ISSN: 2583-5637 (Online)
International Journal of Inventive Research in Science andTechnology Volume (2), Issue (4), April 2023

Crop Recommendation and Crop Disease Classification.

K. Srikanth[1], T. SivaTeja[2], N. Khader Basha[3], M. RaviTeja[4], Dr. C. Bala Subramanian[5]

[1234]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhnagar, Tamilnadu, India.

[5]Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhnagar, Tamilnadu, India.

Abstract-Agriculture is the one of the ancestor occupation in India. Nowadays many farmers are getting loss in agriculture because they are not opting the crop which will be suitable for the soil and to the weather conditions of that particular area. And also one of the other major problem is using more pesticides and insecticides without knowing actual disease. These type of problems can be solved using latest technologies like machine learning for crop recommendation and deep learning for crop disease classification. And dataset from Kaggle which contains attributes like chemical properties of soil and weather conditions for crop recommendation. A real time image dataset for training our deep learning using CNN model for crop disease classification. Deep learning is a subset of AI which will be mostly used to work on image and video and voice data.

Key Words: Machine Learning, Deep learning, CNN, Decision tree, Random Forest, VGG16, SVM.

1. INTRODUCTION

India is the country where all different types of soil available for farming. But still farmers getting loss because they don't know which crop is suitable for that soil based on the properties of soil and weather. So we are developing a model that will predict which crop is suitable for that soil using ensemble learning. And also we used Different machine learning algorithms to find which model will be more efficient for crop recommendation. The dataset contains attributes like chemical properties of soil [sodium, potassium, Nitrogen] and weather conditions like [ph, rainfall, humidity etc]. This paper is not only about crop recommendation but it is useful for machine learning researchers which algorithm is more suitable for this type problem. The algorithms are Decision tree classifier, logistic Regression, Random forest classifier, etc. By finding the accuracy of each algorithm after training and testing we can conclude which algorithm is more suitable.

After crop recommendation other major problem that farmers are facing is using pesticides and insecticides without knowing the proper disease of crop. So that the crop may not produce good yield because more fertilizers or chemicals. For this we can use one of the latest technology Deep Learning for crop disease classification. Because if we are able to find the exact disease that crop has infected we can easily find the medicine for that. So by using CNN and VGG16 algorithms we can classify the crop diseases. For these we are using Infected crop image dataset to train our model.

2. Literature Survey

In Paper[1] they have taken two types of datasets one is Symptom based text dataset and Image based dataset. They developed model based these two datasets using Deep learning algorithms for wheat crop. Some of the algorithms they have used is CNN and in CNN they used ResNet50, VGG16, AlexNet models. They got accuracy of 97.2%. They have taken only one particular crop and developed the model so that it can classify whatever effected by to that particular crop and this whole process done under the supervision some farmers.

In paper[2] they recommended precision agriculture. They have taken real-time dataset from farmers using soil test for crop recommendation. And for that dataset they have used many machine learning algorithms like SVM, ensemble learning, Random forest, naïve bayes. And also they have formed some set of rules for recommendation. They concluded that by using these algorithm they tried to improve crop yield.

In paper[3] they have used combination of machine learning techniques like Random forest, KNN, Naïve bayes, CHAID as ensemble learning and along with that they have used a separate technique which is majority voting technique. That means what majority people are going to cultivate they recommend the same crop for others. For ensemble learning they got accuracy of 88%.

In paper[4] they have tried to develop an automated system which classify the disease in an early stage. For these they have taken wheat crop dataset. Based on disease classification they have done clustering. For that automated system they have used image processing and machine learning. To prepare that automated system using image processing and machine learning they proposed systematic diagram and for that they prepared the paper and published.

3. Methodology

3.1 Dataset Collection

The dataset used for crop recommendation is having attributes like chemical properties and weather conditions. The labels present in our dataset is rice, maize, chickpea, kidneybeans, pigeonpeas, mothbeas, mungbean, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple ,orange, papaya, coconut, cotton, jute, coffee. The attributes in dataset is Nitrogen, phosphorus, potassium, temperature, humidity, PH value, rainfall. The dataset contains 2201 data. According to chemical properties present in the soil the fertility of soil changes and have different crop yielding capacity. So the farmers should know which crop will suitable for their land according to above chemical property.

