IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

Efficient Analysis and Disease Detection System for Paddy Crop Using Machine Learning and Image Processing Techniques

Siva Parvathi V[1], Pavan Gopi Chand Pidikiti[2], Juber Shaik[3], Nandu Rettapalli[4], Jayadeep Mothukuri[5]

Information Technology, Pvpsit, Vijayawada, India[1-5]

Abstract: In India, paddy is one of the most widely grown crops. These days, this crop is facing challenges from diseases that affect its quality and yield. This study presents an effective machine learning and image processing-based analysis and disease detection system for paddy crops. In this work, totally three different disease classes were taken those were brown spot, leaf smut, Bacterial Leaf Blight, and healthy class are taken. The proposed system uses transfer learning from a pre-trained VGG16 convolutional neural network and fine-tunes the model parameters through hyperparameter tuning via grid search to optimize the SVM classifier. By doing so, an accuracy of 98% on the test dataset is achieved. The system also uses image processing techniques, including color thresholding, morphology operations, and contour detection, to analyse and quantify the affected area of diseased leaves. Moreover, the system provides remediation guidance for each disease, utilizing text-to-speech synthesis for multilingual accessibility.

Keywords: VGG16, SVM, Grid Search, Hyperparameter tuning, OpenCV2, Scikit-learn, Morphology operations, Contours, Image processing, gTTs.

I. INTRODUCTION

In India, paddy, or rice, is the most cultivated crop, and it also stands as a cornerstone of agricultural practices worldwide. This paddy cultivation in India extends from the fertile plains of Punjab and Haryana in the north to the coastal regions of West Bengal and Andhra Pradesh in the east and south, which cover up to 47–48 million hectares of land. In the financial year 1980, India produced 53.6 million of tons and was raised up to 120 million of tons by 2020–21, In 2023 financial year, it is estimated to be 135 million of tons. This steady growth in production happened in 2017.

In 1990–91, agriculture's share of India's GDP was 35%, but it's declined to 15% in 2022–2023 due to various reasons, like growth in industrial and service sectors and a decrease in the quality and yield of the crop because of diseases. The agriculture share of India's GDP may increase if we increase the quality and yield of the crop by ensuring disease-free cultivation. Since paddy is the most cultivated and acts as the cornerstone of agriculture in India, it also plays a major role in rural livelihoods by providing employment opportunities to the farmers. So, all their lives will depend on this crop, so it is important to increase the yield and quality by ensuring disease-free cultivation.

There are several diseases that can affect the paddy crop, but the most common are brown spot, leaf smut, and bacterial leaf blight. Therefore, it is important to predict these diseases in the early stages and start remediation. Each disease has its own characteristics and appearance, which helps to classify the diseases. In this work, the prediction of three main disease classes (Brown Spot, Leaf Smut and Bacterial Leaf Blight) has been focused along with the other class (Healthy) to make the differentiation between healthy leaf and diseased leaf. Not only predicting the diseases, the suitable remediation is also important, and the analysis of the area affected by the disease is also important to know the severity of the disease. So along with prediction, this work also focuses on giving the suitable remediation for each disease and also analysis the disease affected area of the leaf which help in knowing the severity of the disease.

Generally, this identification of diseases is carried out by the people by looking at the affected area and analyzing that leaf by the experts, and after that, suitable treatment and remediation will be given by the experts. It takes a lot of time and much more people to look at the entire farm for any affected leaves. So, to solve this problem, a system is created using machine learning and image processing techniques that will give instant results without any delay, so the required action can be taken very quickly without any further delay. For building this efficient analysis and prediction system, the machine learning model should be trained on the images of the leaves of the four classes, and since each disease has its own characteristics, appearance, and color, we can use color thresholding and morphology operations to quantify the affected area percentage of the diseased leaves.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

India is a diverse country with people of different cultures and traditions. According to the Peoples Linguistic Survey of India, India has 780 number languages. Most of the farming is done in the rural parts of India; they have their own regional language, and some of the farmers don't know how to read the scripts or the text. To overcome this barrier, the work is also focused on including the different languages, so the three languages have been introduced in the system: English, Hindi, and Telugu, and a greater number of languages will be introduced in the future development of the system. Also, text-to-speech synthesis was also developed in the system, so it will be useful for the ones who can't read so that they can listen to the predicted disease and affected area percentage and its suitable remediation in their required language.

