Rice Leaf Disease Detection Using Machine Learning Techniques

Conference Paper · December 2019

DOI: 10.1109/STI47673.2019.9068096

4 authors:

Kawcher Ahmed
North South University
1 PUBLICATION **15** CITATIONS
SEE PROFILE

Tasmia Rahman Shahidi
North South University
2 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Syed Md. Irfanul Alam
North South University
1 PUBLICATION **15** CITATIONS
SEE PROFILE

Sifat Momen
North South University
45 PUBLICATIONS **308** CITATIONS
SEE PROFILE

2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 24-25 December, Dhaka

Rice Leaf Disease Detection Using Machine Learning Techniques

Kawcher Ahmed, Tasmia Rahman Shahidi, Syed Md. Irfanul Alam and Sifat Momen

Department of Electrical and Computer Engineering
North South University
Bashundhara, Dhaka - 1229, Bangladesh
Email: {kawcher.ahmed, tasmia.shahidi, syed.irfan, sifat.momen }@northsouth.edu

***Abstract*—As one of the top ten rice producing and consuming countries in the world, Bangladesh depends greatly on rice for its economy and for meeting its food demands. To ensure healthy and proper growth of the rice plants it is essential to detect any disease in time and prior to applying required treatment to the affected plants. Since manual detection of diseases costs a large amount of time and labour, it is inevitably prudent to have an automated system. This paper presents a rice leaf disease detection system using machine learning approaches. Three of the most common rice plant diseases namely leaf smut, bacterial leaf blight and brown spot diseases are detected in this work. Clear images of affected rice leaves with white background were used as the input. After necessary pre-processing, the dataset was trained on with a range of different machine learning algorithms including that of KNN(K-Nearest Neighbour), J48(Decision Tree), Naive Bayes and Logistic Regression. Decision tree algorithm, after 10-fold cross validation, achieved an accuracy of over 97% when applied on the test dataset.**

***Index Terms*—Dataset, Disease detection, machine learning, rice leaf, supervised learning**

I. Introduction

Bangladesh achieved its highest GDP, BDT 10.73 billion in 2019 [1], from agricultural sector. Half of the agricultural GDP is provided by rice production. This consequently also contributes towards almost half of the rural employment (48%) [2]. While providing a vital role in the country' s economy, rice serves as a staple food for the mass population and provides two-thirds of the per capita daily calorie intake. As per the USDA' s report, total rice yielding area and corresponding production are projected to be 11.8 million hectares and 35.3 million metric tons respectively for 2019-2020 (May to April) [3]. These economic turnouts clearly indicate that proper rice cultivation is a high priority for Bangladesh. Disease free rice cultivation would play a dominant role in ensuring stable economic growth and maintainin the desired targets.

Moreover, to keep pace with the emerging fourth industrial revolution, Bangladesh needs to work for its industrial advancements which will involve smart systems that can take decisions without any human interventions. To that end, we have come up with an automated system using machine learning techniques, a system that will contribute in country' s agricultural development by automatically identifying and classifying diseases from the images of rice leaves.

Twenty rice diseases were revealed in Bangladesh from a survey conducted in 1979-1981 [4], among which 13 diseases were detected as the major ones. Rice blast and brown spot were considered as the most prominent diseases then, but now brown spot and bacterial blight are considered as the most prominent and dangerous rice diseases [5]. In this paper, we have focused on the identification of three rice leaf disease detection (bacterial blight, brown spot and leaf smut). The reason for choosing these three diseases is the prevalence of these diseases in Bangladesh.

These three different diseases have their characteristic patterns and shapes. The features of the diseases [6] is described below and illustrated in Fig.1:

- Leaf smut: small black linear lesions on leaf blades, leaf tips may turn grey and dry.
- Bacterial blight: elongated lesions near the leaf tips and margins, and turns white to yellow and then grey due to fungal attack.
- Brown spot: dark brown colored and round to oval shaped lesions on rice leaves.

Fig. 1. (a) Leaf Smut, (b) Bacterial leaf blight, (c) Brown Spot disease

Monitoring the diseases, their occurrences and frequencies are very important for early detection of the affected plants, their timely treatment, and most importantly, for planning future strategies to prevent the diseases to minimize the losses. Traditionally, crop disease management in Bangladesh is carried out by manual detection of any irregularity in plants, then classification of that irregularity as disease by experts and finally recommending appropriate treatment. This series of tasks becomes very challenging while considering large farms. It causes additional time and labour as well. On the contrary,

978-1-7281-0827-8/19/$31.00 ©2019 IEEE

taking the images of the affected area of the plants and testing with a pre-trained model gives a way better detection and classification of diseases.

