Using Deep Learning to assess Phenotypical Differences in Rice Seed Image data of Genotypically Identified Seeds of Indica Subpopultaion

Marvin Jerald F. Villador, Jeffrey Detras and Val Randolf Madrid

***Abstract*— Rice variety groups Indica, Japonica, Aus, and Basmati are genetically proven to have genomic variation. The study aims to check if the genomic variation between these groups significantly translate with its phenotype or appearance using image data. Deep Learning was implemented in training, validation, and prediction process. Convolutional Neural Network (CNN) was used to build the classifier with InceptionV3 architecture.**

***Index Terms*— deep learning, convolutional neural network, machine learning, genomic variation**

I. Introduction

Rice research contribute greatly in food security. Rice is considered as one of the most important foods around the world. It is one of the most consumed foods for a large part of the human population. Large amount of various rice varieties are circulating worldwide, being imported and exported in different countries. These huge amount of demand and transactions give support to the economy of these countries. With the continuous growth of technology and demand for knowledge in rice research to produce sustainable rice varieties that will support the future population, large amount of data are produced [1].

International Rice Research Institute (IRRI) is one of the global institutions that are actively conducting researches to provide a solution for the limited food supply, specifically with rice. Under the field of Bioinformatics, one medium for data analysis is using SNP data [2]–[4]. Another one is the use of digital images. To provide initial data, manual rice seed classification serves an important part in rice research. This method is considered tedious, time-consuming, and prone to error [5]. Thus, studies for rice analysis using image data will be of great help to maximize the potential of rice research given available data to provide information for future works.

Rice seed classification is one of researches under Bioinformatics. Different methods are introduced using biological and chemical techniques to develop rice variety classification. These methods includes the use of genetic markers for rice variety classification. Though these methods are proved accurate for classification, these uses techniques that are costly and time-consuming. This leads to using small amount of data sampling which produces less accuracy than classification of the whole batch. These methods cannot still secure the purity of a rice variety under inspection. To solve the problem of limitation, methods using digital image processing techniques and machine learning have been proposed.

With a large amount of data for classification, exploring deep learning will have an advantage in data analysis. Deep learning is a subset of Artificial Intelligence (AI) and machine learning. It employs principles of machine learning but does its computations in layers. The objective of the study is to explore and develop a system by creating a Convolutional Neural Network classifier model using TensorFlow.

Neural Network model is a machine learning model based on the neural circuitry of the human brain. It has three layers of nodes; input or feature layer, hidden layer, and output layer. The feature layer holds the data feature or input. The hidden layer extracts and manipulate data from feature layer using algorithms giving the output layer data to work with. The output layer holds the results. TensorFlow is a machine learning library for Python developed by Google. It uses tensor, a multidimensional array, as data structure that flows through operations. TensorFlow serves as a stable platform for deep learning in research and applications [6]. TensorFlow uses visualization tool called TensorBoard. It can be used to visualize TensorFlow graph and generate data about the execution of the program. This helps to develop the system easier. The artificial intelligence system supports Convolutional Neural Network (CNN) model, a recognition method that simplifies complex image pre-processing efficiently compared to a regular neural network [7]. This neural network model is widely used for image classification in different studies.

The proposed classifier will give an output of classification accuracy for each class from a given group of rice varieties given the seed images as input. The classifier will be tested on image data from the International Rice Research Institute (IRRI).

A. Objectives of the Study

The study aims to classify rice seed varieties in a phenotypical approach using image data with the application of Deep Learning.