II. LITERATURE REVIEW

The paper titled "Paddy Crop Disease Detection Using Machine Learning" by PrajwalGowda B.S et al. developed algorithm using machine learning and deep learning techniques like convolutional neural network to classify the two diseases bacterial leaf blight and rice blast [1].

"Detection and classification of Paddy crop Disease Using Deep Learning Techniques" proposed by Kiruthika et al. achieved accuracy of 93.33% uses the artificial neural networks and GLCM (Gray-Level-Co-occurrence Matrix) which helped in getting high accuracy and accurate disease detection.[2].

"Efficient Disease Detection of Paddy Crop using CNN" proposed by P.A. Harsha Vardhini, S. Asritha, Y. Susmitha Devi in this paper uses CNN to detect the disease in uploaded image by comparing with the images in the database they used and IoT component raspberry pi in their work [3].

The cutting-edge technologies deep learning, especially convolutional neural network are used to develop the algorithm in their work [4]. The comparison of the performances of different machine learning and deep learning algorithms using evaluation metrics accuracy and f1-score for the diseases brown spot, sheath blight, bacterial leaf blight, leaf blast, Sheath Rot, Leaf Smut [5].

The comparison between five pre-trained CNN architectures is made and ensemble those models to detect blast, bacterial leaf blight and brown spot and achieved an accuracy of 95.54 % is for the final ensemble algorithm. The algorithms used in this work are Wide ResNet, Shuffle Net, ResNeXt, ShuffleNet, GoogleNet[6]. Image Processing is used to classify the diseases leaf smut, brown spot and bacterial leaf blight using Otsu's method and utilizing the "Local Binary Patterns (LBP)" and "Histogram of Oriented Gradients (HOG)" to get the features and then features are classified using polynomial Kernel SVM and HOG achieved an accuracy of 94.6 % [7].

"Rice Leaf Disease Prediction using Machine Learning" authored by Bhartiya et al. In this work the features have been extracted from the diseased leaves and applied various machine learning techniques to classify the images and got an accuracy of 81.8 % when using Quadratic SVM Classifier [8]. This paper presents a machine learning based framework for detection of diseases. Naïve Bayes, SVM and are used for machine learning frameworks and for preprocessing and feature extraction histogram equalization and PCA algorithm is used respectively [9]. A three pre-trained CNN models which EfficientNet V2S, DenseNet201 and InceptionV3 are used by applying transfer learning to detect the five disease classes and the highest accuracy is achieved by the DenseNet201 which is 92.05% [10].

III. PROPOSED WORK

The proposed system takes the images of three different disease classes and one healthy class of paddy leaf which are (Brown Spot, Leaf Smut, Bacterial Leaf Blight, and Healthy) and preprocesses those images to do feature extraction using a pre-trained VGG16 convolution neural network with the weights of the ImageNet dataset. It then fine-tunes the model parameters through hyperparameter tuning via grid search to optimize the SVM classifier to increase the accuracy of disease detection. Image processing techniques include color thresholding, morphology operations, and contour detection to analyse and quantify the affected area percentage of the leaf, give a suitable recommendation for each disease, and use text-to-speech synthesis to help regional farmers.

A. Dataset

The dataset for the three disease classes has been taken from Kaggle [11] and the fourth healthy class has been taken from [12] and some of the images were collected randomly from the internet for each disease. A sample images or dataset example is shown in Fig.1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

