2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering | 978-1-6654-1569-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ECTIDAMTNCON51128.2021.9425696

Image Classification of Rice Leaf Diseases Using Random Forest Algorithm

Panuwat Mekha*
Department of Computer Science,
Faculty of Science, Maejo University
ChiangMai, Thailand
panuwat_m@mju.ac.th

Nutnicha Teeyasuksaet
The Fifth Regional Livestock Office, Department of Livestock Development
ChiangMai, Thailand
nutnichavet@gmail.com

***Abstract*—The problem of rice diseases around the world make to damage and fall into a large number of rice. Caused by many of types, such as; fungi, Bakteri and Viruses. which are the main causes of rice disease affected to farmers. The classification of rice can be classified into several methods. In this research, image classification is used to classify the data set of rice leaf diseases, such as; Brown Spot Rice disease (BSR), Brown Spot Rice disease (BSR), Bacterial Leaf Blight disease (BLB), which is the rice leaf disease with severe outbreaks around Thailand. Moreover, image processing technology in the classification types of rice leaf disease, such as; Random forest classification algorithm, Decision tree classification algorithm, Gradient Boosting classification algorithm and Naïve-Baye classification algorithm, which is measured by the accuracy, precision and recall of each algorithms. The best result of performance in the image classification of rice leaf diseases is random forest algorithm equal to 69.44 percent.**

***Keywords*—Rice Leaf Diseases, Algorithm, Image processing, Classification, Random Forest**

I. Introduction

Image classification is a statistical data technique. It aims to find the distinctive features of each image into class or category. By using mathematical principles and statistics to create a model to determine differential in features of the image point, this is a similarity of image point to the same class [7].

The data classification method is divided into 2 methods: Supervised Learning Classification and Unsupervised Learning Classification. This research purposes to extract the characteristics or features from image data. The information is used to identify class of images by supervised learning algorithm. Learning of image based on information and label [16], including the data mining process. We learn from the characteristics of image classification. We can apply a classification model for new pictures to predict that the class of data [14]. In particular, we take to extract the vector of features numerical image before applying the machine learning algorithm and apply to classified information application [10].

The main principal of classification for image data: 1) Abundance of images that are linked to data requires statistical and data mining model. 2) There are many tools to facilitate data mining users that perform to analysis within packages. It supports high-level programming language, such as python and R. However, the most important and necessary to adjust the parameters of data analysis[21].

Thus, our researcher developed an approach to image classification of rice leaf diseases[1] using random forest algorithm. And we compared with other classification methods, such as; decision tree, gradient boosting and naive bayes. Our experimental results show high accuracy score for image classification method, which simplifier and faster than other methods [17].

II. Materials and Methods

Knime image processing is technology of data analysis for the classification of images. And the most important in process can be link node of data to step by step. Start with loading data into visual system. It can be modify properties of the image and extract specific features (feature extraction) [20] to create a data table (Table of attributes and value data). This is applied to the machine learning algorithms [3]. Our research focus on classification of rice leaf Diseases. Using a database from UC Irvine Machine Learning Repository and available at following link: (https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases). It classified the rice leaf disease into 3 classes:

A. Brown Spot Rice disease: BSR

BSR has been widely overlooked as one of the most widespread and most destructive rice diseases. Brown spot is a fungal disease that infects coleoptiles, trees, sheaths of the leaves, roots of panicles. It is caused by fungus (Helminthosporium oryzae) and fungus can be blown away with the wind and drop on young rice seeds will destroy the seeds and make them white and quality is not good. The most prominent damage is the many broad spots on the leaves that can destroy the whole crop. When infection occurs in seed,

unfilled grains are shaped, or spotted or discolored seeds.

Fig. 1. Brown spot of rice disease

B. *Leaf Smut Rice disease: LSR*

It is a common epidemic in high humidity. It is caused by fungus (Ustilaginoidea virens). This fungus destroys rice seeds then creates many fibers and seeds in that seed until it emerge out of the seed group of fibers and reproduction on this seed to be dark green color. However, it found on some grass species and destroyed by the fungus.