1. Use image processing techniques to extract and enhance image data to improve training

2. Train a classifier model for Indica variety group's subpopulation (XI-1A, XI-1B, XI-2, XI-3, and XI-adm) given rice

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

seed images as input data from IRRI using a Convolutional Neural Network classifier using Keras and Google TensorFlow

3. Test and evaluate the results in terms of classification accuracy of the model

B. Significance of the Study

Genotype refers to the genetic code of an organism while phenotype is the expression of genotype that can be observed. Genomic variation shows differences of organisms by the use of DNA sequences. Rice variety groups Indica, Japonica, Aus, and Basmati are genetically proven to have genomic variation [8]. With this information that these variety groups varies genotypically, the study aims to check if the genomic variation between these groups can be significantly distinguished with its phenotype or appearance using image data. This will show if the genotypical differences between the classes tranlates significantly to its appearances. The rice variety groups that will be used as classes are the subpopulation of Indica variety group (XI-1A, XI-1B, XI-2, XI-3, and XI-adm). These are the only subpopulation with sufficient information that are available from the limited image data in IRRI. The neighbour-joining tree shows how near the rice seed varieties from each other.

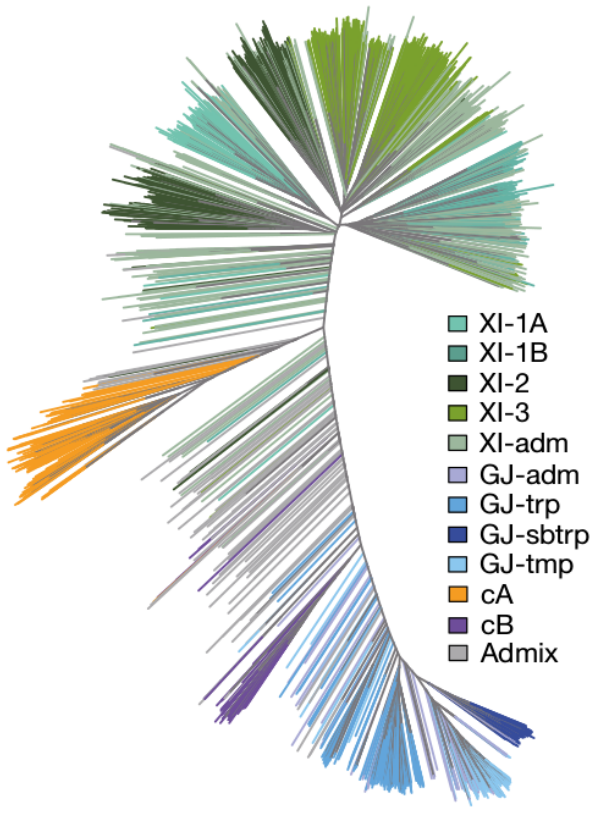

Fig. 1. Unweighted neighbour-joining tree based on 3,010 samples [8]

Though many different studies are conducted concerning rice seed classification, few are done using Deep Learning or with Convolutional Neural Network (CNN). This will also be the first in IRRI. With the large amount of image data, CNN will be great use in image analysis for its powerful features.

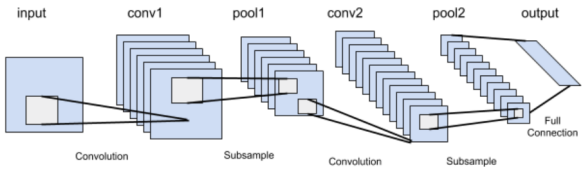

Fig. 2. Sample convolutional neural network model

Even though the institute has a large amount of image data of rice seed accessions, useful information such as variety group are missing for other accessions. Using the trained classifier, information can be provided to complete missing data. The provided data can again be used to improve the trained model.

II. Review of Related Literature

Institute of Computer Science of the University of the Philippines, Los Baos (ICS-UPLB) conducted researches concerning rice seeds using image data. A rice seed variety classification through color image analysis was done under the institute. It uses threshold based clustering in seed color profiling . Since rice seeds share a very similar color, the study resulted a model with low reliability in classification of rice seeds. To improved this approach, another study was conducted with partnership of IRRI, providing rice seed images. Additional feature was introduced to improve the classification. Color and shape features are assessed under the study to classify rice seed varieties. The features used that resulted with great variety were Hue, Saturation, and Intesity (HSI) for color and Length, Major Axis, and Perimeter for shape. With a great improvement from the previous study, difficulty in classifying almost similar varieties still exist. The reason for misclassification of rice seeds is still the similar features shared by some varieties [5].

