ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

Rice Leaf Disease Detection and Crop Yield Prediction Using Random Forest

[1] Shilpa Ajeesh M, [2] Anagha C, [3] Fathima Sifa, [4] Hilma T, [5] Deepthi P, [6] Jijina M T

[1] [2] [3] [4] Student, KMCT College of Engineering for Women, Kozhikode, Kerala, India.
[5] [6] Assistant Professor, KMCT College of Engineering for Women, Kozhikode, Kerala, India.
Corresponding Author Email: riceplantleafdiseasedetection@gmail.com

Abstract*— Rice production contributes a considerable amount to national income. Rice production can be affected by various diseases like brown spots, bacterial leaf blight, leaf smut caused by fungi, bacteria, etc. In this research, the diagnosis of rice plant leaf disease is done using random-forest classification and Digital image processing. The random forest classifier is efficient and accurate on a large dataset. The image is uploaded to the system by following digital image processing steps and using a random forest algorithm to perform on the processed image which outputs disease name, cause, symptoms, and remedy respectively. The proposed method also predicts the crop yield based on temperature, rainfall, humidity, and soil pH level. Overall, the model achieves 90% of accuracy.

***Index Terms*—Machine learning, Image processing, Random forest classification, Crop yield**

I. INTRODUCTION

Rice is the deliberate source of food in India so the detection of rice leaf disease plays an important role in the Indian economy. The current methods in use are not much effective. Methodologies with high costs are not affordable to common people.

Image processing is the technique used to process digital images using machine learning algorithms with the help of a digital computer. Digital image processing is the process of manipulating of digital images. Its focus is on a computer system that can process an image. It uses computer algorithms. Digital image processing is used to obtain some useful information from the digital image. Image processing includes steps like image acquisition, image enhancement, image restoration, color image processing, wavelets and multiresolution processing, compression, morphological processing, segmentation, representation and description, and object recognition.

This research intended to extract features from a given image. Digital image processing takes the digital image as input and generates a useful information as output.

Crop production is affected by the input parameters of soil and climate. Parameters vary from area to area and user to user. It may also be a tiring challenge to collect such information in an even larger area. The large information set is used to predict crop production. There are different methodologies developed and evaluated by researchers.

Crop yield prediction can be used to determine the upcoming production yield so that the farmers can take necessary actions to increase the yield. Some climatic and soil parameters are used for the prediction.

The early detection of rice leaf disease helps to reduce the impact on production. The earlier we detect the disease, the earlier we can take precautions by considering the current year parameters of soil and climate as a reference to overlook the characteristics of the coming year's yield.

Machine learning (ML) is a type of artificial intelligence (AI) it will help to become more accurate the software applications. The input of machine learning algorithms is historical data. From that, we can predict the new outputs. Machine learning is very important because it gives more accurate predictions and it's a competitive differentiator for many companies like Facebook, Google, and Uber.

Different types of machine learning algorithms are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

In supervised learning machine learning, data scientists supply the training data labeled with algorithms and define the variables they want the algorithm to assess for correlations. Input and the output of the algorithm are already known.

Unsupervised learning is a machine learning algorithm that trains on unlabeled data. The main goal of unsupervised learning is to Underlying the structure of the dataset and group this data based on similarities and represent it in a compressed format.

Semi-supervised learning is the mix of the two types. It was introduced to counter the disadvantages of both supervised and unsupervised machine learning. The semi-supervised learning algorithm is trained upon a combination of labeled and unlabelled data. It will contain large unlabelled data and a small amount of labeled data.

Reinforcement learning is one of the machine learning algorithms is used in areas such as: Robotics: They can perform tasks in the physical world. Example: A rabbit is an agent exposed to the environment. Fig.1 shows the model of reinforcement learning. Video gameplay: it is used to play several video games.

Random Forest could be a well-known machine learning algorithm. Machine learning algorithms are supervised and unsupervised learning and the random forest algorithm is belonging to the supervised learning technique. It can be used for both Regression and Classification problems in machine

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

learning. It's supported the concept of ensemble learning, which may be a process of mixing multiple classifiers to unravel a fancy problem and boost the performance of the model.