The dataset used for crop disease classification is having the images of cotton crop with various types of infected images. At first we have to develop deep learning model for training the image dataset. The dataset is having 6 different classes[Aphids, Army worm, Bacterial Bright, Healthy, Powdery Mildew, Target Spot]. By using these model we can justify which disease does the crop was infected.

3.2 Data Augmentation

Data Augmentation is used to increase the size of the image dataset for better training by using techniques like rotating the image, zoom in, zoom out etc. We use data augmentation for crop disease classification dataset. By using these technique we can get more number of images for training and testing in less time with less manpower.

3.3 Data Pre-processing

Data pre-processing is used to clean the data, remove the unwanted data, selecting the required features etc. For crop recommendation we have a theoretical dataset but we didn't have any noises to change and the dataset contains the required features as attributes. So for crop recommendation dataset data pre-processing is not required because it was already a pre-processed dataset.

Data pre-processing is not only used for theoretical data but also for image dataset it can be used. Since we are using image dataset for crop disease classification, but all images are not in same size. We are scaling down the images to get into same size so that we can train the model effectively.

3.4 Algorithms used for crop recommendation

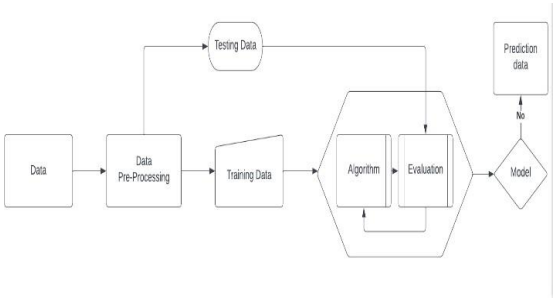

Figure1. Architecture of machine learning

3.4.1 Logistic Regression

Logistic Regression is a supervised machine learning algorithm. It is mostly used to find the categorical dependent variable. Logistic regression is used to solve classification based problems. It is one of the best algorithm for predicting categorical target variable. We have created a logistic regression model and train the model with our dataset. And the before predicting we have given numbering to the label from [1:10] i.e, we are having 10 different types of categorical labels but predicting can be done if it is numerical data only. So we have given numbering to the categorical data and according to that we have trained the data. After testing we got 95.6% accuracy. The classification report for logistic regression is

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	34
blackgram	1.00	0.83	0.91	36
chickpea	1.00	1.00	1.00	35
coconut	1.00	1.00	1.00	31
coffee	0.97	1.00	0.99	33
cotton	0.85	0.97	0.90	29
grapes	1.00	1.00	1.00	30
jute	0.75	0.81	0.78	26
kidneybeans	0.94	1.00	0.97	32
lentil	0.91	1.00	0.95	31
maize	0.97	0.86	0.91	36
mango	0.96	1.00	0.98	26
mothbeans	0.85	0.96	0.90	23
mungbean	0.97	0.97	0.97	32
muskmelon	1.00	1.00	1.00	33
orange	1.00	1.00	1.00	31
papaya	0.96	0.96	0.96	24
pigeonpeas	1.00	0.93	0.96	28
pomegranate	1.00	1.00	1.00	25
rice	0.88	0.75	0.81	28
watermelon	1.00	1.00	1.00	28
accuracy			0.96	660
macro avg	0.95	0.96	0.95	660
weighted avg	0.96	0.96	0.96	660

Figure2. Classification report for logistic regression

3.4.2 Decision Tree

Decision tree is a supervised machine learning algorithm. It can be used for both classification and regression problems. Decision tree produce a tree-structure as an output. In that tree representation the internal nodes represents features of dataset, branches represent decision rules and leaf node represent outcome of the problem. It works based on the decision suitable for solving the problem at each stage. The tree build by the model can be easily understandable so mostly we use Decision tree algorithm. Decision tree will build based on CART algorithm. We created a decision model and trained the model using our dataset and got accuracy upto 98.5%. The tree graph of our model is

ISSN: 2583-5637 (Online)
International Journal of Inventive Research in Science andTechnology Volume (2), Issue (4), April 2023

