Detection of Paddy Diseases using Deep Learning Methodologies

R. Jeya Bharathi
Research Scholar
Dr. A. P. J. Abdul Kalam University
Indore, MP, India

Dr. Arpana Bharani
Research Supervisor
Dr. A. P. J. Abdul Kalam University
Indore, MP, India

Abstract **-----Now a days, Farmers are facing loss in cropproduction due to many reasons one of the major problem forthe above issue is crop diseases.To ensure healthyand proper growth of the paddy plants it is essential to detectany disease in time and prior to applying required treatment tothe affected plants. Since manual detection of diseases costs alarge amount of time and labour, it is inevitably careful to havean automated system. This paper presents a paddy disease detection system using deep learning approaches. Three of the most common paddy diseases namely leaf smut, bacterialleaf blight and brown spot diseases are detected in this work. Clear images of affected paddyused as the input. After necessary pre-processing, the dataset wastrained on with a range of different learning algorithmsincluding that of K-Nearest Neighbour, Decision Tree,Naive Bayes,Logistic Regression and Convolutional Neural Network.Convolutional neural network algorithm (CNN), achieved an accuracy of over98.80%**

Key Words----- Disease detection, Deep learning, paddy, supervised learning, CNN

I. INTRODUCTION

In India most of its economy comes from Agricultureitself and it is the second largest producer of wheat and Rice.It provides employment to 60% of the Indian population and generates 17% to the total GDP of India.This consequently also contributes towards almost half of the rural employment (49%). While providing a vital role in the country's economy, paddy 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 paddy 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) [1]. These economic gatherings clearly indicate that proper paddy cultivation is a high priority for Indonesia. Disease free paddy cultivation would play a foremost role in ensuring stable economic growth and maintainin the desired targets. Moreover, to keep pace with the emerging fourth industrial revolution, Indonesia needs to work for its industrialadvancements which will involve smart systems that can takedecisions without any human interventions. To that end, wehave come up with an automated system using deeplearning techniques, a system that will contribute in country' sagricultural development by automatically identifying andclassifying diseases from the images of paddy.Paddy blast and brown spot were considered as the most prominent diseases then, but nowbrown spot and bacterial blight are considered as the most prominent and dangerous paddy diseases [2]. In this paper, we have focused on the identification of three paddy disease detection namely bacterial blight, brown spot and leaf smut.

The features of the diseases [3] 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

This paper proposes such an approach that makes disease prediction and classification of the three mentioned paddy diseases. The novelty of the paper lies in the detection of paddy diseases using deep learning approaches with high accuracy.The proposed solution of thispaper has been described in section III .The comparative study among thefive has been
analyzed in section IV for better representationand understanding of the efficiency and accuracy of the modeltrained with different algorithms.

II. LITERATURE REVIEW

Sladojevic and colleagues [4] 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 [5]. For achieving a higher accuracy in evaluating a predictive model, the authors used 10-fold cross validation techniqueon their dataset. The accuracy of prediction of this model is 96.5%.Depending on only the extracted percentage of the RGBvalue of the affected area of rice leaf using image processing, amodel was developed in [6] to classify the disease. The RGBpercentages were fed into Naive Bayes classifier to finallycategorize the diseases into three disease classes: Bacterialleaf blight, Rice blast and Brown spot. The accuracy of themodel to classify the diseases is over 89%.In another study [7], the affected parts were separated fromthe rice leaf surface using K-means clustering and the modelwas then trained with SVM using color, texture and shape asthe classifying features.Maniyath et al. used random forest, an ensemble learningmethod, to classify between healthy and diseased leaf [8]. Forextracting the features of an image, the authors used Histogramof Oriented Gradient (HOG). Their work has claimed anaccuracy 92.33%.Image Processing and machine learning techniques werealso used in [9] for the detection and classification of riceplant diseases. Authors of this paper used K-means clusteringfor the segmentation of the diseased area of the rice leavesand Support Vector Machine (SVM) for classification. Theyachieved a final accuracy, 93.33% and 73.33% on training andtest 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 paddy diseasedetection model using deep learning algorithms that canbe helpful for disease recognition. The data for this task iscollected from the UCI Repository [10].Python [11]an open source, has been used toapply different deep learning algorithms to train our model.

Classifiers used

Supervised classification algorithms were applied on paddy Disease Dataset to detect three diseases of paddy.In this work, four classification algorithms were applied todetect the diseases. At first we applied classification algorithmsbefore attributes selection and achieved different results forfour algorithms. After that we applied classification algorithms
using five selected relevant attributes with applying 10-foldcross validation and achieved better results.