This paper proposes such an approach that makes disease prediction and classification of the three mentioned rice diseases. The novelty of the paper lies in the detection of rice leaf diseases using machine learning approaches with high accuracy.

The dataset used in this work has been collected from [7]. We trained our model with four suitable machine learning classification algorithms. The comparative study among the four has been analyzed in section IV for better representation and understanding of the efficiency and accuracy of the model trained with different algorithms. The proposed solution of this paper has been described in section III including the detailed analysis of the dataset including data collection, selection, preprocessing and attributes selection.

II. Literature Review

Sladojevic and colleagues [8] aimed to detect plant diseases using Deep Learning techniques that will help the farmers to quickly and easily detect diseases which in turn would enable the farmers to take proper steps at early stage. They used 2589 original images in performing tests and 30880 images for training their model using the Caffe deep learning framework [9]. For achieving a higher accuracy in evaluating a predictive model, the authors used 10-fold cross validation technique on their dataset. The accuracy of prediction of this model is 96.77%.

Depending on only the extracted percentage of the RGB value of the affected area of rice leaf using image processing, a model was developed in [10] to classify the disease. The RGB percentages were fed into Naive Bayes classifier to finally categorize the diseases into three disease classes: Bacterial leaf blight, Rice blast and Brown spot. The accuracy of the model to classify the diseases is over 89%.

A higher accuracy was found in paper [11] where a plant disease detection model was developed using CNN. This model can identify 13 different types of diseases of plants. The final accuracy achieved from this model is 96.3%.

In another study [12], the affected parts were separated from the rice leaf surface using K-means clustering and the model was then trained with SVM using color, texture and shape as the classifying features.

Maniyath et al. used random forest, an ensemble learning method, to classify between healthy and diseased leaf [13]. For extracting the features of an image, the authors used Histogram of Oriented Gradient (HOG). Their work has claimed an accuracy 92.33%.

Image Processing and machine learning techniques were also used in [14] for the detection and classification of rice plant diseases. Authors of this paper used K-means clustering for the segmentation of the diseased area of the rice leaves and Support Vector Machine (SVM) for classification. They achieved a final accuracy, 93.33% and 73.33% on training and test dataset respectively. The same dataset was also used in our work but our methodology resulted in a higher accuracy both in training and test dataset.

III. Proposed Work

The main idea of this work is to create a rice leaf disease detection model using machine learning algorithms that can be helpful for disease recognition. The data for this task is collected from the UCI Machine Learning Repository [14]. WEKA (Waikato Environment for Knowledge Analysis) [15], an open source machine learning software, has been used to apply different machine learning algorithms to train our model.

A. Block Diagram

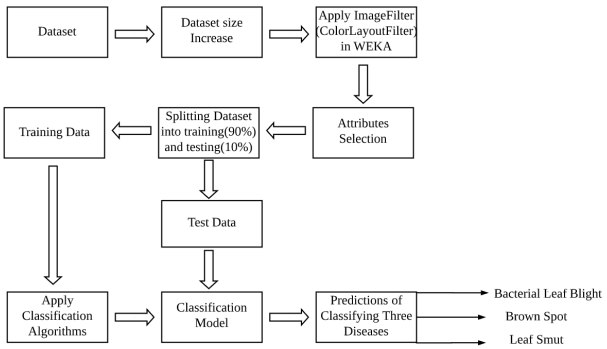

Fig. 2. Block Diagram of the entire work

Fig.2 represents the workflow of the entire work. The rice leaf disease dataset was created by Harshadkumar, et al. [14]. The dataset was created manually by separating infected leaves into three different disease classes. There are three diseases: Bacterial leaf blight, Brown spot, and Leaf smut, each having 40 images. The format of each image is .jpg. The size of the dataset was increased to 480 by image augmentation. After that, the *ColorLayoutFilter* image filtering was used to convert the images into features which added 35 attributes in the dataset. Then correlation based attributes selection technique was used to determine prominent attributes. The dataset was then split into two parts: training set, which comprises of the 90% data and test set, which comprises of the remaining 10%. Finally, four different classification algorithms were used which generated different results.

B. Dataset

1) Dataset Example: Fig.3 shows images of three types of rice leaf disease of our dataset.