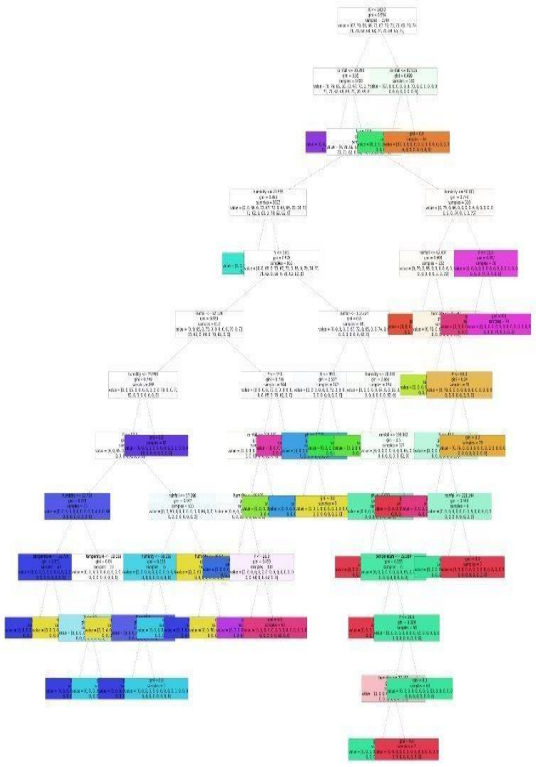

Figure3. Decision tree classification

3.4.3 Random Forest classifier

Random Forest algorithm is a supervised machine learning algorithm. It is also used for both classification and regression problems. It works as ensemble learning to produce high accuracy and efficiency. Random forest classifier will create various decision trees for each subset of data and takes average of that to increase accuracy. Random forest classifier will take less time when compared to other algorithms irrespective of size of dataset. At first it select x number of data points and build some d number of decision trees and based on the average of output of decision tree it predict the output. We created a model of random forest classifier and trained using our dataset. Among all other algorithms we got high accuracy for this model. The accuracy of this model is 99.09%. The classification report of Random forest classifier

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	34
blackgram	1.00	1.00	1.00	36
chickpea	1.00	1.00	1.00	35
coconut	1.00	1.00	1.00	31
coffee	1.00	1.00	1.00	33
cotton	1.00	1.00	1.00	29
grapes	1.00	1.00	1.00	30
jute	0.81	1.00	0.90	26
kidneybeans	1.00	1.00	1.00	32
lentil	1.00	1.00	1.00	31
maize	1.00	1.00	1.00	36
mango	1.00	1.00	1.00	26
mothbeans	1.00	1.00	1.00	23
mungbean	1.00	1.00	1.00	32
muskmelon	1.00	1.00	1.00	33
orange	1.00	1.00	1.00	31
papaya	1.00	1.00	1.00	24
pigeonpeas	1.00	1.00	1.00	28
pomegranate	1.00	1.00	1.00	25
rice	1.00	0.79	0.88	28
watermelon	1.00	1.00	1.00	28
accuracy			0.99	660
macro avg	0.99	0.99	0.99	660
weighted avg	0.99	0.99	0.99	660

Figure 4. Classification report for Random forest Classifier

3.4.4 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm used for classification problems. SVM creates boundaries to segregate n-dimensional space into classes. The boundary is called hyperplane. SVM can also work for image dataset. We created SVM model and trained using our dataset. The accuracy of SVM for the dataset is 97.42%. The classification report for SVM is

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	34
blackgram	1.00	1.00	1.00	36
chickpea	1.00	1.00	1.00	35
coconut	1.00	1.00	1.00	31
coffee	1.00	1.00	1.00	33
cotton	0.88	1.00	0.94	29
grapes	1.00	1.00	1.00	30
jute	0.72	1.00	0.84	26
kidneybeans	0.94	1.00	0.97	32
lentil	0.97	1.00	0.98	31
maize	1.00	0.89	0.94	36
mango	1.00	1.00	1.00	26
mothbeans	1.00	0.96	0.98	23
mungbean	1.00	1.00	1.00	32
muskmelon	1.00	1.00	1.00	33
orange	1.00	1.00	1.00	31
papaya	1.00	1.00	1.00	24
pigeonpeas	1.00	0.93	0.96	28
pomegranate	1.00	1.00	1.00	25
rice	1.00	0.64	0.78	28
watermelon	1.00	1.00	1.00	28
accuracy			0.97	660
macro avg	0.98	0.97	0.97	660
weighted avg	0.98	0.97	0.97	660

