Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 291–300

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Leaf Disease Detection and Classification

B.V. Nikith[a], N.K.S. Keerthan[b], Praneeth M.S[c], Dr. Amrita T[d,*]

[a]*Department of Computer Science, Amrita Vishwa Vidyapeetham, Bangalore-560035, India*
bDepartment of Computer Science, Amrita Vishwa Vidyapeetham, Bangalore-560035, India
[c]*Department of Computer Science, Amrita Vishwa Vidyapeetham, Bangalore-560035, India*
[d]*Department of Chemistry, Amrita Vishwa Vidyapeetham, Bangalore-560035, India*

Abstract

The notion of smart farming is gaining traction in the agricultural industry these days, and it makes use of sensors and a variety of machine learning based technologies. According to recent surveys, 56 percent of the agricultural industry is facing significant losses because of diseases developing on plant leaves. It's critical to keep track of the disease's spread and enhance agricultural yields. To prevent the disease from spreading, we must first recognize it on time and prevent it. As a result, we may solve this problem by putting in place some algorithms for detecting sickness on leaves. This paper presents a comparative analysis between support vector machines (SVM) model, K-Nearest Neighbor (KNN) model and convolution neural network (CNN) model. The three different models are presented and examined in this research, and they can detect eight different leaf diseases. The CNN model has achieved an accuracy of 96 percent when trained with the images of soyabean leaf disease dataset, outperforms the KNN and SVM models, which have accuracy of 64 percent and 76 percent, respectively.

Keywords: SVM; KNN; kernels; HOG; CNN.

1. Introduction

India is the world's second-largest country, making feeding a huge population a difficult undertaking. In addition, we are experiencing a food crisis as well as a sharp rise in food prices. The main cause of the scarcity is the spread of illnesses in the crops, which has an impact on agriculture as well as eroding the soil and rendering the land infertile. Fungi and bacteria are the causes of leaf diseases. Early detection of diseases can help to halt the transmission of the

* Corresponding author. Tel.: +91 9901685326.
E-mail address: t_amrita@blr.amrita.edu

1877-0509
10.1016/j.procs.2023.01.011

disease and prevent it from escalating. The earlier we detect the disease, the more time we have to treat it and prevent the loss of the crop.

The infections in the leaves might affect the plant's survival, reducing the plant's life span to only 2-3 years. Plant illness can damage the plant's reproduction rate, resulting in inedible seeds. These seeds affect the soil, making it infertile for the plant's growth. The disease has also impacted fresh plants sown in the field, and the disease has been passed down through the generations in the soil, causing crop failure. During illness, the metabolism and transport of nutrients are disrupted.

It used to take a long time to figure out which disease was present in the plant, and by the time we figured it out, the sickness had spread throughout the entire crop. To prevent crop loss, we must adopt modern technologies such as AI and machine learning. We'll write the code using SVM, KNN, and CNN approaches, and we'll give a dataset for training and testing the algorithms. Various types of soyabean leaf illnesses are induced by temperature fluctuations or other bacterial infections, according to the dataset utilized in our investigation [1].

This is a multi-class classification problem due to the large number of labels to be classified. The diseases being classified are as follows:

- Bacterial Blight: It's a common soybean disease that occurs more frequently in cool and damp conditions. This disease is mostly found at low levels, and the infection can be transferred by seeds.
- Brown Spot: This illness is caused by a fungal or bacterial infection on the leaves, as well as inconsistent plant watering. The disease can be diagnosed by the numerous large spots that appear on it.
- Copper Phytotoxicity: The disease is caused by a high concentration of copper in plant tissues, which is sprayed frequently across a vast area, as well as a lack of rain in that location.
- Downy Mildew: It's a foliar disease caused by a fungus-like organism. It is spread from plant to plant via airborne spores. It is a wet-weather sickness because the infection is aided by prolonged leaf moisture.
- Healthy: This class contains a set of healthy leaves that can be used to categorise the leaf when it is free of disease.
- Powdery Mildew: White mold is caused by high humidity and a lack of airflow. Planting your vegetation too close together, preventing sufficient air circulation, or overwatering your garden or potting soil might encourage the growth of white mold.
- Southern Blight: Southern blight is caused by the fungus Sclerotium rolfsii. Because this fungus is only active during hot weather, it can probably attack all herbaceous perennials., plants can grow healthily in infested soil throughout the growing season and are only destroyed during the hottest portion of the summer.
- Soyabean Mosaic Virus: This illness appears in the winter and then vanishes in the summer. The virus is spread by aphids and via seed. If the virus is present and can be propagated, environmental circumstances that support aphid proliferation can favour this disease.

