MB	72%	362 MB
Average	25%	398 MB	100%	793 MB	51%	285 MB	63%	329 MB

Table 15 shows the CPU and RAM used by the compression applications during compression. It can be noticed that MS Zip has the lowest CPU utilization and 7Zip got the lowest RAM utilization. On the other hand, WinZip utilizes the maximum CPU available and use a large amount of RAM.

4.4 Weissman Score

Since the value one (1) is use as the Alpha in Weissman Score, all value less than one (1) is considered lower compare to the base compression application. A value greater than one (1) performs greater than the base application.

Table 16. Weissman Score of Documents

	WinZip	7-Zip	WinRAR
.docx	1.1	2.3	1.3
.pdf	1.0	0.8	1.1
.pptx	1.1	0.9	1.1
.xlsx	0.9	1.4	1.3
Total	4.2	5.5	4.8

Table 16 shows Weissman score of document files and their corresponding total. 7Zip got the highest score in this group of file but it only performs higher in .docx and .xlsx and fails in .pdf and .pptx. On the other hand, WinRAR even though score below the 7Zip ensure that it perform better in all document files.

Table 17. Weissman Score of Image

	WinZip	7-Zip	WinRAR
.bmp	0.8	1.3	1.4
.gif	0.9	2.4	1.0
.jpg	1.0	0.7	1.0
.png	1.0	0.7	1.1
Total	3.7	5.0	4.5

Table 17 shows Weissman score of Image files. 7Zip gain the highest score, but in .bmp and .gif file only, an opposite result is gain by WinZip. While WinRAR shows same and above performance from the base compression application.

Table 18. Weissman Score of Music

	WinZip	7-Zip	WinRAR
.mp3	1.1	0.7	1.0
.wav	1.1	0.9	1.8
Total	2.2	1.7	2.8

Table 18 shows the Weissman score of music files. WinRAR got the highest score in terms of audio compression besides the fact that it is not the fastest in terms of compression speed *(Table 12)*. But it tops at the compression ratio *(Table 7)*. WinZip follows it with above performance than the base compression application. However, 7Zip did not perform well in this group of files.

Table 19. Weissman Score of Movie

	WinZip	**7-Zip**	**WinRAR**
.avi	1.1	0.9	1.0
.mkv	1.1	0.8	0.9
.mp4	1.1	0.8	1.0
Total	3.3	2.6	2.9

Table 19 shows the Weissman score of movie files. WinZip dominate in this kind of file with above performance than the base compression application. The compression speed which is twice faster than the two *(as shown in Table 13)* become the key in gaining the highest score. 7Zip did not perform well in any movie file due to the slow compression speed.

Table 20. Weissman Score Overall Result

	Document	**Image**	**Music**	**Movie**	**WS**	**Rank**
WinRAR	4.8	4.5	2.8	2.9	15.0	1
7-Zip	5.5	5.0	1.7	2.6	14.7	2
WinZip	4.2	3.7	2.2	3.3	13.4	3

Table 20 shows the overall Weissman Score results. WinRAR got the total score of 15. Although it did not dominate on the document, image, and movie file group, it remains second from each category. Additionally, it surpasses other application in music file group. Therefore, WinRAR with its proprietary compression algorithm is considered the best compression algorithm evaluated using the said metric.

LZMA algorithm even though it is the slowest it dominates in compressing Microsoft Office files *(.docx, .pptx and .xlsx)*. and some image files (*.bmp and .gif*). Even though it just followed to WinRAR proprietary algorithm with a difference of .3 it proved that compressed files could be recompressed further.

Deflate algorithm in the built-in compression of Windows is the best in resource efficiency, suited for slow machines; Unlike WinZip that implement the same algorithm but utilizes a lot of resources. Aside from their weaknesses Deflate is the fastest compression algorithm and suited for a quick and typical compression jobs.

5. CONCLUSION

It is found that there are several compression algorithms utilized by compression applications. The three essential factors in compression are the amount of space that can be saved, the time needed to perform the task and the resources it will consume. Since there are several studies and benchmark done, Weissman Score provides a new perspective in the evaluation of those algorithms.

It is found out that LZMA algorithm is the best regarding compression ratio. Deflate is the fastest compression algorithm that can be resource efficient or its opposite depending on the application where it is embedded. Lastly, WinRAR with its proprietary algorithm got the highest score in Weissman Score compression algorithm metric.

It is recommended that WinRar is the first option if there are different kinds of files to be compressed. 7Zip is recommended for Microsoft Office files. Microsoft Zip is best if using a slow computer while WinZip is recommended if compression time is the number one priority to finish the job. In the end, each application could release versions soon that may include several compression algorithms and select which algorithm would be best suitable for the job before the compression happen.

REFERENCES

[1] Sun, Y., Yan, H., Lu, C., Bie, R., Zhou, Z. (2014) Constructing the Web of Events from Raw Data in the Web of Things - Mobile Information Systems Volume 10, Issue 1, Pages 105-125 http://dx.doi.org/10.3233/MIS-130173

[2] Schaub, K. (2014, February 12) 80% of Your Customer Data Will be Wasted. Retrieved from http://techmarketingblog.blogspot.com/2014/02/80-of-your-customer-data-will-be-wasted.html

[3] PureStorage (2017) How Data Reduction Works. Retrieved from https://www.purestorage.com/products/purity/flash-reduce.html

[4] DVD.HQ (2017) Data compression basics. Retrieved from http://dvd-hq.info/data_compression_1.php

[5] Dimitrov, P. (2014) The History of Data Compression. Retrieved from http://techmeup.net/history-data-compression-infographic/

[6] Wolfram., S. (2002) A New Kind of Science. Retrieved from https://www.wolframscience.com /reference/notes/1069b