1) Logistic Regression:

Logistic regression can only be appliedif the target class has categorical values. As the aim wasto predict and categorize the disease of the affected rice leaf,logistic regression was a suitable model to train our datasetwith. This paper works on predicting three distinct diseases,so we used multiclass logistic regression. In multiclass logisticregression, for given i classes, idifferent binary classifiers$h_\theta^{(i)}x$ re trained for each class i to determine the probabilityof y, the target class [12]. Then, a new input x can be predictedto belong to the class i if it maximizes max $_\theta{}^{(i)}$ x

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

where,

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

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:

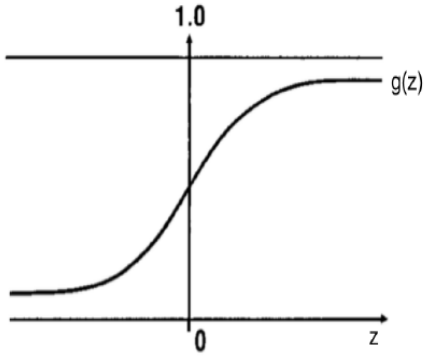

Fig. 2. S-shaped sigmoid function

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:

It calculates the distances of the query point from each of the instances and finds the K minimum distances that is, it determinesthe K nearest neighbours for the query point from whichit can predict the class of the query point. The value ofK needs to be chosen by inspecting the data; in case, wefound when K = 1 the accuracy is 98.8426% on trainingset and 91.6667% on testing set after performing 10-foldcross validation.And when K=3 the accuracy is 85.6481% ontraining set and 72.9167% on testing set after performing 10-fold cross validation. We found, if the value of K is increasedthen accuracy is decreased.

3) Decision Tree:

Decision tree [13] is one of the mostcommonly used machine learning classifiers. Taking the bestsuitable attribute at the root, this algorithm breaks the datasetinto partitions. The goal of the partition is to unmix the dataset.The splitting iterates until eventually the partitions group thedata such that they are homogeneous. Iterative dichotomiser3 (ID3), which uses a greedy approach, is the core algorithmfor decision tree. In this approach, entropy and informationgain, concepts borrowed from information theory, are usedfor constructing the tree. Entropy measures

the impurity ofarbitrary 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.

$$E = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (3)$$

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 mostinformation gain would be selected for this purpose.

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

Here, A is the known attribute and Sv is the subset of Afor which, A has the value v.Using five selected attributes decision tree algorithm wasable to correctly classify 94.9074% data on training set where10-fold cross validation is performed. The model achieved 97.9167% accuracy on test data.

4) Naive Bayes Classifier:
Naive Bayes [14] algorithm is aprobabilistic algorithm that is based on Baye's theorem. Basedon this theorem, the best hypothesis [16] is chosen based onequation 5

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

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

5) Convolutional Neural Network:
CNN[15] consists of an input layer, multiple hidden layer andan output layer. In hidden layer consist of Convolution layer,Rectified Linear Unit, pooling layer and fully connectedlayer.The input layer takes the resized, gray scaledimage and output layer produces the detection of the diseaseand provides remedies. The detailed explanation of the remaining layers as follows,
(i) Convolutional Layer
The training data (images of the diseased and healthy riceplant) was sent to input layer of CNN. The convolutionoperation is then performed on input samples; the input isconvolved with filters called kernels, that is, a number offilters slide over the feature map of the previous layer, toproduce output feature maps.
(ii)Rectified Linear Unit (ReLu)
In this layer is usually called as activation function layer,types of activation function available such as sigmoid, Tanh,ReLU, Softmax, etc. In our model ReLu activation function isused in hidden layers. It is the most widely used activationfunction. In ReLu layer the image with negative pixel valuesare replaced with pixel value 0 and remaining pixel retain asit is. The ReLU function can be written in the mathematicalform in equation 1,0

$$f(x) = \begin{cases} 0 & \text{if } x<0 \\ x & \text{if } x>=0 \end{cases} \quad (1)$$

where x is a pixel value.
(iii) *Pooling Layer*
A pooling layer performs reduction operation along thedimensions of image (Width, Height), resultingdimensionality reduction. The primary aim of poolingoperation is to reduce the size of the images as much aspossible. This scans across the image using a window andcompresses the image extracting features. Average poolingand Max pooling are the most commonly used methods inpooling layers. In max pooling largest value of the pixel istaken from the selected window of the image, while averagepooling takes the average of all pixel values within thewindow.
(iii) Fully Connected Layer
After the convolution + RELU+ Pooling layers, we stack upthese layers many times until the image is reduced to avector. In this layer actual classification will going to happen.In this layer all the neurons are interconnected; this layerproduces an N-dimensional vector, where every neuron inthis layer contains the vectors of the features extracted fromthe image.The proposed system has concentrated on detecting the paddydiseases and provides the suitable remedies, thus leads toincrease in paddy crop production. In this system it detectsthe most common and frequently occurring paddy diseases(Rice blast and bacterial Blight) and provides pesticides orinsecticides as a remedy to control the disease. The type ofpaddy disease is detected by CNN algorithm.