The paper starts with the work relating to disease detection in leaves using machine learning algorithms, followed by a few fundamental definitions that should be helpful to understand the theory discussion, implemented various algorithms for disease identification in leaves, and then present our findings.

2. RELATED WORKS

Paper [2] briefs about the leaf detection in tomato plant using CNN, for that they used transfer learning concept and imported ResNet-50 model, they used a dataset consisting of 12006 images and divided the dataset into 80% and 20% for training and validation of the model. They achieved an accuracy of 97%, and due to this high amount of accuracy, the model can detect the disease within the shortest period.

Whereas in [3], they implemented using CNN based Alex Net model and compared with VGG-16 and Lenet-5 models using a dataset consisting of approximately 7000 images. They achieved an accuracy of 96.7% and used some basic ML algorithms like SVM and KNN but they got a lower accuracy than VGG-16 and Lenet-5.

In paper [4] some of the diseases are segmented using Otsu's method and used Local Binary Patterns (LBP) and HOG to separate various features from it. Classified the data using the SVM technique and achieved an accuracy of 94.6% using a polynomial kernel. So early stage of detection of plant disease will help the crop from destroying its yield.

Paper [5] deals with effective recognition of infected leaves using the K-means clustering algorithm, it detects the fainted or diseased part of the leaf. SVM and KNN algorithms are also used to compare the accuracies and for SVM they achieved an accuracy of 95%, whereas for KNN they achieved an accuracy of 85%.

3. Machine learning techniques

Some basic insights on the concepts explored in this project will give a general overview of the algorithm's creation and implementation.

3.1. Support Vector Machines

SVM is a fundamental machine learning model that, in the case of linear data, produces the most optimal Hyperplane and, in the case of non-linear data, uses the concept of Kernels. After the model receives the input photos, it calculates the number of features that can be retrieved from the data. The data is then plotted, and several hyperplanes are drawn to divide the data into classes [6].

SVM kernels are mostly used for non-linear data, and they transfer data from a lower-dimensional space to a higher-dimensional one by using a hyperplane to separate the data. There are primarily two types of kernels that can be used for this type of data; to determine the most accurate kernel, we must first understand the Hyperparameter tuning concept; in this concept, all kernels are used to check the accuracy of the data; the kernel that gives the highest accuracy is said to be the best kernel for that data [7,8].

There are 2 types of Kernels they are:

- Polynomial Kernel
- Radial Basis Function Kernel (RBF kernel)

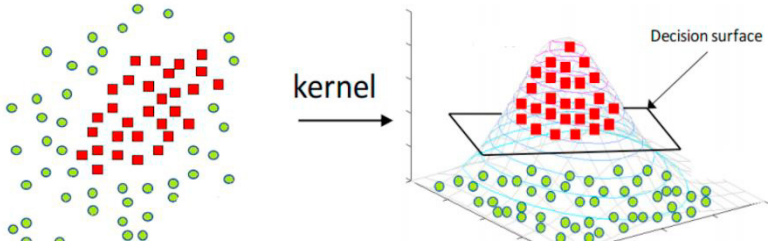


Fig. 1. Kernel procedure [9]

Fig. 1. depicts how a non-linear decision boundary is formed using the kernel method, which converts a 2D image to a 3D image. It finds the best hyperplane to fit the data and then projects it back to 2D.