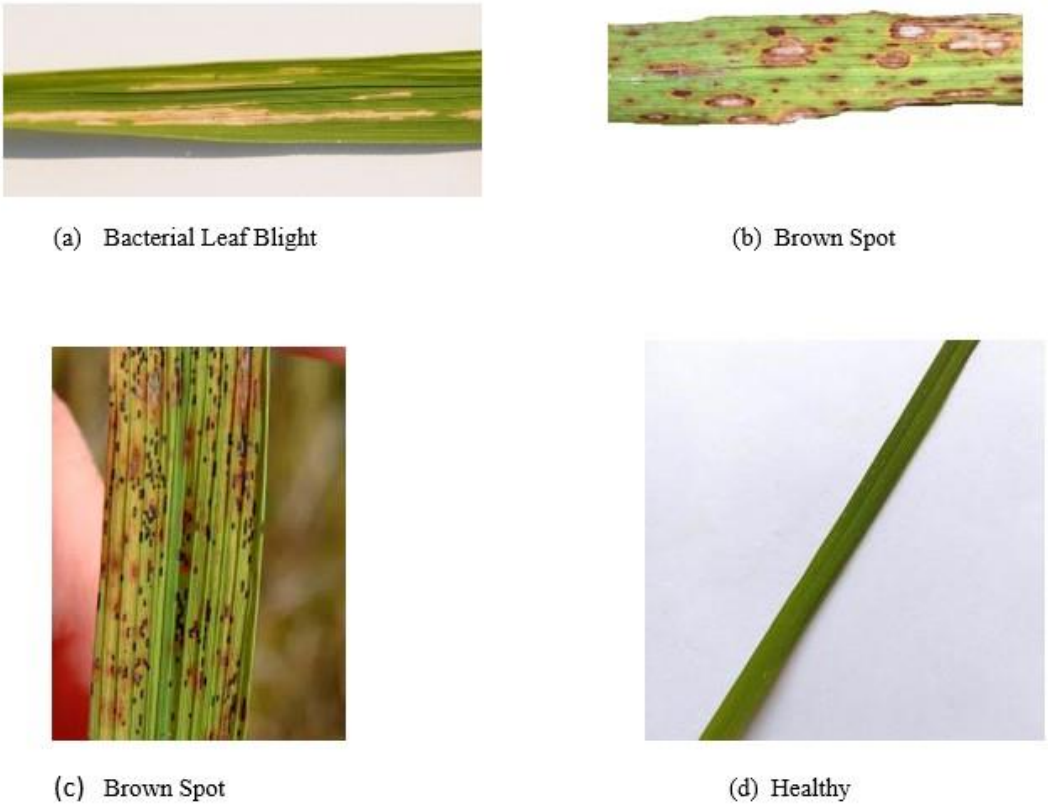

Fig. 1 Dataset Sample

B. Data Pre-processing

The entire dataset is undergoing preprocessing to ensure that the images should fit in the machine learning model. For this preprocessing, the OpenCV2 library is used to read the images, resize the images into 224x224 pixels, and normalize the pixel values to the range [0, 1] to ensure consistency in the data so that the model can train more effectively, which also helps in improving the convergence and performance of the machine learning model. After preprocessing every image in the dataset, The dataset needs to be split into two parts, which are the training dataset and the testing dataset.

C. Data Splitting

The dataset splits into two parts: training data and testing data. Thus, 20% of the images are separated into testing data, which are used to assess the model's performance, and the remaining 80% of the images are divided into training data, which are used to train the model.