Fig. 2. Leaf smut of rice disease

C. *Bacterial Leaf Blight disease: BLB*

BLB is caused by Xanthomonas Oryzae. It causes seedlings to get wilted and leaves to yellow and die [15]. The disease most often occurs in regions where the contaminated plants have weeds and stubbles. It may occur in both tropical and temperate habitats, particularly in lowland irrigated and rainfed regions. The disease usually encourages temperatures at 25−34 °C with relative humidity above 70 percent. It is widely observed when there are high winds and persistent heavy rains, enabling the disease-causing bacteria to quickly spread on lesions of infected plants by ooze droplets. Bacterial blight may be serious on high nitrogen fertilization in susceptible rice varieties [4].

Fig. 3. Bacterial leaf blight of rice disease

Image series of each type for rice leaf disease equal to 120 images, show the characteristics of each type of rice leaf disease in JPEG format and divide to the training dataset and testing dataset equal to 70:30. The data management of training datasets will be partitioned randomly to create a model for predicting the type of rice leaf disease [19] in which the images filename will be define with the first 3 characters to help identify the type of rice leaf disease and create the target column in the learning process. As follows, Brown Spot Rice disease (BSR), Leaf Smut Rice disease (LSR) and Bacterial Leaf Blight disease (BLB), as well as management of training dataset and testing dataset [6], which has increased processing processes such as identification the position of the point indicating the type of disease on the rice leaf disease image, etc. [18]. The predictive model is created and evaluated the effectiveness of the model as Fig. 4.

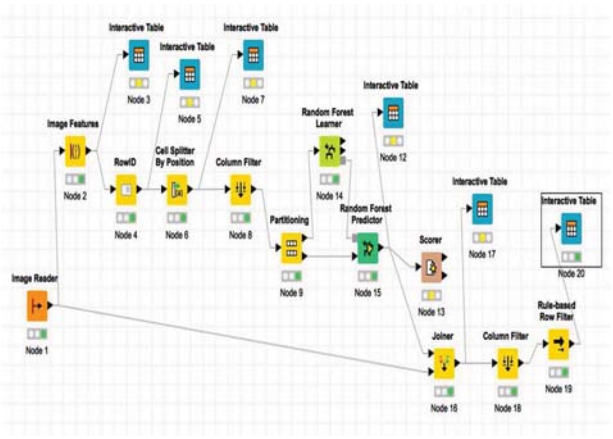

Fig. 4. Flow algorithm for image classification of rice leaf diseases

III. Algorithm for Image Classification

- Random forest

Random forest or collection of decision tree is a combined regression learning method [9] for classification and other methods [22]. Random forest algorithm creates by a many of decision trees for training and showing the results in the prediction of classification (regression) of each tree [2]. Random forest algorithm is a predictor consisting of properties based on random regression of trees. $\{\boldsymbol{r_n(x, \theta_m, D_n)}, \boldsymbol{m \geq 1}\}$, where $\boldsymbol{\theta_1, \theta_2}$, ... with ID is the result of random variable $\boldsymbol{\theta}$ in which the randomized trees will be combined for regression estimation [13].

$$\bar{\boldsymbol{r}}_{\boldsymbol{n}}(\boldsymbol{X}, \boldsymbol{D_n}) = \boldsymbol{E_\theta}[\boldsymbol{r_n}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{D_n})], \quad (1)$$

Where $\boldsymbol{E_\theta}$ is the expectation of random parameters according to conditions of given X and $\boldsymbol{D_n}$ dataset.

- Decision tree

Decision tree is a tool that supports decision making in tree pattern, which considers the possible of data such as; data event, data type etc. It is a way to show the algorithm under control conditions [14].
Decision tree is the most popular algorithms used in research and specific operations of data analysis. Including decision support to identify data trends that will achieve the target by using information gain, which is the criteria used to evaluate the information value by each attribute with key point from data. In addition, the measurement of certainty or uncertainty random variables X define as entropy for classification problems of data can be calculated according to equation:

$$\boldsymbol{H(X) = \mathbb{E}_X[I(x)] = -\sum_{x\epsilon X} p(x) \log p(x)} \quad (2)$$

By calculating the entropy value of each attribute. We can calculate the criteria based on the data classification model from the training dataset to classify the data.