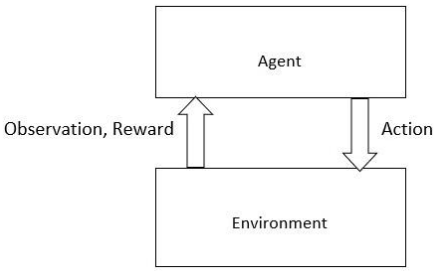

Fig. 1. Reinforcement learning

Random Forest may be a classifier that contains a variety of decision trees on various subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset. Decision trees are the backbone of a random forest algorithm. A choice tree may be a decision support technique that forms a tree-like structure. The random forest employs the bagging method to come up with the desired prediction. It can handle large datasets efficiently. The random forest algorithm provides the next level of accuracy in predicting outcomes over the choice tree algorithm.

The data can be classified in 2 ways: Supervised and Unsupervised learning. The data mining process can be used to identify the class of data. A random forest algorithm is used in both rice leaf disease detection and crop yield prediction. The random forest algorithm is very effective because it can perform both classification and regression at the same time with a very large data set.

The purpose of this paper is the detection of rice leaf disease and the prediction of crop yield by using a random forest algorithm. This paper is arranged as; Section II is the literature review. Section III describes the proposed method for detection and prediction. Section IV shows the experimental arrangements for the implementation of the system. Section V concludes and summarizes the paper while providing suggestions for future work.

The rest of this paper has the following sections, Section II consists of the associated works. Section III introduces the proposed method. The experimental outcomes and critiques are presented in Section IV. Finally, phase V summarizes the paper.

II. LITERATURE REVIEW

In this paper, a random forest algorithm is used to detect rice leaf disease and crop yield prediction. Image segmentation is performed on the test image and features are extracted. A random forest classification algorithm is applied to the segmented image and the disease class is obtained. The crop yield prediction mechanism takes the necessary input data such as weather parameters and performs a random forest algorithm and predicts the crop yield for the upcoming production.

In [1] the classification algorithm is used to detect papaya disease. A Naive Bayes classifier (NBC) is a machine learning algorithm that is used to differentiate objects based on particular features. This paper aims to detect papaya disease using Fuzzy Naive Bayes Classifier (FNBC). The resource person's knowledge is given as input to the system so that the farmers can detect disease without the help of an expert. The resource knowledge is then processed using FNBC to detect the disease class. is used for classification because it is easy and fast in computation, good enough to handle data in qualitative and quantitative forms, and does not require large training data.

In [2] it shows a detailed knowledge of how the machine learning algorithm and digital image processing are implemented in corn plant disease detection. For image, the classification image is processed using Synthetic Minority Oversampling Technique (SMOTE) and using k-means cluster method. The main function of digital image processing in this paper is to upload an image then enlarge the inputted image and from that form its binarization and then image balancing using SMOTE technique. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After getting classified images Gradient Boosting Decision Tree (GBDT) is processed.

In [3] apple fruit infections are detected by using image processing technology. A random forest classifier algorithm is used to extract color and texture features of test images for disease classification. Texture features like Local Binary Pattern, Complete Local Binary Pattern, Gabor Features, and color features like Global Color Histogram, and Color Coherence Vector. The K-means clustering technique is also used in this paper for detecting infected parts of the fruit. The proposed method is made by accompanying some state of the color and texture features are extracted from the test image, then color and texture features are combined together and random forest classifier is used for disease classification. If the fruit is infected by any disease then the infected part is segmented using the k-means clustering technique.

In [4] random forest algorithm is used for the prediction of crop yield. Predictions are done before the cultivation. Climatic and soil parameters are the inputs used for the prediction. This predicts the production of upcoming years. The result obtained from the prediction is useful for making decisions for the upcoming production.

III. PROPOSED METHOD

The proposed system is a methodology for rice leaf disease detection and crop yield prediction using a random forest algorithm. The random forest algorithm is used because of its accuracy. This is a simple, user-friendly system and it consumes less time. Admin, user, and system are the three modules that make up this system. The system is accessible to both admins and users. The administrator has access to the list of registered users, as well as the ability to add disease information and see who has been rated. Users

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

can create accounts and upload photographs to the system. Both administrators and users will be handled by the system.

Input: Affected image or Crop yield parameters.

Output: Disease identification or yield prediction.