Deep learning has been used for researches and applications. Some of them include deciphering spoken search queries and object recognition from images [6]. There are already existing programs that predict quantitative and categorical data using deep neural networks. One such study is an approach for generalized speech animation [9] where speech animation is produced given phoneme labels. The study predicts the lip and jaw movement of a speaker. Another example of the use of Deep Learning is a food recognition from images [10]. Given a smart-phone camera image, foods are recognized and calories are computed to produce a calorie-computed meal.

In Computational Biology, neural networks have been recognized since 1980 [6]. It was used for element structure prediction, and developed in the following years. Other studies focused on molecular activities [11], disease diagnosis through images, and effects-prediction directly from DNA sequences. Alipanahi [12], introduced a convolutional neural network for predicting sequence specificities of DNA and RNA-binding proteins prediction. DeepSEA was also introduced to predict regulatory sequence code from chro-matin-profiling data [13]. A study was conducted to classify rice types using CNN with transfer learning [14]. Rice types are categorized based on lengths, edge details, color, shape, etc. This study used image

data of of scanned rice seeds as inputs. Which is why the proposed convolutional neural network model will use rice images as input data for training and classification.

Rice researches are also conducted at the University of the Philippines Los Baos of the Institute of Computer Science.

Going with the trends in neural networks, Google Brain created a computational model for TensorFlow to introduce it as a platform in deep learning [15]. TensorFlow represents computations through dataflow graphs. It uses tensors as a data structure that flows through graphs and handles variables in computation that affects performance positively. To allow researchers to focus on the flow of the process in graphs, it uses a declarative programming paradigm [6]. Using TensorFlow as a platform in building the proposed neural network will contribute in developing the proposed system in terms of computation and visualization.

The existence of these studies and backgrounds provides sufficient ways and ideas to create a classifier model for rice image classification using a convolutional neural network (CNN) from Google TensorFlow.

III. Materials and Methods

The program was be implemented under the Ubuntu 18.04 operating system in an Anaconda environment using the Python 3.7 programming language, Keras for image preprocessing and TensorFlow. Along with these, openCV and numpy will be used for the image processing part for data inputs. The study is composed of three key phases in the program for rice-image classification: image processing for image data, training process, and analysis and evaluation.

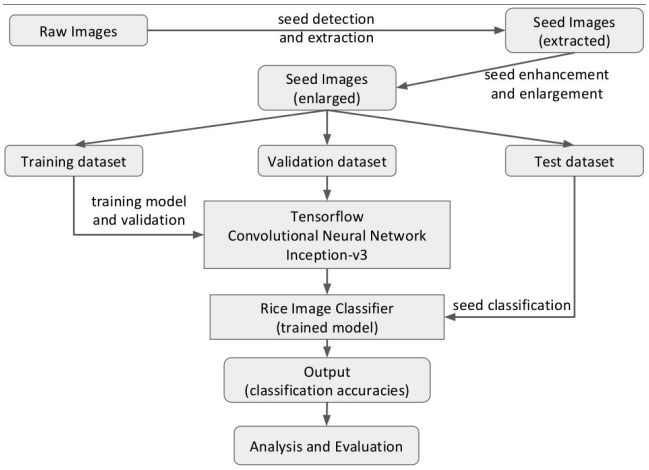

Fig. 3. Flowchart of the proposed program for image classifier

The program was tested on rice seed images captured by a scanner of IRRI. A raw image that is used as input includes ID, BARCODE, 5 grains without husk, and 10 grains with husk for a single accession. The subject for classification in this study are the rice seeds or grains with husk. Data are identified as feature data (rice seed images) and target data (classes). Datasets are categorized according to classes. Subpopulation from the variety group Indica (XI-1A, XI-1B, XI-2, XI-3, and XI-adm) was used as classes for the classifier model.