D. VGG16

A pre-trained VGG16 convolutional neural network is used to extract the features from the images. This pre-trained model is imported from the TensorFlow library with the weights of the ImageNet dataset without including the top layers of the model [13]. This VGG16 comprises 16 convolutional layers, followed by fully connected layers, and topped with a SoftMax layer for classification. The summary of the loaded VGG16 model, which is used for feature extraction, is shown in Fig. 2.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

```
Model: "vgg16"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 224, 224, 3)]     0

 block1_conv1 (Conv2D)       (None, 224, 224, 64)      1792

 block1_conv2 (Conv2D)       (None, 224, 224, 64)      36928

 block1_pool (MaxPooling2D)  (None, 112, 112, 64)      0

 block2_conv1 (Conv2D)       (None, 112, 112, 128)     73856

 block2_conv2 (Conv2D)       (None, 112, 112, 128)     147584

 block2_pool (MaxPooling2D)  (None, 56, 56, 128)       0

 block3_conv1 (Conv2D)       (None, 56, 56, 256)       295168

 block3_conv2 (Conv2D)       (None, 56, 56, 256)       590080

 block3_conv3 (Conv2D)       (None, 56, 56, 256)       590080

 block3_pool (MaxPooling2D)  (None, 28, 28, 256)       0

 block4_conv1 (Conv2D)       (None, 28, 28, 512)       1180160

 block4_conv2 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_conv3 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 14, 14, 512)       0

 block5_conv1 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv2 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv3 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_pool (MaxPooling2D)  (None, 7, 7, 512)         0

=================================================================
Total params: 14714688 (56.13 MB)
Trainable params: 14714688 (56.13 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 2 Overview of architecture and configuration of used VGG16

E. Optimizing SVM Parameters using Grid Search

A support vector machine (SVM) classifier is used for image classification. Here, we are using the grid search technique for hyperparameter tuning the SVM, and the best-performing model is selected based on the cross-validated results obtained from the grid search [14] after getting the best SVM model training that model with the extracted training features.

F. Image Processing for quantifying affected area

The process of applying several procedures to an image is known as image processing. It is used to make the image better or to extract important information from it [15], so in this image processing technique, analyzing and calculating the affected area of a disease is done by using certain techniques, such as color thresholding, morphology operations, and contour detection. Color thresholding is applied in the HSV (Hue Saturation Value) color space. The lower and upper color threshold values will depend upon the characteristics of the disease, and this color threshold changes the pixels within the range into white and the remaining pixels into black [16]. After that, morphology operations are done on the obtained binary mask from the color thresholding to remove all the noise or small gaps in the detected region [17], and a later contour detection technique is used so that it finds the contours in the binary mask and draws those contours in the original image, and based upon these contours, the total affected area is calculated [18].

IJARCCE A. ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

G. Text-to-Speech Synthesis

We used the gTTS (Google Text-to-Speech) Python library for generating the audio to help the farmers who can't read and implemented the system in three languages: English, Hindi, and Telugu, so that they could use their required regional language to get the remediation details and information about the affected area's percentage of leaf affected by the disease.

The entire process of the system is shown in the flowchart shown in Fig. 3.

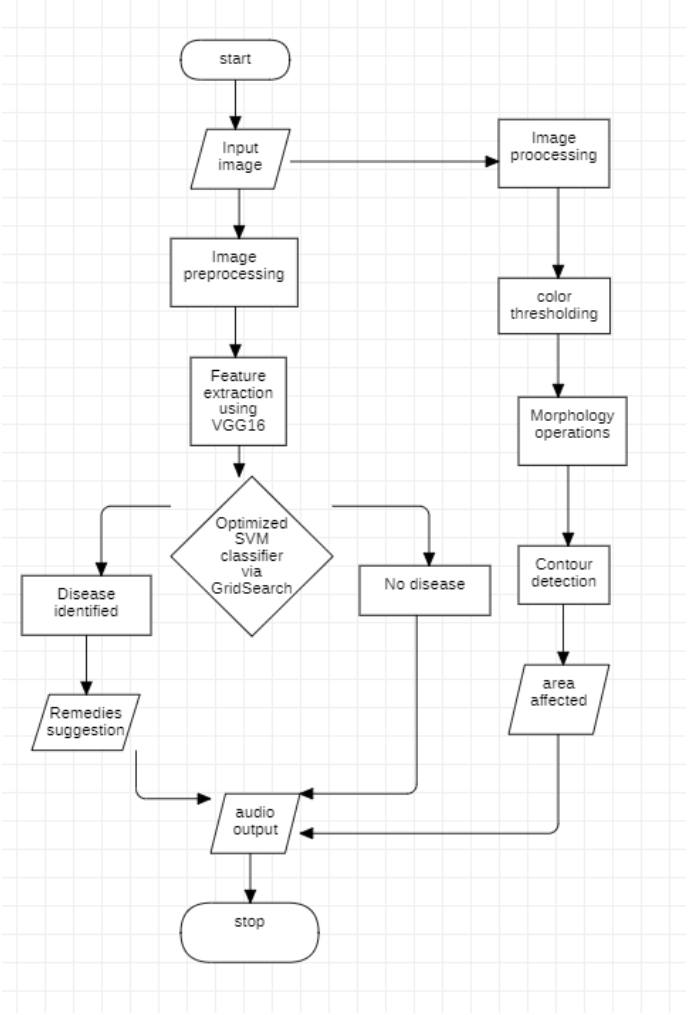

Fig. 3 Flow chart of the proposed system

IV. RESULTS AND DISCUSSION

A. Model Performance and Evaluation