2) Attribute Selection: From the added attributes after applying image filter, five attributes were selected using Correlation Based Feature Selection [16] (CorrelationAttributeEval) technique in Weka which were relevant to making predictions. CorrelationAttributeEval technique calculates the correlation between each attribute and the class. It can only be used with a Ranker Search Method, [17] that lists the attributes in a rank order after evaluating each of them. In this work, we achieved maximum accuracy after selecting best five attributes. Fig.4 shows the selected attributes in this work.

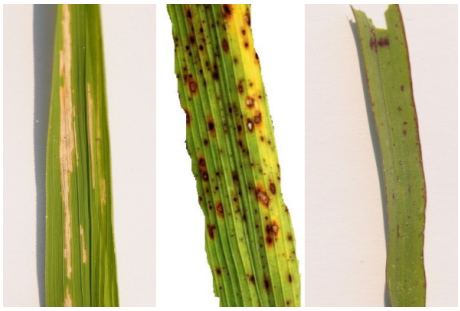

Fig. 3. (a)Bacterial blight, (b)Brown spot, (c)Leaf smut

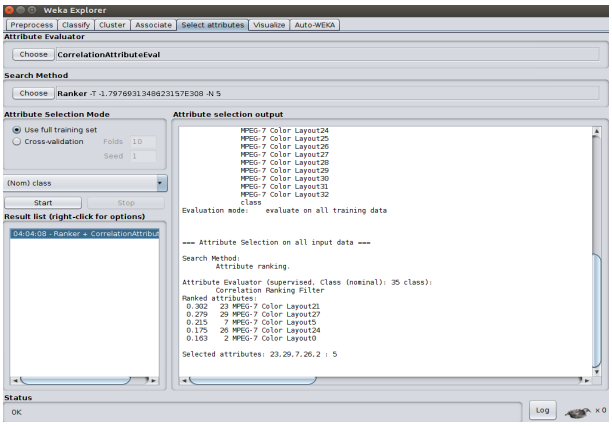

Fig. 4. Selected Attributes

3) Splitting Dataset: In our work, the dataset size was 480. Using *resample filter*, the dataset was divided into two parts: training and test sets where training data contains 432 instances (90% of the 480) and test data contains 48 instances (10% of the 480). The *resample filter* ensures that no instance in the test dataset is present in the training dataset.

C. Classifiers used

Supervised classification algorithms were applied on Rice Leaf Disease Dataset to detect three diseases of rice leaf. In this work, four classification algorithms were applied to detect the diseases. At first we applied classification algorithms before attributes selection and achieved different results for four algorithms. After that we applied classification algorithms using five selected relevant attributes with applying 10-fold cross validation and achieved better results.

1) ***Logistic Regression****:* Logistic regression can only be applied if the target class has categorical values. As the aim was to predict and categorize the disease of the affected rice leaf, logistic regression was a suitable model to train our dataset with. This paper works on predicting three distinct diseases, so we used multiclass logistic regression. In multiclass logistic regression, for given i classes, i different binary classifiers $h_\theta^{(i)}x$ are trained for each class i to determine the probability of y, the target class [18]. Then, a new input x can be predicted to belong to the class i if it maximizes $maxh_\theta^{(i)}x$:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

where,

$$g(\theta^T x) = g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

h_θ is the hypothesis that determines the predicted output; y is predicted to be 1 if h_θ is greater or equal to 0.5 and it is predicted to be 0 if h_θ is less than 0.5. $g(z)$ maps real valued numbers within a range of 0 to 1 and it plots an S-shape curve as Fig.5:

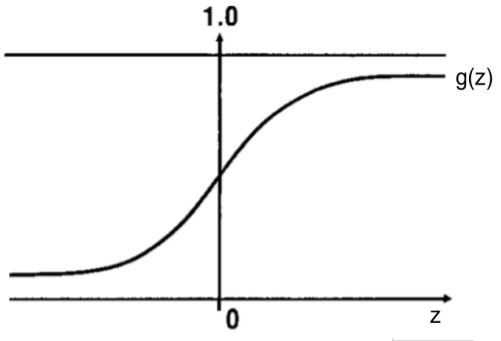

Fig. 5. S-shaped sigmoid function

Here we performed 10-fold cross validation using logistic regression algorithm and achieved 75.463% accuracy on training set and 70.8333% accuracy on test set in detecting three diseases.