Fig. 4. Sample raw image data

Contrast-Limited Adaptive Histogram Equalization (CLAHE) and Bicubic Interpolation are used to enhance and enlarge, respectively, the input image for the classifier model. To get a better result, the input image was enhanced first before it was enlarged [16]. CLAHE is a contrast enhancement technique derived from Adaptive Histogram Equalization (AHE). It was first used for medical purposes to enhance low-contrast images such as portal films. Like a regular adaptive histogram equalization, CLAHE works with small partition of the image, called tiles, rather than the whole image. It applies histogram equalization on each tile of the image. This even the distribution of gray values to give importance of other features of the image. The output histogram of each tile will be shaped using the 'Distribution' parameter. This will provide a shaped histogram to produce better quality result. This will limit the contrast to avoid the amplifying noise that might be produced or present during the process. Neighboring tiles will then be combined using bilinear interpolation.

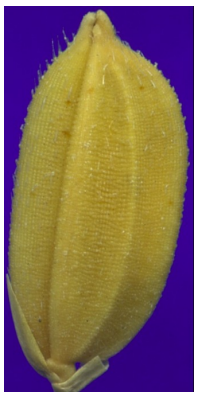

Fig. 5. Sample extracted image data

Bicubic Interpolation is being used for image resampling. It is an extension of bilinear interpolation that is widely used for digital image enlargement. Bicubic interpolation estimates and assigns the value of the unknown pixel by considering the nearest 16 pixels (4x4) into calculations. In this approach, 16 equations will be generated to determine 16 coefficients to assign a new pixel value. This method produces a smoother generated new and enlarged image [17].

The proposed Convolutional Neural Network Model are composed of three major layers: feature layer, hidden layers,

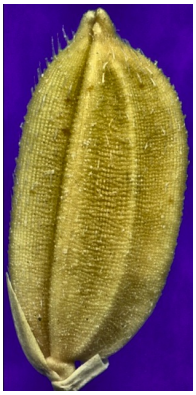

Fig. 6. Sample enhanced image data

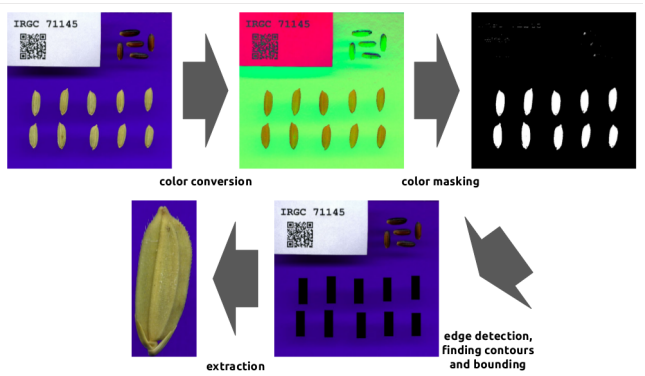

Fig. 7. Rice seed image detection and extraction

and output layer. Each layer is built with nodes that holds and manipulates data that are connected by edges where weights flow. The feature layer holds feature/input data (images).

The hidden layers are layers of the convolutional neural network where most of the computations will be done. The hidden layer transforms the inputs to be used by the output layer. Inception-V3 for CNN will be used as architecture for the classifier.

The output layer will then hold the result after the computations have been made by the neural net.

1) Image Detection and Extraction: This part will include the extraction, enhancement, and enlargement of the target images (rice seeds) from the raw images. Since the setup of the raw image intends to highlight the rice grains from its background, seeds can be detected within the raw image using colors. Color scale conversion will be helpful to separate the seeds from the its background. To detect colors accurately the image will be converted to HSV format from RGB color space. The converted image will be efficiently used in separating the seeds from the background through masking within the threshold of the background color of the target image. To get the seeds, points bounding the polygon will be detected and the rectangle bounding the seeds will be extracted from the raw image.

