Identification of Paddy Leaf Diseases using Machine Learning Techniques

Pemasiri S.S.B.P.S.
Department of Computing and Information Systems
Faculty of Applied Sciences
Wayamba University of Sri Lanka

Vidanagama V.G.T.N.
Department of Computing and Information Systems
Faculty of Applied Sciences
Wayamba University of Sri Lanka

ABSTRACT
Many Sri Lankan as well as rice farmers in other countries have trouble identifying diseases in paddy leaves.Only tedious manual methods are used to identify those diseases. But results are not accurate and effective as they would expect due to lack of knowledge. Using a proper computerized approach those diseases can be identified quickly and accurately which can save time and crop yield.

Automatic identification and diagnosis of paddy leaf diseases is a welcome task in the agricultural field. Using a dataset of 800 natural images of diseased and healthy rice plant leaves and stems captured from the rice experimental field, machine learning models and Convolutional Neural Networks (CNN) models are trained to identify diseases in paddy leaves. The logistic regression, decision tree and CNN models were applied to the dataset. Thereafter, CNN techniques were chosen for the experiment. This study proposes a novel paddy leaves diseases identification method based on a deep CNN model. The proposed CNN model achieves the highest training accuracy of 80.25% with the training data set.

Keywords
Paddy agriculture, CNN techniques, machine learning models, logistic regression, decision tree.

1. INTRODUCTION
Rice is a staple food in most countries in the world. Most of the farmers in Sri Lanka also depend on rice cultivation. Here, approximately 10% to 30% of rice production is wasted due to rice diseases. There are numerous rice diseases throughout the world such as Rice Blast, Rice Sheath Blight, Brown Spot, False Smut, Crain, Spotting and peaky rice, Leaf Scald, Marrow Brow Leaf Spot, Sheath Rot, Root Knot,Bacterial Leaf Streak, Brown Spot, Leaf Blast are mainly caused by bacteria, virus or fungi. The damage of these diseases is caused to the plant, which can result in a significant drop in yield due to low photosynthesis. Brown Spot and Leaf Blast are two main diseases that mostly harm rice leaves. [1]

Farmers judge the diseases by their experience but this is not accurate and effective. Sometimes farmers seek the help of agricultural experts for detecting the diseases resulting in more time and cost. However, it is an important task for farmers to find out these diseases immediately in the early stages for minimizing the loss of yield.

As a problem statement how unhealthy paddy leaves can be separated from healthy leaves affected by various paddy diseases. Conventional Neural Network (CNN) methods can easily separate unhealthy paddy leaves from healthy leaves effectively.

Therefore, this study aims to use an optimized neural network to identify paddy leaf diseases targeting the farmers who involve in the paddy sector. It will enhance their yield by solving one of the major problems of identifying unhealthy paddy leaves. By using machine learning models properly, byidentifying those diseases within few seconds it is possible tosave timeand cost.

In this new era with science and technology, many computer-aided diagnosis systems have been developed for agricultural sciences along with plant disease recognition systems. Nowadays, deep learning techniques have attracted the attention of researchers due to their great performance in image classification. The advantage of the deep learning technique is that it avoids the extraction of complex hand-crafted features unlike traditional machine learning techniques and provides end-to-end learning.Among different deep learning techniques, the deep convolutional neural network (CNN) has been used mostly for image classification.

2. LITERATURE REVIEW
When considering the possibility of using Machine Learning Techniques for Rice Plant Disease Detection in Agricultural Research Daniya and Vigneshwari has reviewed it. Convolution Neural Network (CNN) was identified as a top-level-performing methodology for recognition and was potentially used in several applications. Each ML and image processing technology comprises its benefits and specific features. But with the comparison of accuracy, it is concluded that the CNN classifier has a higher accuracy rate. [2]

As well as Ahamad et al. have focused their work to create a rice leaf disease detection model using machine learning algorithms that can be helpful for disease recognition. Image pre-processing and applying methods have to be perfect to obtain a good result. Logistic regression, KNN, decision tree, and naïve bayes algorithms were used. KNN has scored a higher accuracy while naïve bayes scored the lowest accuracy. The decision tree has also performed well in this work. It is needed to perform k -cross-validation to improve the accuracy of the model. Since they have used a small dataset, it is difficult to identify the attributes. Therefore, a large dataset will result in better performances with a clear selection of attributes. [16]