- Gradient boosting

Guardian boosting is the machine learning technique for problems with regression analysis and data classification[12], which is a model development based on predictions in decision trees of creating models for increased efficiency in each function [5].

Input: training dataset $\{(\boldsymbol{x_i, y_i})\}_{i=1}^{n}$, Differential loss function: $\boldsymbol{L\big(y, F(x)\big)}$, number of repetition M

Algorithm:

1. Start with static number

$$\boldsymbol{F_0(x) = arg_\gamma \, min \sum_{i=1}^{n} L(y_i, \gamma)} \quad (3)$$

2. Value of m=1 to M

2.1 Calculate:

$$\boldsymbol{r_{im} = -\left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n} \quad (4)$$

2.2 Adjust the learning based to tree $\boldsymbol{h_m(x)}$ to calculate the remaining values by the training dataset. $\{(\boldsymbol{x_i, r_{im}})\}_{i=1}^{n}$,

2.3 Calculate $\boldsymbol{\gamma_m}$ with optimize to 1D problem

$$\boldsymbol{\gamma_m = arg_\gamma \, min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))} \quad (5)$$

2.4 Format update:

$$\boldsymbol{F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)} \quad (6)$$

Display $\boldsymbol{F_M(x)}$

- Naive Bayes

Machine learning with the Naive Bayes is a probability characteristic based on reliable Bayesian theory and independent assumptions between characteristics of data. Naive Bayes classifier can adjust the size of the data to analysis using by linear parameters for a number of variables (characteristics and prediction) to learn from the training dataset. Maximum-likelihood can estimate to parameters from the probability of data. Naive Bay algorithm is the use of conditional probability models, which applies to classification problems by vector $\boldsymbol{x = (x_1, \dots, x_n)}$and represented by $\boldsymbol{n}$ features (Independent variables) can determine the probability.

$$\boldsymbol{p(C_k|x_1, \dots, x_n)} \quad (7)$$

with probability of each $\boldsymbol{k}$ or type of $\boldsymbol{C_k}$

This formula is the number of attributes n. It is a large of the characteristics can be used in many values and reference model to the probability table [23]. Therefore, we define format that will make it more flexible using by Bayes's theory for conditional probability can be expressed as follows:

$$\boldsymbol{p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}} \quad (8)$$

Bayesian probability as follows:

$$\boldsymbol{posterior = \frac{prior \times likelihood}{evidence}} \quad (9)$$

IV. Results

Our research show the algorithms used to classify type of rice leaf diseases for image classification: random forest algorithm, decision tree algorithm, gradient boosting algorithm and naive bayes algorithm [11] which the efficiency of the accuracy (precision and recall) of each algorithm [8]. It can be seen that the random forest algorithm has the highest accuracy equal to 69.44 percent, followed by the gradient boosting algorithm with accuracy Equal to 66.67 percent and naive bayes algorithm has the lowest accuracy equal to 36.11 percent. And the highest precision of each classes: Gradient Boosting has equal to 69.23 in BLB and Random Forest has equal to 80.00 in BSR and Decision Tree has equal to 70.00 in LSR. Moreover, the highest recall of each classes: Naive Bayes has equal to 91.67 in BLB and Random Forest has equal to 66.67 in BSR and Random Forest has equal to 83.33 in LSR as in TABLE I

TABLE I PREDICTION PERFORMANCE OF IMAGE CLASSIFICATION METHODS

Methods	Precision			Recall			ACC
	BLB	BSR	LSR	BLB	BSR	LSR	
Random Forest	63.64	**80.00**	66.67	58.33	**66.67**	**83.33**	**69.44**
Decision Tree	60.00	63.64	**70.00**	75.00	58.33	58.33	63.89
Gradient Boosting	**69.23**	63.64	66.67	75.00	58.33	66.67	66.67
Naive Bayes	34.38	0.00	66.67	**91.67**	0.00	16.67	36.11