Algorithm:

Steps:

1. log in to the system
2. Upload the image otherwise go to step 7
3. Load training images
4. Read the testing image
5. Perform the random forest algorithm for disease detection
6. Identify the disease, suggest remedies, list out symptomsand causes and go to step 10
7. Enter parameters for crop yield prediction
8. Perform the random forest algorithm for crop yield
9. prediction
10. Predict the yield and go to step 10 10. End

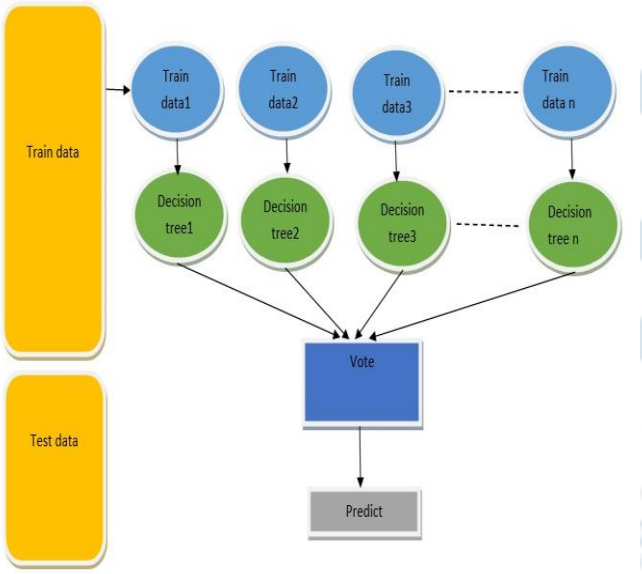

Fig. 2. Working of Random forest algorithm

Rice leaf disease detection and crop yield prediction system using random forest algorithm. The programming language used is python. Here random forest works differently in both scenarios. Digital image processing has a keen role in disease detection[5]-[7]. Fig. 2. shows working of random forest algorithm. Random forest uses a train-test split. Here, 80% is training and 20% testing. The 20% testing is used for prediction.

Fig 3. shows the Architecture of the proposed method for both disease detection and crop yield prediction. For disease detection, the user can upload the training image that is the affected rice leaf into the system, and feature level extraction will happen. Finally using the random forest algorithm to detect the affected disease. In crop yield prediction users can enter the details such as Temperature, Humidity, Rainfall, and Soil pH. The data are extracted from these details and crop yield prediction will happen by using the Random Forest algorithm.

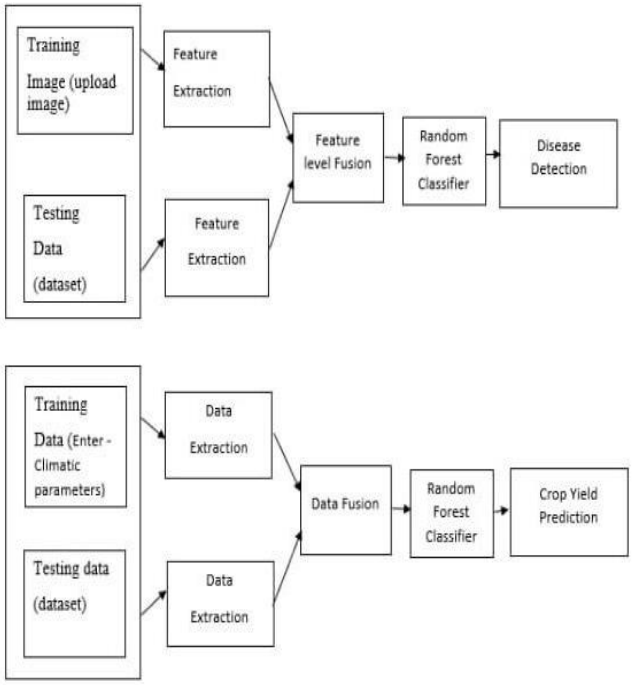

Fig. 3. Architecture of the proposed method

A. Disease Detection

The initial step is to collect images and create a data set that contains the images of various rice leaf diseases. A .csv file is formulated from the created data set. Some of the commonly seen rice leaf diseases are Brown spot, bacterial blight, leaf smut,tungro, blast, etc. The other dataset is a collection of nondisease data. Every time a prediction is made, it is compared to the dataset that has been provided. The more massive the dataset, the more accurate the result.