Rice Blast Disease Recognition Using a Deep Convolutional Neural Network is generally consideredas the better methodology in image recognition. A relatively better dataset was used and both quantitative and qualitative analyses were used for evaluating the proposed method. Feature extraction was thoroughly performed because it is required the features to be sufficiently discriminated. Dimension reduction can use for effective pattern recognition and image analysis tasks. The

SVM classifier is capable of converting nonlinear separable problems into linear separable problems. SVM with the suitable kernel function and CNN combination can achieve higher accuracy. So that, SVM is used to compare feature extraction methods and improve the performance of the classification. [5]

A novel rice diseases identification method based on deep convolutional neural networks (CNNs) techniques was implemented by Zeng et al. Without using the CNN algorithm directly, a CNN-based model was introduced. It has performed well with a record of 95.41% accuracy. K- cross-validation method was used to identify ten common diseases. One of the major issues with this task is the size of their dataset. It is a relatively small set of natural images. Also, it is better to work with several different algorithms rather than depending on a certain algorithm so that the suitability of the model can be assured with the comparison. [4]

When considering the build of the CNN model for rice leaf disease detection, Minhaz et al have used a process with four major stages: preparation of dataset, developing the model, deep feature extraction, and finally classification for recognizing rice leaf diseases. The proposed custom CNN-based model is more focused on reducing the network parameters. Images with diverse backgrounds are augmented to improve the generalization of the model. Images were passed in batches to train the model while learning and optimizing the network parameters in convolution, pooling, and dense layers. Binary classification can use to compare the effectiveness of the model for each class of rice leaf disease Even though the proposed method reached for higher accuracy it is suggested that the model is more effective concerning memory storage. [6]

3. METHODOLOGY

Fig. 1 shows the basic steps involved in the implementation process for decision tree and logistic regression. These steps were followed sequentially, solid basic knowledge and understanding of these steps were required for accurate completion of the work.

For this research, data of the paddy leaf images were collected which were taken on white background. The data set contained 800 images where 400 images were healthy and 400 images were diseased. The data set was divided into healthy and diseased data sets and labelled. Again, the data set was divided into training and test sets. Data sets were tested where the tested data set contains 20% of the total data. Divided data sets were uploaded to google drive and used the google co-laboratory for further analysis.

The paddy leaf-related diseases were listed out in advance. The images were read and displayed using google co-labs. Data prepossessing was performed with the features of google co- lab where all images were resized into 255*255 pixels and greyscale. Oncethe data pre-processing is finished, a decision tree and logistic regression were applied to the dataset and compared the accuracy of the models with the test data set.

The basic steps involved in the implementation process for building CNN models is shown in fig2.In CNN model 1, the first layer rescales the images. This model consists of three convolution blocks and three max-pooling 2D layers. Each Conv2D layer in this model is configured as applying the same padding for all images. [2] The CNN model 1 includes one flatten layer. Adam function was used as the model optimizer and the binary cross-entropy function is used as the loss function when fitting the model before training the model. After fitting, the model was trained with 50 epochs.

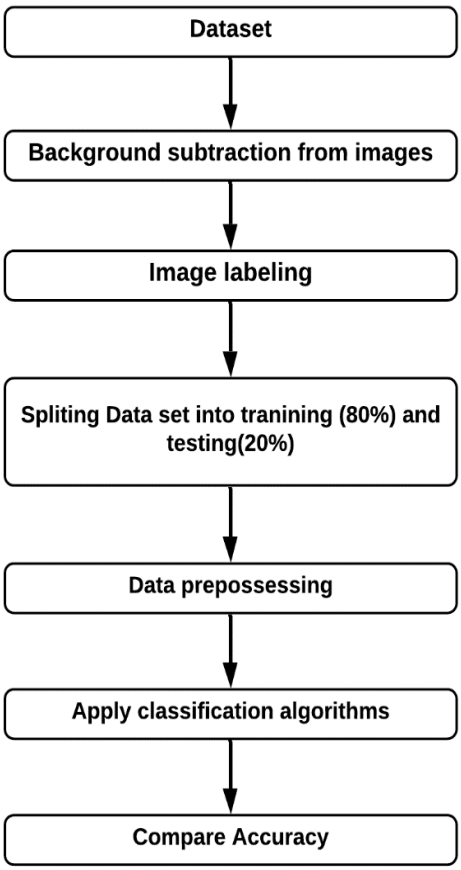

Fig 1: Basic steps involved in the implementation process for Decision Tree and Logistic regression

CNN model 2 is an improvement of CNN model 1 and to achieve CNN model 2, compile method parameters were changed. In that creation, the loss function was changed to Sparse Categorical Cross-entropy (from_logits=True). The batch size was changed to 32 and finally, the input shape of the rescaling layer was removed. This model was also trained with 50 epochs.