3.2. K – Nearest Neighbors (KNN)

This supervised learning model divides the data into classes so that when a new data point enters the model, it first sees its K-nearest neighbours, then the neighbour with the highest frequency among those neighbours, and the data point is aligned with the class. This step is repeated for all data points that have been properly categorized and entered the model [10].

We need run the model for some k values to get an accurate k value. It is simple to categorize the data if the k value is odd. Then we can visualize the data or, more commonly, determine the optimal value based on the accuracy of all the numbers [11].

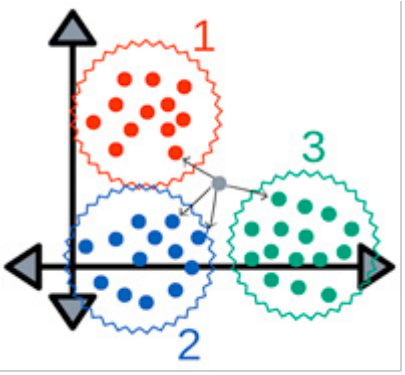


Fig. 2. K – Nearest Neighbors [12]

In Fig. 2. the data has been divided into 3 classes red, blue, green and when a new data point of color grey enters the model it sees its nearest neighbors from that class, and the point is classified to a class with maximum neighbors in this model the grey point classifies to blue class.

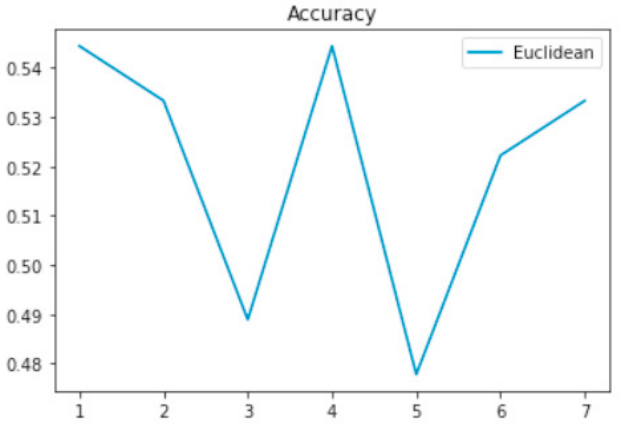


Fig. 3. Accuracy plot for best K-NN

Fig. 3. shows the graph for the data represented in Fig. 2. Here for the K = 4, it shows good accuracy of 60% and also the graph has been measured based on Euclidean distance. The HOG feature is the factor that is extracted from the image.

3.3. Convolutional Neural Network

It is a subdomain of Deep Learning algorithms in which these algorithms, when implemented in a real-time environment, view the world through the eyes of humans and work in a manner like humans, and these algorithms are trained in all the possible ways humans can think about a given scenario in each situation. These machines have grown into all aspects of life, from our everyday mobile phones to accurate supercomputers. They are replacing humans with advanced machines to increase the productivity of the labour [13].

There are 3 layers in this CNN:

- Input Layer
- Hidden Layer
- Output Layer

•Input Layer: The input layer consists of images provided as input to the model; in this layer, image data is saved in pixels and stored in nodes, and every operation in CNN occurs in nodes.

•Hidden Layer: This layer's hidden layer handles with computations such as data processing, feature extraction, and data transformation. The more hidden levels in the CNN design, the more complex the architecture becomes.

The data processing in the hidden layer allows the model to learn features and perform quickly with real-time data, making it more equivalent to the CNN models currently in use. The CNN's activation functions determine whether a neuron should be activated based on the node's value.

In this project, we used the features retrieved from the hidden layer to recognize and predict the disease presence on the leaves.

•Output Layer: The output layer is a fully linked layer where the hidden layer's outputs are flattened and taken as an input. This layer accepts input and turns it into the desired classes [14].