Extracted seed images will be enhanced using CLAHE. Conversion from RGB to Lab color space will be done in order to apply the adaptive histogram equalization. The lightness channel from the converted image will be enhanced using CLAHE to make hidden features of the image more visible. Then it will be combined again with the other channels from the converted Lab color space. The image will be reverted to RGB color space format. The enhanced image will undergo through enlargement using Bicubic Interpolation.

The enhanced and enlarged rice seed images will be used to train the convolutional neural network model. Rice seed images will be separated according to their designated classes (rice seed varieties).

2) Building Model and Training Process: Once the image processing for rice seed images is done, image data in each class (XI-1A, XI-1B, XI-2, XI-3, and XI-adm) will be divided into training, validation, and testing data; these are independent from each other to prevent biases for training, evaluation

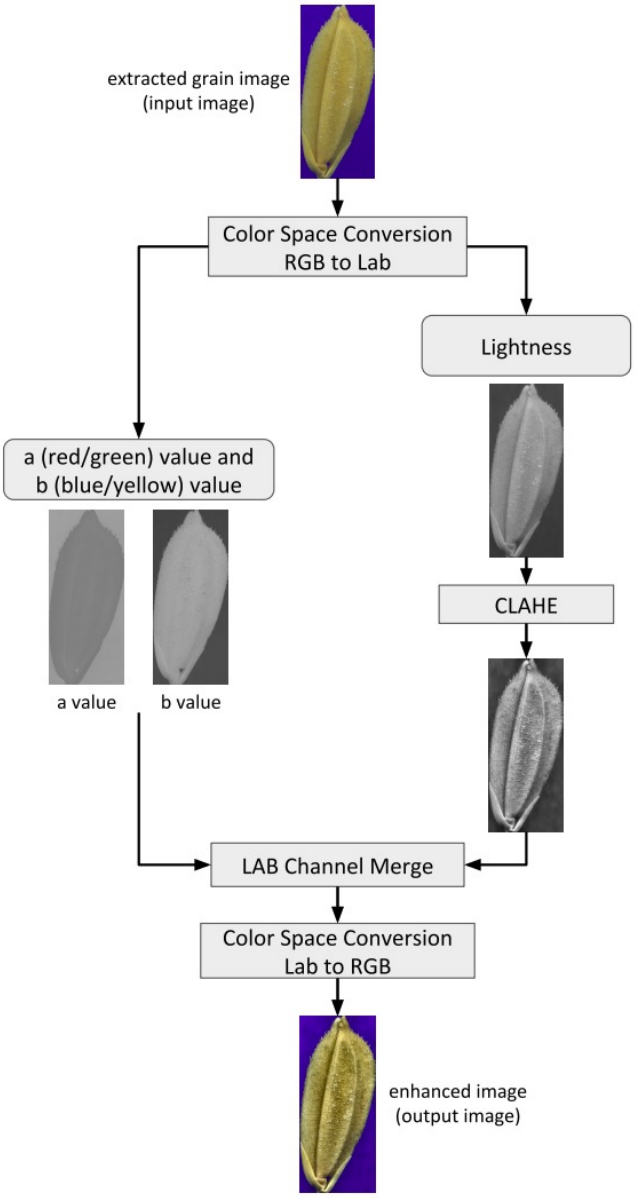

Fig. 8. Enhancement of extracted seed image using CLAHE

and prediction. 20 images form each class are reserved for the the test or prediction data. The remaining images are separated between training and validation data, 80% for train data and 20% for validation data. There are 5,544 image data for training (752, 448, 1296, 1144, and 1904 respectively) while there are 1,386 image data for validation (188, 112, 324, 286, and 476 respectively).

The program will assign input data (rice seed images) and target data (classes) for classification. Feature data will serve as the feature layer in the neural network. The data from the feature layer with weights and biases from connecting edges, will pass through the hidden layers and produce results for the output layer. The target data will serve as values to be classified. The training dataset will be fitted to the convolutional neural network to generate a learned model that will be used for classification. The resulting trained model classifier will then be used for rice seed image classification.