2) ***K-Nearest Neighbour****:* Like the logistic regression, K-NN [19] also works well for discrete target classes. It calculates the distances of the query point from each of the instances and finds the K minimum distances that is, it determines the K nearest neighbours for the query point from which it can predict the class of the query point. The value of K needs to be chosen by inspecting the data; in case, we found when K = 1 the accuracy is 98.8426% on training set and 91.6667% on testing set after performing 10-fold cross validation.And when K=3 the accuracy is 85.6481% on training set and 72.9167% on testing set after performing 10-fold cross validation. We found, if the value of K is increased then accuracy is decreased.

3) ***Decision Tree****:* Decision tree [20] is one of the most commonly used machine learning classifiers. Taking the best suitable attribute at the root, this algorithm breaks the dataset into partitions. The goal of the partition is to unmix the dataset. The splitting iterates until eventually the partitions group the data such that they are homogeneous. Iterative dichotomiser 3 (ID3), which uses a greedy approach, is the core algorithm for decision tree. In this approach, entropy and information gain, concepts borrowed from information theory, are used for constructing the tree. Entropy measures the impurity of arbitrary attributes; zero entropy means all instances belong to

the same class. As entropy becomes more and more positive, the instances become more and more heterogeneous.

$$E = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{3}$$

Here c is the number of classes.

Information gain allows to determine attribute to be selected as the next node in the tree. The attribute with the most information gain would be selected for this purpose.

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \tag{4}$$

Here, A is the known attribute and S_v is the subset of A for which, A has the value v.

Using five selected attributes decision tree algorithm was able to correctly classify 94.9074% data on training set where 10-fold cross validation is performed. The model achieved 97.9167% accuracy on test data.

*4) **Naive Bayes Classifier**:* Naive Bayes [21] algorithm is a probabilistic algorithm that is based on Baye's theorem. Based on this theorem, the best hypothesis [16] is chosen based on equation 5

$$\hat{y} = argmax P(y) \prod_{i=1}^{n} (P(x_i|y)) \tag{5}$$

In this work Naive Bayes algorithm achieved the lowest accuracy to correctly classify three diseases.

IV. Results and Discussion

Training and Test dataset contains 432 and 48 instances respectively and 5 attributes were chosen. Table I shows the accuracy of four classification algorithms after performing 10-fold cross validation on training data (90% of dataset)and test data (10% of dataset) where best five attributes were selected.

TABLE I
ACCURACY ON TRAINING AND TEST DATASET

Algorithms	***Accuracy On Training Set***	***Accuracy On Testing Set***
Logistic Regression	75.463 %	70.8333 %
KNN(K=1)	98.8426 %	91.6667 %
KNN(K=3)	85.6481 %	72.9167 %
Decision Tree (j48)	94.9074 %	97.9167 %
Naive Bayes	58.7963 %	50 %

The comparison between the accuracy of the four classification algorithms are represented in Figure 6. Besides accuracy, other performance measures like TPR(True Positive Rate), FPR(False Positive Rate), Precision value (Positive Predictive Value), Recall value (Sensitivity), F-Measure and AUC(Area Under ROC) are also evaluated to compare among the four algorithms and it reveals in Table II and Table III that in each case, decision tree algorithm outperforms all other algorithms in detecting and classifying the diseases.

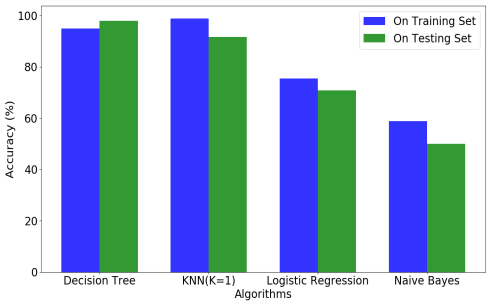

Fig. 6. Comparison between algorithms

TABLE II
DETAILED EVALUATION ON TRAINING SET

Algorithms	***TP Rate***	***FP Rate***	***Precision***	***Recall***	***F-Measure***	***Area Under ROC***
Logistic Regression	0.755	0.123	0.757	0.755	0.754	0.882
KNN(K=1)	0.988	0.006	0.988	0.988	0.988	0.987
KNN(K=3)	0.856	0.072	0.858	0.856	0.857	0.978
Decision Tree (j48)	0.949	0.026	0.950	0.949	0.949	0.980
Naive Bayes	0.588	0.207	0.670	0.588	0.580	0.816