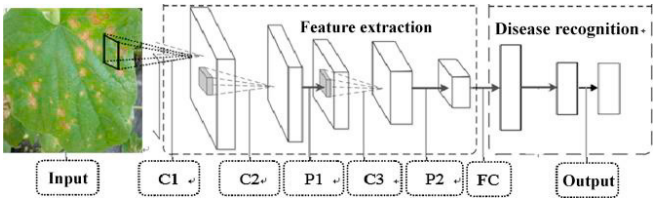


Fig. 4. Methodology used [15]

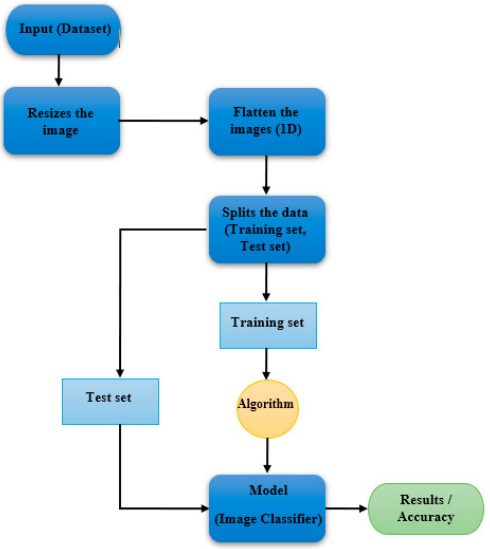
Fig. 4. demonstrates how the features are extracted and then the disease is identified using CNN.

4. IMPLEMENTATION

Initially for the soyabean dataset obtained from reference [13] various machine learning algorithms are imported, and then SVM, KNN, and CNN algorithms build in python are used to identify the type of disease present in the leaf.

4.1. Support Vector Machines

a)



b)

```
GridSearchCV(estimator=SVC(probability=True),
             param_grid={'C': [0.1, 1, 10, 100],
                         'gamma': [0.0001, 0.001, 0.1, 1],
                         'kernel': ['rbf', 'poly']})
```

Fig. 5. (a) SVM Implementation; (b) SVM model fitting parameters

Fig. 5. (a)demonstrates how SVM is used. Following the import of the dataset, all photos are resized and flattened, transforming 2D images to 1D images. The dataset is then divided into train data and test data in a 70:30 ratio, with the train data serving as an input to the algorithm and the test data serving as a check on the algorithm's accuracy,

before the SVM algorithm is run. Once the accuracy has been anticipated to determine disease in the leaf, test images are presented as sample input, and (b) shows the model parameters used for fitting the images in the SVM model.

4.2. K Nearest Neighbors (KNN)

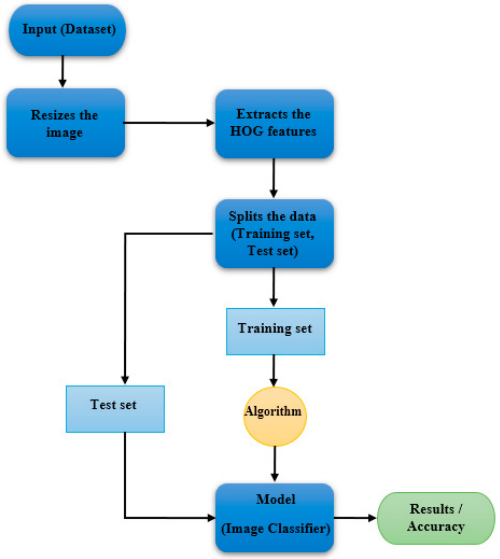
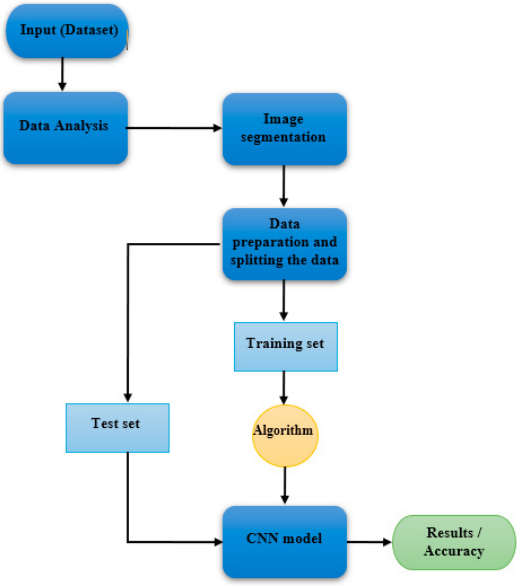