Before fitting training images to the model, images went through data generator for data augmentation. This will increase the diversity of available data for training since this study works with a limited amount of image data.

The architecture Inception-v3 using ImageNet was used for transfer learning. This will serve as the Base model. The pre-trained model using ImageNet weights is widely used for image classification because of its computational efficiency and use of less layers for less error rate.

The layers of the pre-trained model is freezed in training to keep the weights of the network unchanged. Freezing the layers of the pre-trained model is essential to be able to use the full potential of its knowledge from previous dataset. Following the Base model is the Fully-Connected Layers after flattening. Flattening operation transforms the shape of the tensors from the base model, ready to be fed to the Fully-Connected Layers. 1024 by 1024 fully-connected layers was used in the model. The following Dense layer has a ReLU activation function. The activation function Rectified Linear Unit (ReLU) defines a non-negativity constraint for the computation, this activation function is commonly used in neural network models, especially in Convolutional Neural Network. Dropout of 0.5 was also introduced to prevent overfitting.

For Prediction layer, Dense layer is set-up based on the number of classes available for predictions. Softmax activation function was used because it is great for probability analysis.

The final Fine-tuned Model has layers for Training Part and a Prediction Layer. Adam optimizer was used with learning rate of 0.00001. Compared to other compatible optimizer to our model, Adam optimizer is quicker. Categorical Cross-Entropy Loss, which is the combination of Softmax activation and Cross-Entropy Loss, was used for a multi-class classification. This loss function was used since for each image from the dataset represents only a single class. For every epoch of the training process, saving checkpoints was done.

3) Testing and Evaluation: The final part of the study is the testing and evaluation of the results. The program will be tested for classification of different rice seed images from different rice seed varieties. Test images from each class will be classified. The output accuracy of classes will be compared to the true value of class of the variety. The class with the highest accuracy from the result will be prioritized. Classifications will then be evaluated using Confusion Matrix.

IV. Results and Discussion

The classes for classification that used in this study was limited into Indica subpopulation (XI-1A, XI-1B, XI-2, XI-3, and XI-adm) as a result of limited image data. Extraction of rice seeds from each scanned raw image data expects 10 extracted images that will be used for the classification. For almost each extraction from raw images, 10 whole extracted rice seeds were produced. Rarely, there are instances that random parts of seed were extracted. These are the effect of near color scale of seeds from the background. Upon color scale conversion and masking, incomplete polygon was produced, thus it create separate images from a single seed. Few were not been detected properly that resulted cropped seed images.

Fig. 9. Sample image of cropped rice seed

Since this study works with difficult (almost identical images) data, image enhancement and enlargement will contribute to improve the data for a better classification. Extracted seed images from raw data was enhanced. CLAHE gave more depth and details to the images that help the classifier model get a better features. Enlarging the image data gave a more image values to be considered and computed for data generation for training.

Fig. 10. Sample images of rice seeds for each class after image enhancement

For Training part, the training and validation process reached 40 thousand epochs. The training process took about 35 days. Training and validation history were saved. The maximum training and validation accuracy and the minimum training and validation loss were recorded for every 500 epoch range. Training accuracy trends to gradually improves as training runs. Validation accuracy goes flat playing between 34%-37%. This may be improved by lowering the learning rate of the model for a slow but more detailed learning. Training and validation loss started nearby with each other until 20,000

epochs is reached. The second half of the line graph (see Fig. 13) shows training and validation loss separated. Validation loss is over the the training loss, thus, it indicates overfitting. Increasing the dropout can improve the results of this to lessen overfitting. The results seems not improving anymore, thus training the model can be stopped. Final accuracies are 47.65% (training) and 34.92% (validation) while the final losses are 1.236371579 (training) and 1.392561555 (validation).