[7] Salomon, D. (2008) A Concise Introduction to Data Compression. Retrieved from http://www.springer.com/978-1-84800-071-1

[8] Ziv, J., Lempel, A. (1977) A Universal Algorithm for Sequential Data Compression. IEEE Transactions on Information Theory, Vol. It-23, No. 3. Retrieved from www.cs.duke.edu/courses/spring03/cps296.5/papers/ziv_lempel_1977_universal_algorithm.pdf

[9] Deutsch, P. (1996). DEFLATE Compressed Data Format Specification version 1.3. IETF. p. 1. sec. Abstract. RFC 1951. Retrieved from https://tools.ietf.org/html/rfc1951#section-Abstract

[10] WinZip (2016) Choosing a Compression Method. Retrieved from http://kb.winzip.com/help/help_compression.htm

[11] Petrovic (2014) 15 Archivers Tested to Find the Fastest Speeds and Smallest File Sizes. Retrieved from https://malwaretips.com/threads/15-archivers-tested-to-find-the-fastest-speeds-and-smallest-file-sizes.22664/

[12] Pavlov, I. (2004). LZMA spec? [Online forum comment]. Message posted to https://sourceforge.net/p/scoremanager/discussion/457976/thread/c262da00/

[13] TechMediaNetwork.com (2011) 7-Zip Review. Retrieved from https://web.archive.org/web/20121025200631/http://file-compression-software-review.toptenreviews.com/7-zip-review.html

[14] Masiero, M. (2013) Compression Performance: 7-Zip, MagicRAR, WinRAR, WinZip. Retrieved from http://www.tomshardware.com/reviews/winrar-winzip-7-zip-magicrar,3436-13.html

[15] Storer, J., Szymanski, T. (1982). "Data Compression via Textual Substitution". Journal of the ACM. 29 (4): 928–951. doi:10.1145/322344.322346.

[16] Mahoney, M. (2013) Data Compression Explained. Retrieved from http://www.mattmahoney.net /dc/dce.html#Section_521

[17] Hasan, M., Ibrahimy, M., Motakabber, S., Ferdaus, M., Khan, M. (2013) Comparative data compression techniques and multi compression results. 5th International Conference on Mechatronics (ICOM'13) doi:10.1088/1757-899X/53/1/012081

[18] Smith, M. (2012) What's the Best File Compression Method? MakeUseOf Tests Zip, RAR & More. Retrieved from http://www.makeuseof.com/tag/best-file-compression-method-zip-rar-and-more/

[19] GPSoftware (2016) File Type Groups. Retrieved from https://www.gpsoft.com.au/help /opus12/index.html#!Documents/Groups.htm

[20] Duff Johnson Strategy and Communication (2015) The 8 most popular document formats on the web. http://duff-johnson.com/2014/02/17/the-8-most-popular-document-formats-on-the-web/

[21] Beal, V. (2005) Common Audio Formats. Retrieved from http://www.webopedia.com/ DidYouKnow/Computer_Science/digital_audio_formats.asp

[22] Kumar, V., Barthwal, S., Kishore, R., Saklani, R., Sharma, A., Sharma, S., (April 2016) Lossy Data Compression Using Logarithm. Retrieve from https://www.researchgate.net/publication/ 301876041_Lossy_Data_Compression_Using_Logarithm

[23] Wikibooks (2017) Data Compression/Evaluating Compression Effectiveness. Retrieved from https://en.wikibooks.org/wiki/Data_Compression/Evaluating_Compression_Effectiveness

[24] Salomon, D. (2002) A Guide to Data Compression Methods. Springer-Verlag New York Inc. NY.

[25] Mukherjee, S. (2014) Application of Parallel Algorithm Approach for Performance Optimization of Oil Paint Image Filter Algorithm. Signal & Image Processing: An International Journal (SIPIJ) Vol.5, No.2, April 2014

[26] Cebri'n, J., Guerrero, G., García, J., (2012) Energy Efficiency Analysis of GPUs. Parallel and Distributed Processing Symposium Workshops & Ph.D. Forum (IPDPSW), 2012 IEEE 26th International. DOI: 10.1109/IPDPSW.2012.124

[27] Perry, T. (2014) A Fictional Compression Metric Moves into the Real World. Retrieved from http://spectrum.ieee.org/view-from-the-valley/computing/software/a-madefortv-compression-metric-moves-to-the-real-world

[28] Baron, D., Weissman, T. (2012) An MCMC Approach to Universal Lossy Compression of Analog Sources. IEEE Transactions on Signal Processing (Volume: 60, Issue: 10, Oct. 2012)

[29] Zoph, B., Ghazvininejad, M., & Knight, K. (2015). How Much Information Does a Human Translator Add to the Original? In EMNLP (pp. 889-898). Retrieved from https://pdfs.semanticscholar. org/2228/d339e4db65f38e9390313169c1af13e3092a.pdf

[30] Fisher, K. M. (2016). Towards Understanding the Compression of Sound Information. Chicago. Retrieved from http://digitalcommons.trinity.edu/compsci_honors/38/

[31] Hoffman, C. (2016) Should You Use Windows' Full-Drive Compression to Save Space? Retrieved from http://www.howtogeek.com/266472/should-you-use-windows-full-drive-compression-to-save-space/

[32] Masiero, M. (2013) Compression Performance: 7-Zip, MagicRAR, WinRAR, WinZip. Retrieved from http://www.tomshardware.com/reviews/winrar-winzip-7-zip-magicrar,3436-6.html

[33] Dean, M. (2016) 8 best file compression tools for Windows 10. Retrieved from http://windowsreport.com/best-compression-tools-windows-10/