Fig. 6. KNN Implementation

Fig. 6. demonstrates how KNN is used. Instead of flattening the images, KNN extracts HOG features using Euclidean distance and Manhattan distance from them and then splits the dataset in the same way that SVM does, however KNN algorithm is used instead of SVM and a 75:25 split between train and test data, Manhattan distance provides greater accuracy in the KNN algorithm when compared to Euclidean distance. To predict the disease, test images from test data are used as sample input.

4.3. Convolution Neural Networks (CNN)

a)



b)

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 72, 72, 32)	896
max_pooling2d (MaxPooling2D)	(None, 36, 36, 32)	0
conv2d_1 (Conv2D)	(None, 36, 36, 128)	36992
max_pooling2d_1 (MaxPooling2D)	(None, 18, 18, 128)	0
conv2d_2 (Conv2D)	(None, 18, 18, 256)	295168
max_pooling2d_2 (MaxPooling2D)	(None, 9, 9, 256)	0
flatten (Flatten)	(None, 20736)	0
dense (Dense)	(None, 128)	2654336
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 10)	650

Fig. 7. (a) CNN Implementation; (b) CNN model simulation parameters

Fig. 7. (a) demonstrates how CNN is used. Following the import of the dataset, all images are subjected to data analysis, which crops the images and extracts leaf pixels via image segmentation. K means clustering is used for image segmentation. Because the images are in RGB colour format, performing K means clustering to find the best colour channel by splitting the image into various colours available in individual channels is difficult. The b channel is ideal for K means clustering; K means clustering is performed on the b channel, and the component to which the leaf belongs is then discovered. Once the component is identified, the CNN algorithm is used to create a model, and the dataset is divided into train and test sets, which are used to predict accuracy, and (b) shows the model summary describing the various layers used while building the model.

5. RESULTS

5.1. Support Vector Machines (SVM)

The “Soyabean leaf” Dataset when trained with SVM model gave an accuracy of 76%. When predicted with some test images it is predicting with a good accuracy.

```
The predicted Data is :
[0 3 0 8 0 0 9 8 8 0 0 8 8 8 3 5 5 5 3 7 3 3 4 8 3 5 0 3 3 3 8 0 8 5 0 0 5
 9 1 7 8 8 3 8 5 5 8 8 7 3 9 5 5 5 3 8 3 0 5 3 1 8 8 5 8 3 5 8 2 5 4 9 0 0
 8 5 5 0 1 9 5 0 5 8 5 2 8 2 5 3 5 5 0 7 5 5 8 8 5 3 0 0 1 3 0 7 5 0 0 0 0
 8 7 3 8 7 5 3 5 5 8 8 9 8 3 1 8 5 0 2 8 2 0 0 7 0 3 0 5 5 0 0 5 8 5 5 3 0
 7 1 0 8 3 8 2 9 3 8 8 2]
The actual data is:
[2 3 0 3 7 0 9 8 5 2 0 1 3 8 1 5 5 1 3 7 3 3 4 2 1 5 0 5 3 1 8 0 8 5 0 2 5
 9 1 7 8 8 3 8 9 5 8 8 7 3 9 8 5 5 3 8 3 0 5 3 1 8 8 8 8 8 5 8 2 5 4 9 0 0
 8 5 3 0 1 9 9 7 8 8 8 2 8 2 5 3 5 9 0 7 5 5 5 8 8 3 0 0 6 3 2 7 5 0 0 0 0
 8 7 3 8 7 8 3 1 5 8 3 9 8 3 1 5 5 0 2 5 2 0 0 7 0 9 0 5 8 0 0 5 3 0 5 3 0
 7 1 0 8 3 8 2 9 3 8 8 2]
The model is 76.875% accurate
```