V. DISCUSSION

The results of our research, it can be concluded that the random forest algorithm is the highest of accuracy for image classification by compare with the other algorithm classification. And show the effectiveness of the classification of each rice leaf disease using the random forest algorithm that can display picture of various rice leaf diseases, including with target and prediction result as shown in Fig. 5-7

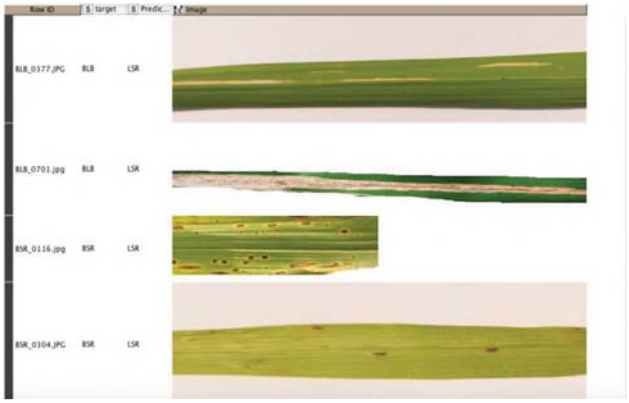

Fig. 5. Target (BLB, BSR) and Prediction (LSR) of Rice Leaf Diseases

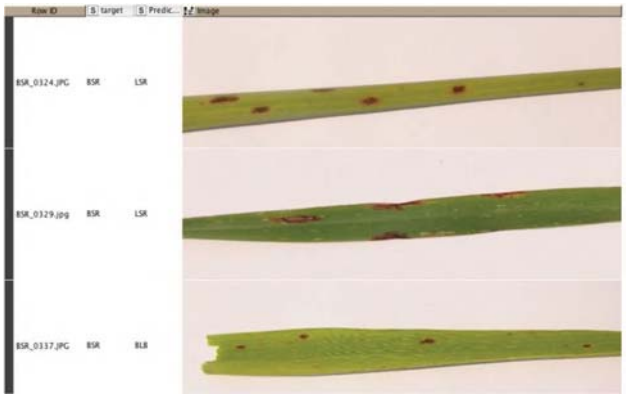

Fig. 6. Target (BSR) and Prediction (LSR, BLB) of Rice Leaf Diseases

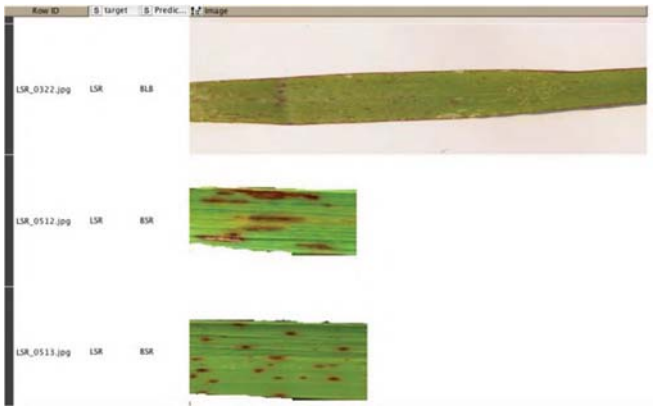

Fig. 7. Target (LSR) and Prediction (BLB, BSR) of Rice Leaf Diseases

VI. CONCULSION

This research presents the algorithm that can be used to classify rice leaf diseases by images. It was found that the random forest algorithm has the highest accuracy equal to 69.44 percent of image classification. Moreover, this research show comparison of efficiency of rice leaf disease classification by image of each algorithm, which can be seen the flow algorithm for image classification of rice leaf diseases image processing by knime method that can be analyzed and extracted the unique characteristics of each type of rice leaf disease.

ACKNOWLEDGEMENT

This our research was supported by Department of Computer Science, Faculty of Science, Maejo University and The Fifth Regional Livestock Office, Department of Livestock Development.