IV. RESULTS AND DISCUSSION

Training and Test dataset contains 432 and 48 instancesrespectively and 5 attributes were chosen. Table I shows theaccuracy of five classification algorithms after performing 10-fold cross validation on training data (90% of dataset)and testdata (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	94.9074 %	97.9167 %
Naive Bayes	58.7963%	50%
CNN	96.81%	98.80%

The comparison between the accuracy of the four classificationalgorithms are represented in Figure 6. Besides accuracy,other performance measures like TPR(True Positive Rate),FPR(False Positive Rate), Precision value (Positive PredictiveValue), Recall value (Sensitivity), F-Measure and AUC(AreaUnder ROC) are also evaluated to compare among the five algorithms and it reveals in Table II and Table III that in eachcase, decision

tree algorithm outperforms all other algorithmsin detecting and classifying the diseases.

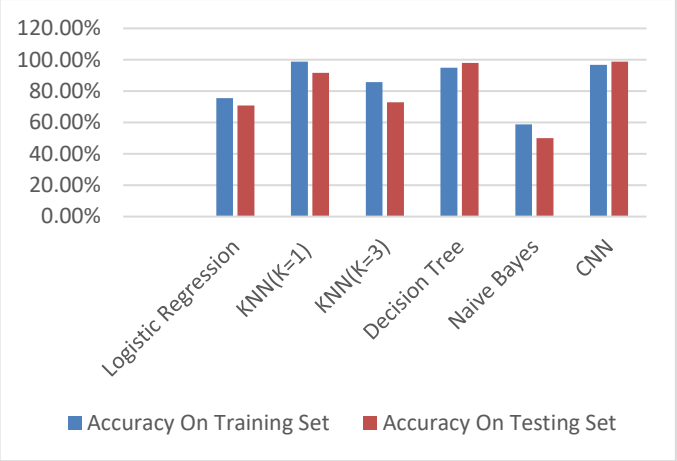

TABLE II DETAILED EVALUATION ON TRAINING SET

Algorithms	TP	FP	Precision	Recall	F-measure	Area under ROC
Logistic Regression	0.755	0.123	0.757	0.755	0.754	0.882
KNN(K=1) KNN(K=3	0.988 0.856	0.006 0.072	0.988 0.858	0.988 0.856	0.988 0.857	0.987 0.978
Decision Tree	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
CNN	0.95	0.027	0.960	0.95	0.96	0.99

TABLE IIIDETAILED EVALUATION ON TEST DATA

Algorithms	TP	FP	Precision	Recall	F-measure	Area under ROC
Logistic Regression	0.708	0.141	0.720	0.708	0.699	0.882
KNN(K=1) KNN(K=3	0.917 0.729	0.042 0.131	0.933 0.755	0.917 0.729	0.915 0.731	0.899 0.931
Decision Tree	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
CNN	0.95	0.008	0.970	0.980	0.98	0.97

The empirical error rate for a classifier is given by equation6. This is then used to determine the accuracy.

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

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 deep learning approach to detectthree different paddy diseases: leaf smut, bacterial leaf blightand brown spot disease. A comparison between five deeplearning algorithms in the realms of paddy disease detection has been made. The algorithms predictedthe paddy diseases with varying degrees of accuracy. It wasfound that CNN performed the best with 98.29% accuracy on test data. Having thus identified a near-optimalalgorithm, we hope to extend this study further as higherquality datasets become available in the future.In Future other Paddy diseases can be trained and Mobile Application can be developed and make it available free on the Google play store. Other diseases can be trained and integrate with the paddy crop model, so that model can be able to predict other diseases.

REFERENCES

[1] "Usda: Rice output continues to see growth." https://www.dhakatribune.com/business/economy/2019/04/09/usda-rice-output-continues-to-seegrowth.Accessed: 2019-08-25.
[2] 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.
[3] "Rice disease identification photo link." www.agri971.yolasite.com/resources/RICE/DISEASE/0IDENTIFICATION.pdf. Accessed: 2019-08-25.
[4] R. Kaur and V. Kaur, "A deterministic approach for disease prediction in plants using deep learning," 2018.
[5] "Caffe." https://caffe.berkeleyvision.org/. Accessed: 2019-08-26.
[6] 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.
[7] 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.E, 2018. [9] "Caffe." https://caffe.berkeleyvision.org/. Accessed: 2019-08-26.
[8] 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.
[9] 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.
[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[11] 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.
[12] 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.
[13] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, no. 1, pp. 81–106, 1986.
[14] 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.
[15] Mrs.Shruti U, Dr.Nagaveni V, Dr.Raghavendra B K "A review on machine learning classification techniques for plant disease detection" 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019.