Fig. 8. SVM Model Results

Fig.8. demonstrates the results obtained with SVM model. The predicted model is tested with some test cases.

a)

b)

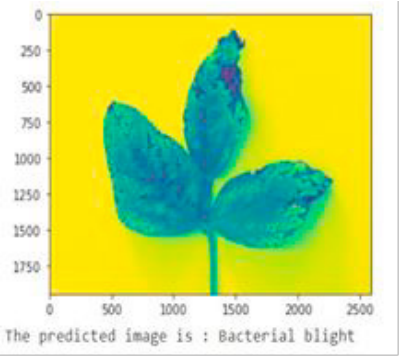


Fig. 9. (a) Input test image; (b) Output predicted image

Fig.9. (a) shows the input test case that was used to identify the disease using the SVM model, and (b) shows the ailment that will be caused by the provided input.

5.2. K Nearest Neighbors (KNN)

When training the KNN model with the "Soyabean leaf" Dataset, we get an accuracy of 64% and the best value of k is 6. Here both Euclidean and Manhattan distance is used to find the value of k. The results obtained are analyzed in the plot.

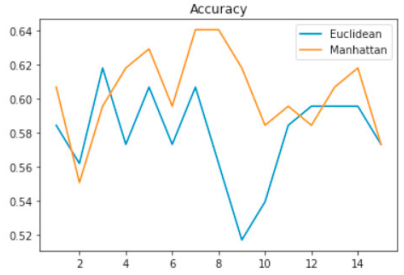


Fig. 10. KNN accuracy obtained at different values of K

We can see from Fig. 10. that in this case Manhattan distance gives better accuracy compared to Euclidean distance. The KNN model predicted is tested with some test cases.

Fig. 11. Input test image

Fig.11. indicates the input test case used to identify the disease using KNN model.

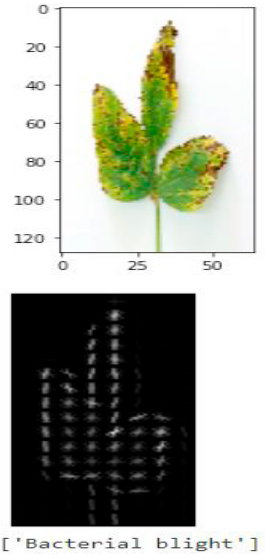


Fig. 12. Output predicted image

Fig.12. indicates the predicted disease for given input.

5.3. Convolution Neural Networks (CNN)

When the CNN model is trained using the "Soyabean leaf" Dataset, we get 96% accuracy rate. The accuracy at different points can be analyzed from the below figure.

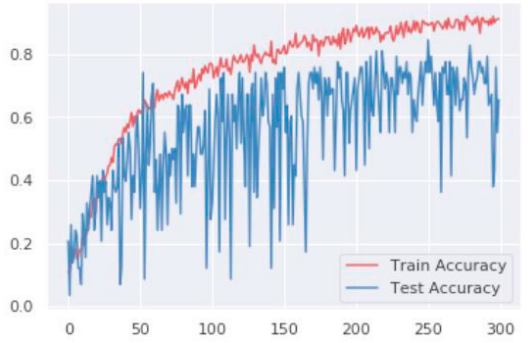


Fig. 13. Training and testing accuracy and accuracy plots

Fig.13. demonstrates the accuracy of the CNN model as well as the accuracy plot obtained for the train and test datasets.