Once CNN model 2 is trained, CNN model 3 was created. Fig 6 shows CNN model 3. In this model, CONV 2D layers were added and data augmentation was used additionally. This model was trained with 50 epochs.

The number of images was increased using data augmentation in CNN model 4. Fig 3 shows the CNN model 4. Two drop-out layers were added on both sides of the flatten layer to reduce the model overfitting. The activated function was changed to the sigmoid function for a better outcome. [6]

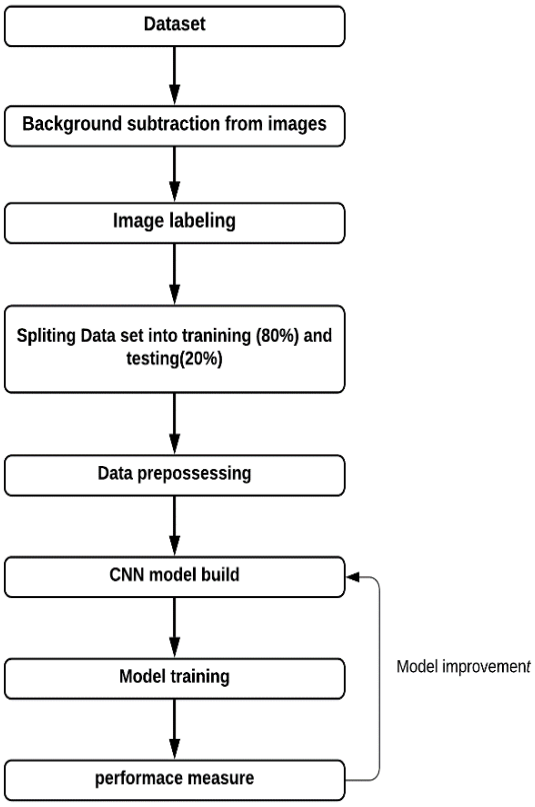

Fig 2: Basic steps involved in the implementation process for build the CNN models

4. RESULT

Even though both decision tree and logistic regression methods were used to obtain results, high accuracy was obtained by the Conventional Neural Network (CNN) model to identify healthy and unhealthy paddy leaves in the field. Results obtained by each algorithm are mentioned in table 1.

Table 1: Validation accuracy of all models

Models	**Validation Accuracy (%)**
Decision Tree	52.14
Logistic Regression	47.85
CNN model 1	47.53
CNN model 2	59.26
CNN model 3	72.89
CNN model 4	80.25