Figure 5. Classification report for SVM

3.5 Algorithms Used for Crop Disease classification

3.5.1 Convolution Neural Network

Deep learning will have neural networks which performs as human. CNN is one of the main model in deep learning. CNN model is mainly used for image classification, object detection etc. CNN has Convolution layers where all inputs pass to these convolution layers filters, Fully connected layers, pooling etc and gives the probability between 0 and 1. The general architecture of CNN is

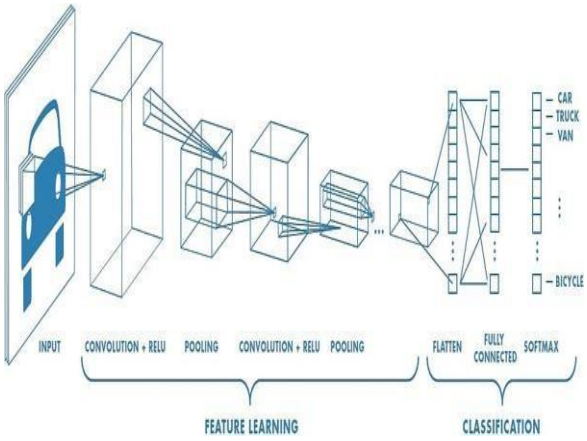

Figure 6. Architecture of CNN

At first CNN will extract the features from the dataset using convolution layer. In CNN we use different activation functions like (RELU, Sigmoid etc) to find the final output. And in CNN we have Striding, padding, Pooling etc. For CNN we have used sequential model for training the dataset using different layers like conv2D, Dense, Flatten, Maxpooling2D. We trained the model by taking 20 epochs. The accuracy of the model is 98.73%. It was trained by taking 6 different classes and now it can able to predict the disease that the cotton crop was infected by. The summary of the sequential model is

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 50, 50, 32)        896

max_pooling2d (MaxPooling2D  (None, 25, 25, 32)        0
)

conv2d_1 (Conv2D)            (None, 25, 25, 32)        9248

max_pooling2d_1 (MaxPooling  (None, 12, 12, 32)        0
2D)

conv2d_2 (Conv2D)            (None, 12, 12, 128)       36992

max_pooling2d_2 (MaxPooling  (None, 6, 6, 128)         0
2D)