6. CONCLUSION

Leaf disease detection is being implemented in this paper for the multi-class classification dataset using three models SVM, KNN, and CNN. It presents a real time image classification and detection for a disease present on leaves which helps farmers for correct use of pesticide for those diseases detected and eradicating it's spread. Based on the results obtained, we can conclude that CNN has an accuracy of 96%, while SVM and KNN have accuracy of 76% and 64%, respectively, so according to our findings CNN is better than SVM and KNN. When the dataset size is increased, the accuracy of the SVM and KNN models may improve. Hence CNN is a better model for detecting diseases in leaves because it is neural network model.

Table. 1. Comparing the accuracies

Model	SVM	KNN	CNN
Accuracy	76%	64%	96%

Table. 1. Compares the accuracy of SVM, KNN and CNN model developed.

7. FUTURE SCOPE

To diagnose the leaf diseases with greater precision, many additional machine learning algorithms can be developed to increase the accuracy. When a huge dataset is being used, systems with additional GPUs can be employed, or a cluster of several systems can be set up. To assist farmers and improve their life, particularly to assess the calibre of output being generated, a real-time application for live photos might be created for the successful identification of leaf disease.

References

[1] https://www.digipathos-rep.cnptia.embrapa.br/jspui/simple-search?filterquery=Soja+%28Soybean%29&filtername=crop&filtertype=equals

[2] N. K. E., K. M., P. P., A. R. and V. S., "Tomato Leaf Disease Detection using Convolutional Neural Network with Data Augmentation," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1125-1132.

[3] P. B R, A. Ashok and S. H. A V, "Plant Disease Detection and Classification Using Deep Learning Model," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1285-1291.

[4] M. E. Pothen and M. L. Pai, "Detection of Rice Leaf Diseases Using Image Processing," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 424-430.

[5] J. N. Reddy, K. Vinod, and A. S. Remya Ajai, "Analysis of Classification Algorithms for Plant Leaf Disease Detection," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-6.

[6] N. R. Bhimte and V. R. Thool, "Diseases Detection of Cotton Leaf Spot Using Image Processing and SVM Classifier," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 340-344.

[7] P. Krithika and S. Veni, "Leaf disease detection on cucumber leaves using multiclass Support Vector Machine," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 1276-1281.

[8] Y. K. Dubey, M. M. Mushrif and S. Tiple, "Superpixel based roughness measure for cotton leaf diseases detection and classification," 2018 4th International Conference on Recent Advances in Information Technology (RAIT), 2018, pp. 1-5.

[9]https://www.google.com/search?q=svm+non+linear%2C+linear%2C+and+kernal+decision+boundary&tbm=isch&ved=2ahUKEwjKycPnpYX5AhXhoukKHYNdA9YQ2-cCegQIABAA&oq=svm+non+linear%2C+linear%2C+and+kernal+decision+boundary&gs_lcp=CgNpbWcQAzoECCMQJ1C6C1jcNGD-NmgBcAB4AIAB1QGIAY4VkgEGMi4xOS4xmAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=atHWYsrsJOHFpgeDu42wDQ&bih=714&biw=1536&rlz=1C1UEAD_enIN991IN991#imgrc=xj-P7V0nvaWoUM

[10] M. P. Vaishnnave, K. S. Devi, P. Srinivasan and G. A. P. Jothi, "Detection and Classification of Groundnut Leaf Diseases using KNN classifier," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-5.

[11] S. Veni, R. Anand, D. Mohan and P. Sreevidya, "Leaf Recognition and Disease Detection using Content based Image Retrieval," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 243-247.

[12] https://images.app.goo.gl/7y9H518kMhF7RcZT6

[13] A. Jenifa, R. Ramalakshmi and V. Ramachandran, "Cotton Leaf Disease Classification using Deep Convolution Neural Network for Sustainable Cotton Production," 2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES), 2019, pp. 1-3.

[14] S. Kumar, K. Prasad, A. Srilekha, T. Suman, B. P. Rao and J. N. Vamshi Krishna, "Leaf Disease Detection and Classification based on Machine Learning," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 361-365.

[15] https://images.app.goo.gl/FfagRNbaTV7mDrCj6