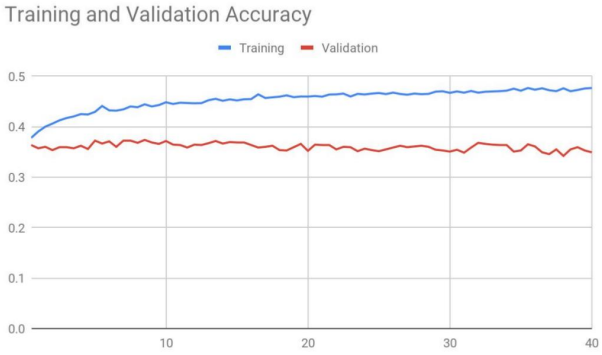

Fig. 11. Training and validation accuracy per 500 epochs (in thousands)

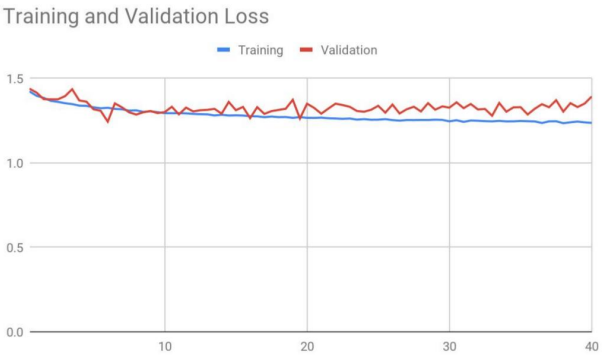

Fig. 12. Training and validation loss per 500 epochs (in thousands)

		prediction					
		XI-1A	XI-1B	XI-2	XI-3	XI-adm	TOTAL
actual	XI-1A	10	0	4	0	6	20
	XI-1B	3	0	4	3	10	20
	XI-2	5	0	0	4	11	20
	XI-3	6	0	1	2	11	20
	XI-adm	2	0	2	4	12	20
	TOTAL	26	0	11	13	50	100

TABLE I
CONFUSION MATRIX OF PREDICTION RESULTS, N=100

Using confusion matrix for the classification/prediction of rice seed images allows to compute for precision and recall of the prediction results for each classes, and the accuracy of the classifier. The prediction using the classifier produced poor results. For each class (XI-1A, XI-1B, XI-2, XI-3, and XI-adm) few to none were correctly classified, 10, 0, 0, 2, 12 respectively. These are out of 100 test images, with 20 images each class. Final prediction accuracy is 24%.

XI-1A	0.3846
XI-1B	0
XI-2	0
XI-3	0.1538
XI-adm	0.24

TABLE II
PRECISION FOR EACH CLASS

XI-1A	0.5
XI-1B	0
XI-2	0
XI-3	0.1
XI-adm	0.6

TABLE III
RECALL FOR EACH CLASS

The classifier resulted 2 classes (XI-1B and XI-2) with no correctly classified seed images. This automatically indicates that the 2 classes resulted a 100% of false negative prediction. The prediction results lean towards predicting Indica admix (XI-adm). Even if it resulted a 60% recall, its precision is low because of the amount of its false positive results from other classes. This is the effect of the limited data the study is working with. Instances of false positives points to XI-adm because of its bias with the data. The class has almost double of data of other classes.

Although the performance of the classification of this study is poor, this can be improved if given unbiased and more image data. The study works only with the Indica subpopulation as classes. With classes under the same parent makes the classification more difficult than having classes of general rice variety groups (Indica, Japonica, Aus).

V. CONCLUSION

In this study to validate if genomic variation of rice seeds translates to the corresponding characteristics or appearance (phenotype) of each seed, image data of Indica subpopultion (XI-1A, XI-1B, XI-2, XI-3, XI-adm) were used. For each raw image data of rice accessions, rice seeds were extracted and enhanced to give more details to improve classification. Convolutional Neural Network (CNN) was used to build the classifier model.