flatten (Flatten)            (None, 4608)              0

dense (Dense)                (None, 256)               1179904

dense_1 (Dense)              (None, 64)                16448

dense_2 (Dense)              (None, 6)                 390

=================================================================
Total params: 1,243,878
Trainable params: 1,243,878
Non-trainable params: 0
_________________________________________________________________
```

Figure7. Summary of sequential model in CNN

3.5.2 VGG16

VGG16 is one of the type of Convolution Neural Network. It has 16 deep layers of CNN. So that the predicting capacity of these model will be high because it uses all its 16 layers for feature extraction and classification. And also this algorithm is mainly used for big image datasets for higher predicting capability. The architecture of VGG16 is

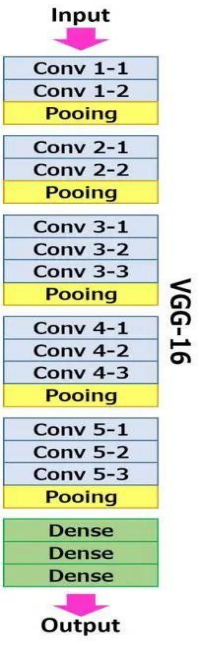

Figure8. Architecture of VGG16

VGG16 is having 13 convolution layers and 5 pooling layers and 3 dense layer of total number of layers is 21. Here we are using VGG16 to get high accuracy and train the model efficiently. For these disease classification dataset it has to find the type of disease that the crop was infected by. For that in VGG16 we have taken weights as ImageNet and model same as CNN sequential. Now The summary of the model is

Model: "vgg16"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 50, 50, 3)]	0
block1_conv1 (Conv2D)	(None, 50, 50, 64)	1792
block1_conv2 (Conv2D)	(None, 50, 50, 64)	36928
block1_pool (MaxPooling2D)	(None, 25, 25, 64)	0
block2_conv1 (Conv2D)	(None, 25, 25, 128)	73856
block2_conv2 (Conv2D)	(None, 25, 25, 128)	147584
block2_pool (MaxPooling2D)	(None, 12, 12, 128)	0
block3_conv1 (Conv2D)	(None, 12, 12, 256)	295168
block3_conv2 (Conv2D)	(None, 12, 12, 256)	590080
block3_conv3 (Conv2D)	(None, 12, 12, 256)	590080
block3_pool (MaxPooling2D)	(None, 6, 6, 256)	0
block4_conv1 (Conv2D)	(None, 6, 6, 512)	1180160
block4_conv2 (Conv2D)	(None, 6, 6, 512)	2359808
block4_conv3 (Conv2D)	(None, 6, 6, 512)	2359808
block4_pool (MaxPooling2D)	(None, 3, 3, 512)	0
block5_conv1 (Conv2D)	(None, 3, 3, 512)	2359808
block5_conv2 (Conv2D)	(None, 3, 3, 512)	2359808
block5_conv3 (Conv2D)	(None, 3, 3, 512)	2359808
block5_pool (MaxPooling2D)	(None, 1, 1, 512)	0

Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0

Figure9. Summary of VGG16 model

And also we have added flatten and dense layer to the model and again checked the summary.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 1, 1, 512)	14714688
flatten_1 (Flatten)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131328
dense_4 (Dense)	(None, 64)	16448
dense_5 (Dense)	(None, 6)	390

Total params: 14,862,854
Trainable params: 14,862,854
Non-trainable params: 0

Figure10. Summary of model in VGG16 after adding flatten and dense layer.

For VGG16 we got accuracy of 99.57% after testing the data.

3.6 Visualization

In crop recommendation system we have different visualization technique to visualize the data and understand the data efficiently. Using seaborn we have visualized the data of each and every attribute along with remaining attributes gives us:

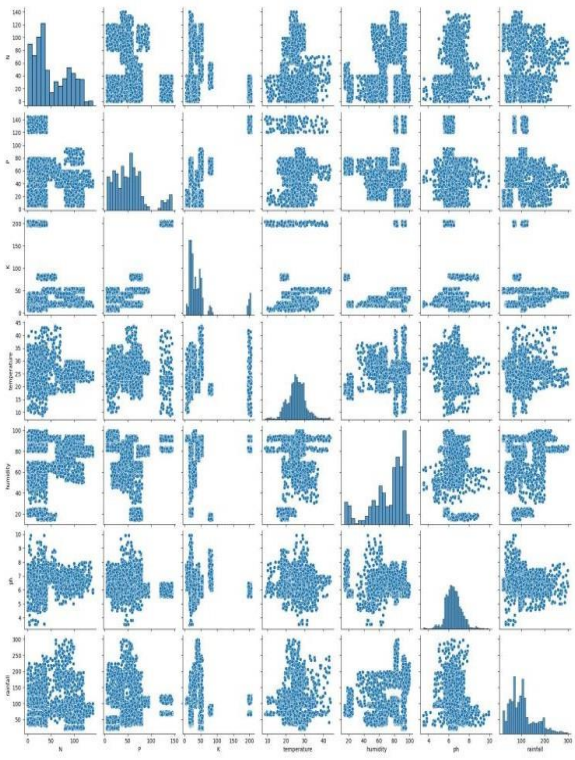

Figure11. Visualization of attributes with respect to other attributes.