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
sequential (Sequential)      (None, 256, 256, 3)       0
_________________________________________________________________
rescaling_1 (Rescaling)      (None, 256, 256, 3)       0
_________________________________________________________________
conv2d (Conv2D)              (None, 256, 256, 16)      448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 128, 128, 16)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 128, 128, 32)      4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 64, 64, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 64, 64, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 32, 32, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 32, 32, 64)        36928
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 16, 16, 64)        0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 16, 16, 64)        36928
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 8, 8, 64)          0
_________________________________________________________________
dropout (Dropout)            (None, 8, 8, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 4096)              0
_________________________________________________________________
dropout_1 (Dropout)          (None, 4096)              0
_________________________________________________________________
dense (Dense)                (None, 128)               524416
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 258
=================================================================
Total params: 622,114
Trainable params: 622,114
Non-trainable params: 0
_________________________________________________________________
```

Fig 3: CNN model 4

Fig 4: Final model test with disease paddy leaf

99.975% accuracy. In detail, it was mentioned that this is a real diseased paddy leaf.

Fig 5:Final model test with disease paddy leaf

Fig.5 shows the output of the final model after inserting a diseased paddy leaf image. According to the results of the model, it was identified as a real diseased paddy leaf with 99.949% accuracy. In detail, it was ensured that this is a real diseased paddy leaf.

5. DISCUSSION

Decision tree and logistic regression models are predefined machine learning models and therefore, these models cannot be accustomed per someone's wish to get better results. One can only adjust the data pre-processing steps in these algorithms. According to table 1, the decision tree algorithm provides an accuracy of 52.14% whereas the logistic regression algorithm provides an accuracy of 47.85%. The reasons for this outcome are thought to be the use of limited image sets and the difficulties in distinguishing between healthy and diseased leaves.

CNN models are used to train data sets since the decision tree and logistic regression cannot be used due to less accuracy. 2D Convolution Layers (Conv2D layers) help to produce a tensor of outputs. CNN model 1 includes one pattern layer which affects the backside of the network.

In CNN model 2 the batch size was changed to 32 hence, the number of samples that spread through the network increased. Therefore, the input shape of the rescaling layer was removed. The CNN model 2 achieved an accuracy of 69.26%. These changes.

In CNN model 2 the batch size was changed to 32 hence, the number of samples that spread through the network increased. Therefore, the input shape of the rescaling layer was removed. The CNN model 2 achieved an accuracy of 69.26%. These changes resulted in the increment of the accuracy of the CNN 2 model. Additional two Conv2D layers were added to increase the accuracy of the CNN model 3. This model was trained with 50 epochs and it was given 72.89% accuracy with validation data set.

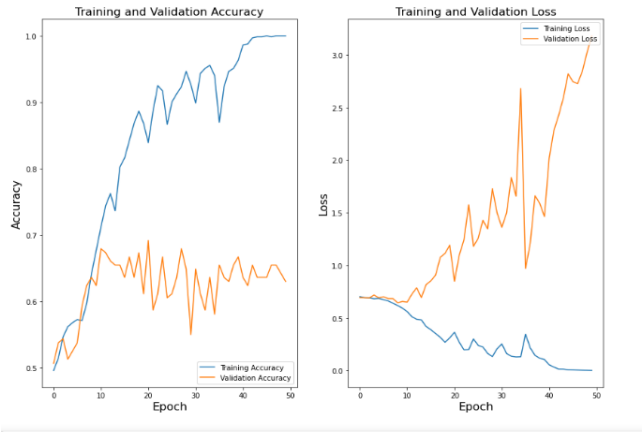

Fig 6: CNN model 3 training and validation accuracy and loss according to epochs

According to fig 6, the training accuracy is almost always higher than validation accuracy. Hence, CNN model 3 is not over-fitted. But the training accuracy and validation accuracy has a significant difference. Even though the training loss is decreased, validation loss is increased over the number of epochs. Validation loss has a very high value. Therefore, this model has to be improved for better accuracy.

Fig 3 represents the CNN model 4. The activation function was changed to the sigmoid function in the last dense layer and data augmentation was used to increase the amount of data. Two drop-out layers were added to reduce the overfitting of the model. This model was trained with 80 epochs and it shows 80.25% accuracy with the validation set.

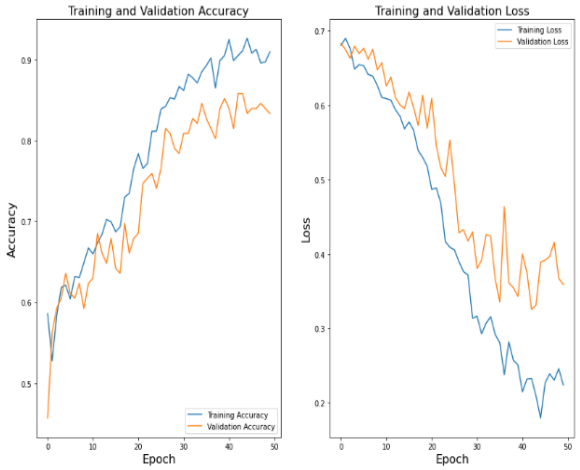

Fig 7: CNN model 4 training and validation accuracy and loss according to epochs

The accuracy of a model can be increased by increasing the epoch size. But training accuracy is increased, the validation loss is also increased over the epochs. That is, CNN model 3 has a 93.74% validation loss with 250 epochs. Therefore, CNN model 3 was rejected.