In the next step, the user can upload an image to the system. The features are extracted using Grey Level Cooccurrence Matrix(GLCM) and calculate the mean value for each feature. Then the mean value is converted into an int array. The attribute and label values for the uploaded images are taken, then inputted into the random forest model and the result is obtained. Along with the result, the system provides symptoms, causes, and remedies for the disease.

Each disease is considered as a different class and each class has unique characters and features according to which the image is classified. The disease class is predicted based on the features of the input rice image. The features we considered are energy, homogeneity, dissimilarity, correlation, and contrast. Features are extracted using Gray Level Co-occurrence Matrix (GLCM). The GLCM functions characterize textures of an image by calculating how often pairs of the pixel with specific values. The disease classes are obtained as output with its symptoms, causes and remedy to solve them.

1) Energy: The Sum of squared elements in the GLCM is also called uniformity or the angular second moment.

2) Homogeneity: It shows the uniformity of the pixels in the image. Also known as inverse difference moment. It

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

measures the density of distribution of elements in the cooccurrence matrix and in its diagonal.

3) Dissimilarity: Dissimilarity is the measure of distance between pairs of pixels in the region of interest. It is the lack of similarity or likeliness in appearance to something else.

4) Correlation: It is the measure of linear dependencies in images. A high correlation means a high number of linear structures present on an image.

5) Contrast: It is also known as inertia, it measures the difference in brightness between the object and its background, and the local variations of gray levels in the co-occurrence matrix. A high contrast causes large variation in the image.

B. Bacterial leaf blight

Bacterial Leaf Blight is a rice leaf disease caused by the bacteria, Oryzae sativa. Fig.4 shows the leaf affected by Bacterial Leaf Blight. Normally the leaf color range is light green to greyish green, when it gets affected by Oryzae sativa it's color changes to yellowish lesions with uneven edges. Then the leaves become yellow they gradually wilt and die. In severe cases, the crop loss may be as high as 75%, and millions of hectares of rice are infected annually.

C. Leaf blast

Blast is a rice leaf disease caused by the fungus Magnaporthe oryzae. Fig.5 shows the image of the leaf blast affected leaf. It affects parts of a rice plant: leaf, collar, node, and sometimes leaf sheath. Blast can occur whenever blast spores are present. Lesions can be found on all parts of the plant, including leaves.

D. Tungro

Tungro is a rice leaf disease caused by the bacteria Rice Tungro Bacilliform Virus (RTBV) and Rice Tungro Spherical Virus (RTSV). Fig .6 shows that tungro diseases are transmitted by leafhoppers. Leaves become yellow or orange-yellow. Controlling the Tungro virus is by spraying insecticides. Spraying of insecticides can reduce the populations of the green leafhoppers which means it will reduce the speed of spreading the virus.

E. Leaf smut

Leaf smut is caused by a fungus and is most prevalent in blue grasses, but can also occur in some cool-season grass varieties. Fig. 7 Leaf smut is caused by the fungus Ustilaginoidea virens. The symptoms are discoloration of grains, grains Transformed into a mass of Yellow fruiting bodies, smut balls bursting and becoming black, and greenish-black smut balls with a velvety appearance.

F. Brown spot

Brown spot is a fungal disease caused by Cochliobolus miyabeanus that infects the coleoptile, leaves, leaf sheath, panicle branches, glumes, and spikelets. The most noticeable damage is the numerous big spots on the leaves which can kill the whole leaf shown in Fig. 8. When infection occurs in the seed, discolored seeds are formed.

Fig. 4. Bacterial blight.

Fig. 5. Leaf blast.

Fig. 6. Tungro.

Fig. 7. Leaf smut.

Fig. 8. Brown spot.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

G. Crop Yield Prediction

The initial step is to collect climatic parameters and soil parameters to create a data set that contains various parameters for crop production[8]. Data pre-processing technique is adapted to convert the raw data to a good data set. A .csv file is formulated from the created data set. Some of the commonly used parameters are temperature, soil pH, rainfall, and humidity. Table.1 shows the normal ranges for these parameters.

Table 1. Normalparameterrange

Sl.no	Temperature	Humidity	Rainfall	pH
1	25 - 35°C	65 – 80 g.kg -1	17 – 30 cm	6.5 – 7.5

The user inputs the parameters and random forest model to compare these values with the values that are present in the data-set and generate a corresponding result such as high, low, or medium production for the upcoming year.