The accuracy score, confusion matrix, and classification report are among the assessment metrics used to assess the performance of the machine learning model. The scikit-learn library's sklearn. metric module is the source of all of these assessment metrics. The detailed analysis of the model performance based on these evaluation metrics is elaborated in the below subsections.

1) *Accuracy Score:* The model's accuracy score is calculated on the test data by using the 'accuracy_score' function from the scikit-learn library. The accuracy score calculates the percentage of accurately predicted instances in the test data based on the total number of instances.

The optimized SVM model via GridSearch achieved an accuracy of 97.6190% indicating that the model is performing well in classifying the images.

IJARCCE A. ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

2) *Confusion Matrix:* This is most cruical tool evaluating the machine learning model performance by comparing the acutal labels of the test dataset with predicted labels.

By examining the values in the confusion matrix helps in calculating or in deriving the various aspects of the model performance those are accuracy, recall, precision, f1-score the values of true positives, true negatives, false positives and false negatives are useful in deriving these metrics from the confusion matrix. The confusion matrix is displayed in Fig. 4 when the generated model is applied to the test data.

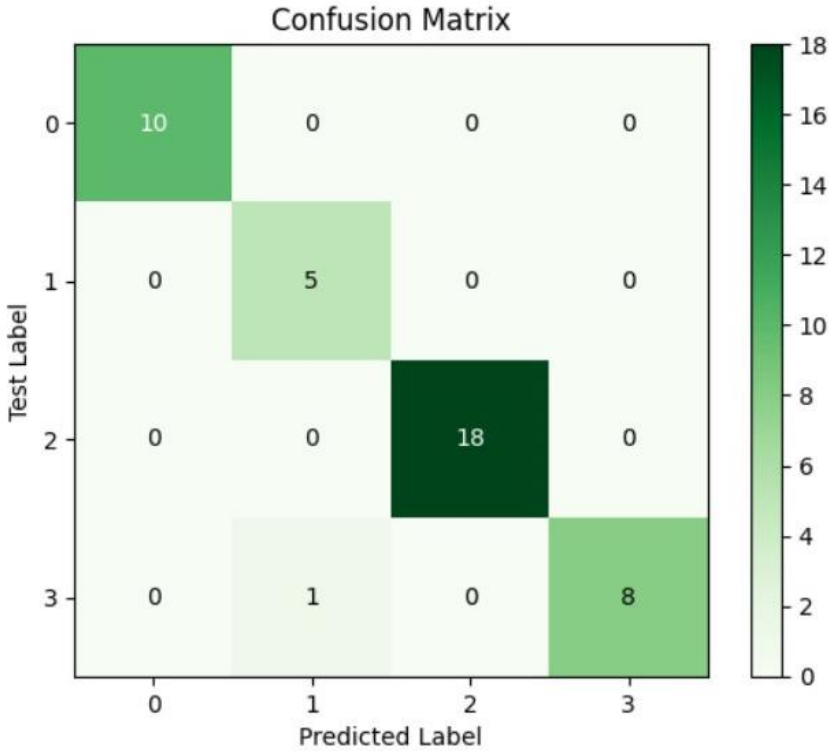

Fig. 4 confusion matrix

3) *Classification report:* The many model performance evaluation parameters for each class are summarized in this classification report, along with accuracy, weighted average, macro average, and precision, recall, F1-score, and support measures. Precision represents the quality of positive predictions, calculated as the ratio of correctly predicted positive observations to the total predicted positive observations.

Recall, also known as sensitivity, measures the proportion of true positive predictions relative to the total number of actual positive instances in the dataset. The F1-score represents the balanced average between precision and recall. The presence of every class in the dataset is known as support. Accuracy can be defined as the proportion of correctly classified cases relative to the total number of examples. The metrics averaged by the number of examples in each class is known as the weighted average. Macro average is the unweighted mean of metrics calculated for each class

The scikit-learn library's "classification_report" function is used to create the classification report, and Fig. 5 displays the classification report for the developed model.

IJARCCE A. ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

	precision	recall	F1-score	support
Bacterial Leaf Blight	1.00	1.00	1.00	10
Brown Spot	0.83	1.00	0.91	5
Healthy	1.00	1.00	1.00	18
Leaf Smut	1.00	0.89	0.94	9
accuracy			0.98	42
macro avg	0.96	0.97	0.96	42
weighted avg	0.98	0.98	0.98	42

Fig. 5 Classification report

V. CONCLUSION

Based upon the classification report, the developed optimized SVM model demonstrates good precision, recall, and f1-scores across all classes, indicating its effectiveness in classifying diseases. The overall accuracy achieved by the model is 98%. So, this paper presents a system that uses machine learning approaches for detecting the three disease classes: brown spot, leaf smut, and bacterial leaf blight, along with the healthy class, and image processing approaches for analysing and quantifying the affected area of the leaf. It also uses text-to-speech synthesis, which helps farmers who can't read, and the system can be accessible in three different languages: English, Hindi, and Telugu, to overcome language barriers.

In the future development of this work, the more diseases are added for classification, the dataset size needs to be increased so that the model can train with a variety of features, and the number of languages the system can access needs to be increased so that it can be used by every regional area in the country. We need to provide access for users to create an account on the website and develop a system that can track all user activities, like which month the user uploaded a diseased leaf, which is useful for alerting the user before the crop gets affected by the disease.