CNN model 4 gave the highest accuracy than any other models that we created. According to fig 7, the training and validation accuracy is increased similarly and the validation loss kept decreasing over the epochs. Also, the training and validation loss is decreased with the increase of the number of epochs. According to the graph, the validation accuracy is lower than the training accuracy. Therefore, CNN model 4

was not over-fitted. Therefore, this model is chosen as the best model.

Although other researchers enhanced their accuracy in the CNN model using large data sets, here there was obtained 80.25% accuracy using about 800 data set. Therefore, this CNN model is specific than others.

To reduce the loss of the model, an increase in the image count and quality of images is suggested. Although image augmentation can be used to increase the count of images, obtaining many real images is the best solution. Obtaining large dataset is suggested because eventually, it will lead to a model with better accuracy.

6. CONCLUSION

The purpose of this research is to identify the diseased images in the paddy leaves using image classification models. An 800 images data set was used for conducting this research by representing 400 healthy images and 400 unhealthy images. This study has obtained 52.14% accuracy in the decision tree algorithm and 47.85% accuracy in logistic regression.

Convolutional Neural Network (CNN) is used for obtaining higher accuracy since the above models did not have adequate accuracy. As a result, the best CNN model was selected among four CNN models which gave the highest accuracy of 80.25% using about 800 data set. The accuracy can be enhanced by increasing the training data set.

Therefore, it ensures that the CNN model is an effective method for classifying the diseases of paddy leaves. As future enhancements, there is an ability to improve this model tospecifically identifying the disease.

7. ACKNOWLEDGMENTS

It was gratifying that I wrote this to express my deepest gratitude to everyone who encouraged me to complete the research project.

I should not forget to appreciate the immense support given by the all-academic staff at the Department of Computing and Information Systems to successfully complete my research project. Finally, my sincere gratitude goes to all my colleagues, my family members for supporting me throughout the whole process.

8. REFERENCES

[1] Department of Agriculture Sri Lanaka, "Rice Research and Development Institute," 2021. [Online]. Available: https://doa.gov.lk/rrdi/index.php?option=com_sppagebuilder&view=page&id=42&lang=en. [Accessed 20 06 2021].

[2] T. Daniya and D. Vigneshwari, "A Review on Machine Learning Techniques for Rice Plant Disease Detection in Agricultural Research," International Journal of Advanced Science and Technology, vol. 28, pp. 49-62, 2019.

[3] Y. Lu, S. Yi, N. Zeng, Y. Liu and Y. Zhang, "Identification of rice diseases using deep convolutional neural networks," Neurocomputing, vol. 267, pp. 378-384, 2017.

[4] W.-j. Liang, H. Zhang, ,. G.-f. Zhang and H.-x. Cao, "Rice Blast Disease Recognition Using a Deep Convolutional Neural Network," Scientific reports, vol. 9, p. 2869, 2019.

[5] S. Md, M. Hossain, M. M. M. Tanjil1, M. A. B. Ali, M. Z. Islam, M. S. Islam, S. Mobassirin, I. H. Sarker and R. Islam, Rice Leaf Diseases Recognition Using Convolutional Neural Networks, Springer, Cham, 2021.

[6] Aniskheloufi, "Preprocess Image Data For Machine Learning," analytics-vidhya, 28 3 2021. [Online]. Available: https://medium.com/analytics-vidhya/preprocess-image-data-for-machine-learning-37df531583d8. [Accessed 2021].

[7] J. Brownlee, "Train-Test Split for Evaluating Machine Learning Algorithms," machinelearningmastery, 26 08 2020. [Online]. Available: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/. [Accessed 04 08 2021].

[8] K. G. Liakos, P. a Busato, D. Moshou, S. Pearson and D. Bochtis, "Machine Learning in Agriculture," Sensors, pp. 1-24, 2018.

[9] N. S. Chauhan, "Decision Tree Algorithm, Explained," kdnuggets, january 2020. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Accessed 26 04 2020].

[10] B. . T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, pp. 20-28, 2021.

[11] D. Brownlee, "Logistic Regression for Machine Learning," 01 04 2016. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. [Accessed 28 04 2021].

[12] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," International Conference on Engineering and Technology (ICET), pp. 1-6, 2017.

[13] www.geeksforgeeks, "Confusion Matrix in Machine Learning," www.geeksforgeeks, 21 06 2020. [Online]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/. [Accessed 05 08 2021].

[14] R. Ruizendaal, "Deep Learning #3: More on CNNs & Handling Overfitting," Towards Data Science, 12 03 2017. [Online]. Available: https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d. [Accessed 06 08 2021].

[15] R. dadsas, "Disease Identification in paddy leaves using CNN based Deep Learning," Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021.

[16] K. Ahmed, T. R. Shahidi, S. Md, I. Alam and . S. Momen, "Rice Leaf Disease Detection Using Machine Learning Techniques," International Conference on Sustainable Technologies for Industry, vol. 4, pp. 24-25, 2019.