IV. EXPERIMENTAL SETUP

The system contains three data sets. The first data set contains almost 10,000 images of 5 different rice leaf diseases. The second data set contains uninfected rice leaf images and other images. The third data set contains parameters for crop yield prediction.

Depending upon the resulting output, the users can able to take necessary action for the coming production so that it is a user-friendly system.

	A	B	C	D	E	F	G	H	I
1572	1570	0.126848	0.011722	14.47323	0.902203	552.8968	Bacterialblight		
1573	1571	0.125079	0.011263	19.45895	0.81613	1126.783	Bacterialblight		
1574	1572	0.160162	0.017427	10.48046	0.840915	315.0896	Bacterialblight		
1575	1573	0.117117	0.012173	16.89018	0.781713	818.9477	Bacterialblight		
1576	1574	0.172573	0.018885	11.70894	0.90587	478.7589	Bacterialblight		
1577	1575	0.092909	0.010011	23.08118	0.718301	1364.277	Bacterialblight		
1578	1576	0.208427	0.019308	11.41108	0.88462	538.5847	Bacterialblight		
1579	1577	0.209564	0.020154	13.40985	0.886898	624.8167	Bacterialblight		
1580	1578	0.238293	0.021823	7.414086	0.96495	194.5378	Bacterialblight		
1581	1579	0.207824	0.017313	10.91862	0.919531	421.5451	Bacterialblight		
1582	1580	0.225647	0.019868	8.4719	0.947463	238.6517	Bacterialblight		
1583	1581	0.189778	0.016286	11.72509	0.907529	457.6619	Bacterialblight		
1584	1582	0.099905	0.011911	24.09338	0.619111	1360.168	Bacterialblight		
1585	1583	0.270986	0.025859	7.598764	0.934722	234.356	Bacterialblight		
1586	1584	0.494922	0.04777	1.562659	0.993901	5.843303	Blast		
1587	1585	0.262963	0.023675	7.041707	0.954199	180.6962	Blast		
1588	1586	0.30597	0.037784	3.165217	0.974737	20.67635	Blast		
1589	1587	0.287163	0.024255	6.419678	0.945017	151.4337	Blast		
1590	1588	0.314882	0.039065	3.083533	0.972015	20.18161	Blast		
1591	1589	0.312165	0.032289	4.324652	0.972091	57.64951	Blast		
1592	1590	0.546231	0.066027	3.063572	0.884365	115.132	Blast		
1593	1591	0.223929	0.01937	7.804877	0.944171	184.1073	Blast		
1594	1592	0.408213	0.049154	2.179621	0.990432	10.95424	Blast		
1595	1593	0.319099	0.027696	4.85147	0.965933	83.07152	Blast		

Fig. 9. Disease detection dataset

	A	B	C	D	E
1	Temperature	Humidity	pH	Rainfall	Label
2	20.87974371	82.00274423	6.502985292	202.9355362	high
3	21.77046169	80.31964408	7.038096361	226.6555374	high
4	23.00445915	82.3207629	7.840207144	263.9642476	high
5	26.49109635	80.15836264	6.980400905	242.8640342	very high
6	20.13017482	81.60487287	7.628472891	262.7173405	high
7	23.05804872	83.37011772	7.073453503	251.0549998	very high
8	22.70883798	82.63941394	5.70080568	271.3248604	medium
9	20.27774362	82.89408619	5.718627178	241.9741949	low
10	24.51588066	83.5352163	6.685346424	230.4462359	very high
11	23.22397386	83.03322691	6.336253525	221.2091958	very high
12	26.52723513	81.41753846	5.386167788	264.6148697	medium
13	23.97898217	81.45061596	7.50283396	250.0832336	very high
14	23.97898217	80.88684822	5.108681786	284.4364567	medium
15	24.01497622	82.05687182	6.98435366	185.2773389	very high
16	25.66585205	80.66385045	6.94801983	209.5869708	very high
17	24.28209415	80.30025587	7.042299069	231.0863347	very high
18	21.58711777	82.7883708	6.249050656	276.6552459	high
19	23.79391957	80.41817957	6.970859754	206.2611855	high
20	21.8652524	80.1923008	5.953933276	224.5550169	low
21	23.57943626	83.58760316	5.85393208	291.2986618	medium
22	21.32504158	80.47476396	6.442475375	185.4974732	very high
23	25.15745531	83.11713476	5.070175667	231.3843163	medium
24	21.94766735	80.97384195	6.012632591	213.3560921	medium
25	21.0525355	82.67839517	6.254028451	233.1075816	medium
26	23.48381344	81.33265073	7.375482851	224.0581164	very high
27	25.0756354	80.52389148	7.778915154	257.0038865	very high
28	26.35927159	84.04403589	6.286500176	271.3586137	very high
29	24.52922681	80.54498576	7.070959995	260.2634026	very high