ISSN: 2583-5637 (Online)
International Journal of Inventive Research in Science andTechnology Volume (2), Issue (4), April 2023

In Crop disease classification we used matplotlib to plot the graph of disease images.

Figure12. Different class of infected crop images

We plotted graph for training and validation accuracy and training and validation loss in CNN and as well as VGG 16.

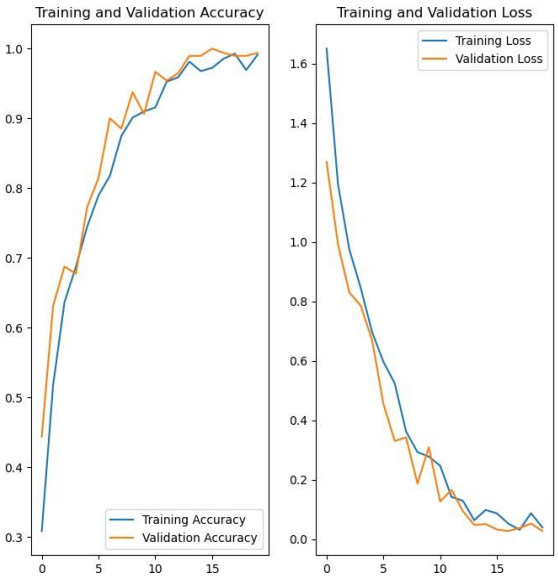

Figure12. Training and validation accuracy and training and validation loss in CNN model

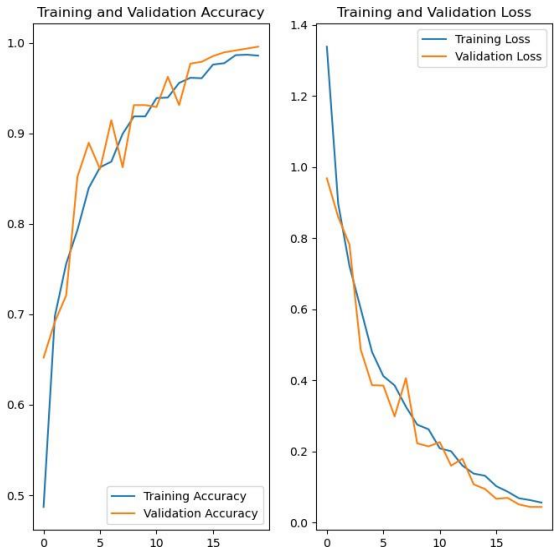

Figure13. Training and validation accuracy and training and validation loss in VGG16 model

4. Conclusion

In agriculture some of the major problems for not getting good yield is unable to find the suitable crop according to soil and using more chemicals and fertilizers without knowing the actual disease that crop was infected. To overcome these problems we have used the latest technologies. By using random forest classifier we can find the proper crop that suit for the properties of soil. And CNN as well as VGG16 model for disease classification. We hope our work may help the farmer for the crop maintenance. And also for machine learning researchers we are concluding that random forest classifier as a ensemble learning will be giving the highest accuracy of 99.09% for predicting and recommending of crop. And our VGG16 model was giving the accuracy of 99.57%. This may help the farmers to find which crop is more suitable for their crop and which chemical should be used by finding the exact disease. And also in future we try improve our model and create some web based applications so that the farmers can access our model and use it more efficiently.

5. References

1) Kumar, V. Vinoth, et al. "Paddy plant disease recognition, risk analysis, and classification using deep convolution neuro-fuzzy network." Journal of Mobile Multimedia (2022): 325-348.
2) Li, Lili, Shujuan Zhang, and Bin Wang. "Plant disease detection and classification by deep learning—a review." IEEE Access 9 (2021): 56683-56698.
3) Shirahatti, Jyoti, Rutuja Patil, and Pooja Akulwar. "A survey paper on plant disease identification using machine learning approach." 2018 3rd International Conference on Communication and Electronics Systems (ICCES). IEEE, 2018.
4) Pudumalar, S., et al. "Crop recommendation system for precision agriculture." 2016 Eighth International Conference on Advanced Computing (ICoAC). IEEE, 2017.
5) Kumar, Avinash, Sobhangi Sarkar, and Chittaranjan Pradhan. "Recommendation system for crop identification and pest control technique in agriculture." 2019 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2019.
6) Katarya, Rahul, et al. "Impact of machine learning techniques in precision agriculture." 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE). IEEE, 2020.
7) Priyadharshini, A., et al. "Intelligent crop recommendation system using machine learning." 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, 2021.
8) Shruthi, U., V. Nagaveni, and B. K. Raghavendra. "A review on machine learning classification techniques for plant disease detection." 2019 5th International conference on advanced computing & communication systems (ICACCS). IEEE, 2019.