REFERENCES

[1] Archana, K. S., & Sahayadhas, A. (2018). Automatic rice leaf disease segmentation using image processing techniques. *Int. J. Eng. Technol*, *7*(3.27), 182-185.

[2] Bosch, A., Zisserman, A., & Munoz, X. (2007, October). Image classification using random forests and ferns. In *2007 IEEE 11th international conference on computer vision* (pp. 1-8). Ieee.

[3] Chaudhary, A., Kolhe, S., & Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, *3*(4), 215-222.

[4] Gnanamanickam, S. S., Priyadarisini, V. B., Narayanan, N. N., Vasudevan, P., & Kavitha, S. (1999). An overview of bacterial blight disease of rice and strategies for its management. *Current Science*, 1435-1444.

[5] Li, P., Wu, Q., & Burges, C. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in neural information processing systems*, *20*, 897-904.

[6] Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823-870.

[7] Mekha, P., Osathanunkul, K., & Teeyasuksaet, N. (2016, December). Gene classification of dengue virus type based on codon usage. In *2016 International Computer Science and Engineering Conference (ICSEC)* (pp. 1-6). IEEE.

[8] Mekha, P., Osathanunkul, K., & Teeyasuksaet, N. (2017). Comparison of Gene Classification Methods for Dengue Virus Type Based on Codon Usage. *FEU Academic Review*, *11*, 7-7.

[9] Mekha, P., & Teeyasuksaet, N. (2015, November). Using random forest based on codon usage for predicting Human Leukocyte Antigen gene. In *2015 International Computer Science and Engineering Conference (ICSEC)* (pp. 1-6). IEEE.

[10] Mekha, P., & Teeyasuksaet, N. (2019, January). Deep learning algorithms for predicting breast cancer based on tumor cells. In *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)* (pp. 343-346). IEEE.

[11] Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60).

[12] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

[13] Ok, A. O., Akar, O., & Gungor, O. (2012). Evaluation of random forest method for agricultural crop classification. *European Journal of Remote Sensing*, *45*(1), 421-432.

[14] Patil, J. K., & Kumar, R. (2011). Advances in image processing for detection of plant diseases. *Journal of Advanced Bioinformatics Applications and Research*, *2*(2), 135-141.

[15] Perumalsamy, S., Bharani, M., Sudha, M., Nagarajan, P., Arul, L., Saraswathi, R., ... & Ramalingam, J. (2010). Functional marker-assisted selection for bacterial leaf blight resistance genes in rice (Oryza sativa L.). *Plant breeding*, *129*(4), 400-406.

[16] Phadikar, S., & Sil, J. (2008, December). Rice disease identification using pattern recognition techniques. In *2008 11th International Conference on Computer and Information Technology* (pp. 420-423). IEEE.

[17] Prajapati, H. B., Shah, J. P., & Dabhi, V. K. (2017). Detection and classification of rice plant diseases. *Intelligent Decision Technologies*, *11*(3), 357-373.

[18] Rastogi, A., Arora, R., & Sharma, S. (2015, February). Leaf disease detection and grading using computer vision technology & fuzzy logic. In *2015 2nd international conference on signal processing and integrated networks (SPIN)* (pp. 500-505). IEEE.

[19] Sanyal, P., & Patel, S. C. (2008). Pattern recognition method to detect two diseases in rice plants. *The Imaging Science Journal*, *56*(6), 319-325.

[20] Shoombuatong, W., Mekha, P., Waiyamai, K., Cheevadhanarak, S., & Chaijaruwanicha, J. (2013). Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *ScienceAsia*, *39*(1), 42-49.

[21] Shoombuatong, W., Mekha, P., & Chaijaruwanich, J. (2015). Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. *International journal of data mining and bioinformatics*, *13*(3), 211-224.

[22] Xu, B., Ye, Y., & Nie, L. (2012, June). An improved random forest classifier for image classification. In *2012 IEEE International Conference on Information and Automation* (pp. 795-800). IEEE.

[23] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Information Sciences*, *179*(19), 3218-3229.