TABLE III
DETAILED EVALUATION ON TEST DATA

Algorithms	***TP Rate***	***FP Rate***	***Precision***	***Recall***	***F-Measure***	***Area Under ROC***
Logistic Regression	0.708	0.141	0.720	0.708	0.699	0.882
KNN(K=1)	0.917	0.042	0.933	0.917	0.915	0.899
KNN(K=3)	0.729	0.131	0.755	0.729	0.731	0.931
Decision Tree (j48)	0.979	0.009	0.980	0.979	0.979	0.985
Naive Bayes	0.500	0.238	0.554	0.500	0.477	0.782

The empirical error rate for a classifier is given by equation 6. This is then used to determine the accuracy.

$$Error = \frac{\Sigma_{i=1}^{n} \delta(h(x), y)}{n} \tag{6}$$

where

$$\delta(h(x), y) = 1 \text{ if } y \neq h(x)$$

and

$$\delta(h(x), y) = 0 \text{ if } y = h(x)$$

V. Conclusion and Future Work

This paper presents a machine learning approach to detect three different rice leaf diseases: leaf smut, bacterial leaf blight and brown spot disease. The work has significant economic

importance for Bangladesh. A comparison between four machine learning algorithms (including that of KNN, Decision tree, Logistic regression and Naive Bayes) in the realms of rice leaf disease detection has been made. The algorithms predicted the rice leaf diseases with varying degrees of accuracy. It was found that decision tree performed the best with 97.9167% accuracy on test data. Having thus identified a near-optimal algorithm, we hope to extend this study further as higher quality datasets become available in the future.

For our future work, we plan to explore the effectiveness of ensemble learning methods on this dataset.

REFERENCES

[1] "Bangladesh gdp from agriculture." https://tradingeconomics.com/bangladesh/gdp-from-agriculture. Accessed: 2019-08-25.

[2] T. Akter, M. T. Parvin, F. A. Mila, and A. Nahar, "Factors determining the profitability of rice farming in bangladesh," *Journal of the Bangladesh Agricultural University*, vol. 17, no. 1, pp. 86–91, 2019.

[3] "Usda: Rice output continues to see growth." https://www.dhakatribune.com/business/economy/2019/04/09/usda-rice-output-continues-to-see-growth. Accessed: 2019-08-25.

[4] S. Miah, A. Shahjahan, M. Hossain, and N. Sharma, "A survey of rice diseases in bangladesh," *International Journal of Pest Management*, vol. 31, no. 3, pp. 208–213, 1985.

[5] S. Miah, A. Shahjahan, M. Hossain, and N. Sharma, "A survey of rice diseases in bangladesh," *International Journal of Pest Management*, vol. 31, no. 3, pp. 208–213, 1985.

[6] "Rice disease identification photo link." www.agri971.yolasite.com/resources/RICE/DISEASE/0IDENTIFICATION.pdf. Accessed: 2019-08-25.

[7] "Rice leaf diseases data set." https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases. Accessed: 2019-09-27.

[8] R. Kaur and V. Kaur, "A deterministic approach for disease prediction in plants using deep learning," 2018.

[9] "Caffe." https://caffe.berkeleyvision.org/. Accessed: 2019-08-26.

[10] T. Islam, M. Sah, S. Baral, and R. RoyChoudhury, "A faster technique on rice disease detection using image processing of affected area in agro-field," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 62–66, IEEE, 2018.

[11] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational intelligence and neuroscience*, vol. 2016, 2016.

[12] F. T. Pinki, N. Khatun, and S. M. Islam, "Content based paddy leaf disease recognition and remedy prediction using support vector machine," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2017.

[13] S. R. Maniyath, P. Vinod, M. Niveditha, R. Pooja, N. Shashank, R. Hebbar, *et al.*, "Plant disease detection using machine learning," in *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, pp. 41–45, IEEE, 2018.

[14] H. B. Prajapati, J. P. Shah, and V. K. Dabhi, "Detection and classification of rice plant diseases," *Intelligent Decision Technologies*, vol. 11, no. 3, pp. 357–373, 2017.

[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[16] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[17] S. Gnanambal, M. Thangaraj, V. Meenatchi, and V. Gayathri, "Classification algorithms with attribute selection: an evaluation study using weka," *International Journal of Advanced Networking and Applications*, vol. 9, no. 6, pp. 3640–3644, 2018.

[18] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[19] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[20] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[21] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.