REFERENCES

[1]. PrajwalGowda, B. S., Nisarga, M. A., Rachana, M., Shashank, S., & Raj, B. S. (2020). Paddy crop disease detection using machine learning. *International Journal of Engineering Research & Technology*, *8*(13), 192-195.

[2]. Kiruthika, S. U., Raja, S. K. S., Jaichandran, R., & Priyadharshini, C. (2019). Detection and classification of paddy crop disease using deep learning techniques. *International Journal of Recent Technology and Engineering*, *8*(3), 4353-4359.

[3]. Vardhini, P. H., Asritha, S., & Devi, Y. S. (2020, October). Efficient disease detection of paddy crop using CNN. In *2020 International conference on smart technologies in computing, electrical and electronics (ICSTCEE)* (pp. 116-119). IEEE.

[4]. Nithjapoopathy, S., Pratheesh, B., Sansayan, N., Abishaalini, S., Rupasinghe, L., & Liyanapathirana, C. (2023, December). Image Classification of Disease Detection in Paddy Crops Using Transfer Learning Approach. In *2023 5th International Conference on Advancements in Computing (ICAC)* (pp. 143-148). IEEE.

[5]. Vasantha, S. V., Kiranmai, B., & Krishna, S. R. (2021). Techniques for rice leaf disease detection using machine learning algorithms. *Int. J. Eng. Res. Technol*, *9*(8), 162-166.

[6]. Acharya, A., Muvvala, A., Gawali, S., Dhopavkar, R., Kadam, R., & Harsola, A. (2020, November). Plant Disease detection for paddy crop using Ensemble of CNNs. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-6). IEEE.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13390

[7]. Pothen, M. E., & Pai, M. L. (2020, March). Detection of rice leaf diseases using image processing. In 2020 fourth international conference on computing methodologies and communication (ICCMC) (pp. 424-430). IEEE.

[8]. Bhartiya, V. P., Janghel, R. R., & Rathore, Y. K. (2022, March). Rice leaf disease prediction using machine learning. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)* (pp. 1-5). IEEE.

[9]. Pallathadka, H., Ravipati, P., Sajja, G. S., Phasinam, K., Kassanuk, T., Sanchez, D. T., & Prabhu, P. (2022). Application of machine learning techniques in rice leaf disease detection. *Materials Today: Proceedings*, *51*, 2277-2280.

[10]. Hosain, A. S., Mehedi, M. H. K., Jerin, T. J., Hossain, M. M., Raja, S. H., Ferdoushi, H., ... & Rasel, A. A. (2022, September). Rice leaf disease detection with transfer learning approach. In *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 1-6). IEEE.

[11]. "Rice Leaf Diseases Dataset." *Kaggle*, 21 Feb. 2020, www.kaggle.com/datasets/vbookshelf/rice-leaf-diseases.

[12]. "Rice_Leaf_Images." *Kaggle*, 16 Nov. 2021, www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images.

[13]. "Tf.keras.applications.vgg16.VGG16 : Tensorflow V2.15.0.POST1." *TensorFlow*, www.tensorflow.org/api_docs/python/tf/keras/applications/vgg16/VGG16. Accessed 21 Mar. 2024.

[14]. C, Bala Priya. "Hyperparameter Tuning: GridSearchCV and RandomizedSearchCV, Explained - KDnuggets." *KDnuggets*, www.kdnuggets.com/hyperparameter-tuning-gridsearchcv-and-randomizedsearchcv-explained.

[15]. Simplilearn. "What Is Image Processing : Overview, Applications, Benefits, and More." *Simplilearn.com*, 11 Oct. 2023, www.simplilearn.com/image-processing-article#:~:text=Image%20processing%20involves%20performing%20operations,important%20details%20from%20the%20image.

[16]. "ImageMagick." *ImageMagick*, imagemagick.org/script/color-thresholding.php#:~:text=Use%20color%20thresholding%20to%20specify,with%20a%20hyphen%20between%20them.

[17]. Rosebrock, Adrian. "OpenCV Morphological Operations - PyImageSearch." *PyImageSearch*, 9 May 2021, pyimagesearch.com/2021/04/28/opencv-morphological-operations.

[18]. Rath, Sovit, and Sovit Rath. "Contour Detection Using OpenCV (Python/C++)." *LearnOpenCV – Learn OpenCV, PyTorch, Keras, Tensorflow With Code, & Tutorials*, 5 Feb. 2024, learnopencv.com/contour-detection-using-opencv-python-c/#:~:text=Using%20contour%20detection%2C%20we%20can,image%20segmentation%2C%20detection%20and%20recognition.