With the limited data that the study is working with, the classifier has a poor performance. There are no significant results that were made. The classifier accuracy is very low from the expectation. Thus, classifying rice seeds varieties

phenotypically using image data using the classifier in this study is having difficulty.

There are possible aspects that causes poor results of the classifier, one is the classes themselves. With only the Indica subpopulation, with the available image data, as classes poses a great challenge. In Fig. 1, the genetic variation of the classes is very near from each other. There are instances that they also overlap, and visually, an average person can not easily identify the rice seeds based on their respective classes. If given more general variation of rice varieties (Indica, Japonica, Aus, Basmati) as classes, it will greatly improve the classification.

Other aspect is the limited data. Table I shows that the prediction leans to XI-adm. Other classes gave amount of false negative to XI-adm. This is because it holds the largest number of image data. If given more and unbiased image data that the study can work with, the classification will produce unbiased results. Unbiased trend of results can be evaluated for further studies.

REFERENCES

[1] J. Li *et al.*, "The 3,000 rice genomes project: new opportunities and challenges for future rice research," *GigaScience*, vol. 3, no. 1, p. 8, 2014.

[2] N. Alexandrov *et al.*, "Snp-seek database of snps derived from 3000 rice genomes," *Nucleic Acids Res.*, vol. 43, no. 1, p. 7, 2015.

[3] L. Mansueto *et al.*, "Snp-seek ii: A resource for allele mining and analysis of big genomic data in oryza sativa," *Current Plant Biology*, vol. 7, no. 8, pp. 16–25, 2016.

[4] ——, "Rice snp-seek database update: New snps, indels, and queries," *Current Plant Biology*, 2017.

[5] T. A. L. M. Y. d. R. V. M. J.Iglesia, J. R. Tarong and T. K. Monserrat, "Assessment of different color and shape features in rice seed variety classification," 2014.

[6] L. Rampasek and A. Goldenberg, "Tensorflow: Biologys gateway to deep learning?" *Cell Systems 2*, 2016.

[7] J. Bankar and N. Gavai, "Convolutional neural network based inception v3 model for animal classification," *Computer and Communication Engineering*, vol. 7, 2018.

[8] W. Wang *et al.*, "Genomic variation in 3,010 diverse accessions of asian cultivated rice," *Nature*, vol. 557, pp. 43–49, 2018.

[9] S. Taylor *et al.*, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–11, 2017.

[10] P. Pouladzadeh and S. Shirmohammadi, "Mobile multi-food recognition using deep learning," *ACM Transactions on Multimedia Comput. Commun. Appl.*, vol. 13, no. 3, pp. 1–21, 2017.

[11] N. J. G.E. Dahl and R. Salakhutdinov, "Multi-task neural networks for qsar predictions," *ACM Transactions on Multimedia Comput. Commun. Appl.*, 2014.

[12] M. W. B. Alipanahi, A. Delong and B. Frey, "Predicting the sequence specificities of dna- and rna-binding proteins by deep learning," *Nat. Biotechnology*, vol. 33, pp. 831–838, 2015.

[13] J. Zhou and O.G.Troyanskaya, "Predicting effects of noncoding variants with deep learningbased sequence model," *Nat. Methods*, vol. 12, pp. 931–934, 2015.

[14] V. Patel and M. Joshi, "Convolutional neural network with transfer learning for rice type classification," *Machine Vision*, 2018.

[15] M. Abadi *et al.*, "A computational model for tensorflow: An introduction," pp. 1–7, 2017.

[16] H. Lin *et al.*, "A study of digital image enlargement and enhancement," *Mathematical Problems in Engineering*, 2014.

[17] R. B. M. Amara and T. Silva, "Slic based digital image enlargement," *Computer Vision and Pattern Recognition*, 2018.

Marvin Jerald F. Villador is the middle child of Perlito and Mina Villador. He is a servant of Christ. His ambition is to be a teacher someday. He loves coffee.