Fig. 10. Crop yield prediction dataset

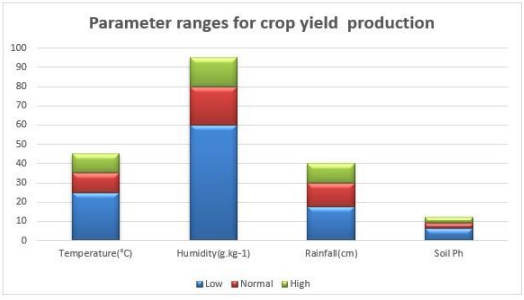

Fig. 11. parameter ranges for crop yield production

Fig. 9. shows the dataset for rice leaf disease detection. Dataset consists of mean values of 5 features in each image in the dataset and corresponding 1 label. The features are energy, homogeneity, dissimilarity, correlation, and contrast. The labels can be bacterial leaf blight, bacterial blight, leaf blast, tungro, and leaf smut. Fig. 10. shows the dataset for rice crop yield prediction. Dataset consists of 4 parameter values and a corresponding 1 label. The parameters are temperature, humidity, rainfall, and pH. The labels are low, high, and medium. Fig. 11. shows the graph representing parameter ranges for crop yield prediction. The graph portrays low, high, and medium ranges for each parameter.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

V. CONCLUSION

This paper presents the algorithm that can be used to classify rice leaf diseases and predict the crop yield using the Random forest algorithm. Image classification using the random forest algorithm is highly accurate when compared with other machine learning algorithms like Naive Bayes, gradient boosting, etc. This provides both regression and classification at the same time so that its prediction is more than 85% accurate and also it predicts the crop yield based on the climate input parameters. This is highly efficient and also less time-consuming.

This proposed system provides the disease name, its symptoms, cause, and remedies following the inputted image and also provides the prediction of crop yield for the coming year. Crop yield prediction is based on some climatic parameters like humidity, soil pH, rainfall, and temperature The built website is user-friendly. The net page is developed to predict rice leaf disease and crop yield by providing data from that area.

VI. ACKNOWLEDGMENT

Our research is supported by the Department of Computer Science and Engineering, Faculty of Department of Computer Science and Engineering, KMCT College of Engineering for Women.

REFERENCES

[1]. Wahyuni Eka Sari, Yulia Ery Kurniawati, Paulus Insap Santosa, “Papaya Disease Detection Using Fuzzy Na¨ıve Bayes Classifier”, 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, doi: 10.1109/ISRITI51436.2020.9315497.

[2]. Tong Xiao, Heyun Liu, Yun Cheng, ”Corn Disease Identification Based on improved GBDT Method”,2019 6th International Conference on Information Science and Control Engineering (ICISCE), 215-219, 2019,doi.org/10.1109/ICISCE48695.2019.00051.

[3]. Bhavini J. Samajpati,Sheshang D. Degadwala, ”Hybrid approach for apple fruit diseases detection and classification using random forest classifier”,2016 International Conference on Communication and Signal Processing (ICCSP), IEEE, doi: 10.1109/ICCSP.2016.7754302.

[4]. Namgiri Suresh, N.V.K.Ramesh, Syed Inthiyaz, “Crop Yield Prediction Using Random Forest Algorithm”. 2021 7th International Conference on Advanced Computing Communication Systems (ICACCS).doi: 10.1109/ICACCS51430.2021.9441871.

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

[6]. 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.

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

[8]. S.Veenadhari, Dr Bharat Misra, Dr CD Singh.2019.”Machine learning approach for forecasting crop yield based on climatic parameters.”. International Conference on Computer Communication and Informatics (ICCCI), 2014.