ASSIGNMENT- 3

CLOUD BASED APPLICATION FOR DISEASED PLANT DETECTION

NAME	STUDENT NUMBER	EMAIL ID
Ragu Sankar Nagendran	6112456	rsn386@uowmail.edu.au
Pedda Babu Asapu	6165527	pba704@uowmail.edu.au
Abhishek Balaji Sharma	6237083	abs105@uowmail.edu.au
Varad Girish Kusurkar	6261358	vgk234@uowmail.edu.au

Introduction

In order to understand semi-supervised learning (SSL), it is very necessary to understand what is Supervised learning (SL) and what is unsupervised learning (UL). Because semi-supervised learning is somehow lying between supervised and unsupervised learning.

Considering the set of 'n' points which are not dependent on each other also which are similarly distributed then the mathematical equation for unsupervised learning is given as

$$X = (x_1, \; . \; . \; . \; . \; . \; . \; , \; x_i)$$

Where, $x_i \in X \, for \; all \; i \in [n] \coloneqq \{1, \; . \; . \; . \; . \; , \; n\}$

The major goal of this learning is to recognize a particular pattern in a given data set say 'X' which will possibly generate again that data set(X). The basic difference between unsupervised learning and supervised learning is the supervised learning uses label data. The training set for supervised learning is (x_i, y_i) whereas y_i are the labels for the x_i. Basic functionality of supervised learning is for the set of unlabelled data Y mapping has been carried out with label data X. So, it can be concluded that mapping is a very important aspect of supervised learning[1].

Supervised learning works on label data on the other hand unsupervised learning uses unlabelled data. Handling of label data is very easy, but it is very difficult to get label data. Moreover, label data is also very costly. On the other hand, it is a very difficult task to predict model from unlabelled data, but unlabelled data is easily available also it is less costly. Semi-supervised learning has very few label data and plenty of unlabelled data. The basic functionality of Semi-supervised learning is using particular patterns in label data it will try to convert given unlabelled data into label data.

The mathematical equation of semi-supervised learning is, consider the data set X = $(x_i)_{i\in[n]}$. There are two major parts of this equation. X_l: = (x_1, . . . ,x_l) which is having mapping Y_l := (y_1, . . . , y_l). That is Y_1 is mapped to X_1. The second part is X_u: = (x_{l+1}, . . . ,x_{l+u}) which is unlabelled data in the given equation. This unlabelled data doesn't have any mapping and using multiple patterns generated by label data it needs to find a mapping for given unlabelled data.

Semi-supervised learning is a very important aspect of machine learning environment because it allows conversion of unlabelled data into label data. As already mentioned, label data is very costly and it's not easily available and unlabelled data is very difficult to handle, so this conversion becomes a very necessary and essential aspect of the machine learning environment. Semi-supervised learning is a very important aspect of machine learning environment because it allows conversion of unlabelled data into label data. As already mentioned, label data is very costly and it's not easily available and unlabelled data is very difficult to handle. The human efforts that have been Involved in finding label data have been minimized by semi-supervised learning[1].

There are different forms of semi-supervised learning. These forms mainly classified as a probabilistic generative model of semi-supervised learning, the second type is a graphical model of semi-supervised learning, support vector machine, self-training, co-training, etc.

Though semi-supervised learning is a very effective method in order to generate label data, but it has some fundamental limitations. Unlike supervised or unsupervised learning, it must deal with both labels as well as unlabelled data at a given time which will increase the complexity of the system. While building the model some assumptions must be considered for SSL, but the huge error will occur when this model fails.

There are various applications associated with semi-supervised learning.

- Classification of web page
- Transcription of telephone conversations.
- Video surveillance
- Natural language processing.

The aim of this project is to develop an app for the small landholder farmers so that by analyzing plant by that app the disease associated with the plant can be determined. Symptoms of these diseases can be found by observing various physical changes associated with the plants. Once disease have been identified it can be further sent to experts for further future investigation. Food safety is a very critical issue now a days; by using this machine learning based solution concentrate solution can be found.

Background Theory

Semi-Supervised Learning

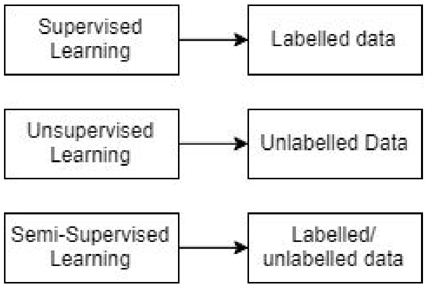


Figure 1 - Different kinds of machine learning

From above figure it can be seen that supervised learning uses labelled data on the other hand unsupervised learning uses unlabelled data. But semi-supervised learning lies between supervised and unsupervised learning which uses combination of labelled and unlabelled data. Basically, semi-supervised learning uses small amount of labelled data and from the patterns of labelled data it predicts unlabelled data. As labelled data is very difficult moreover it is very costly, so this conversion become very essential[1].

Support Vector Machine (SVM)

SVM is a classifier that is used for the feature identification in convolutional neural networks. The features can be distinguished by a hyperplane that brings in a set of features together considering the factors such as intensity, threshold, colour, texture and shape of the image. It is an automatic classifier and works efficiently and consumes less time for the feature identification process.

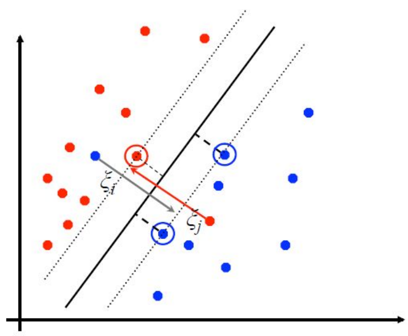


Figure 2 - Support Vector Machine (SVM)

Red dots indicate dataset of class X comprises of data $x_1, x_2, \ldots x_n$
Blue dots indicate dataset of class Y comprises of $y_1, y_2, \ldots y_n$

Best classifier is obtained when the distance between the closest point of x_i and the line as well as distance between closest point of y_i and the line is equal.

Convolutional Neural Networks (CNN)

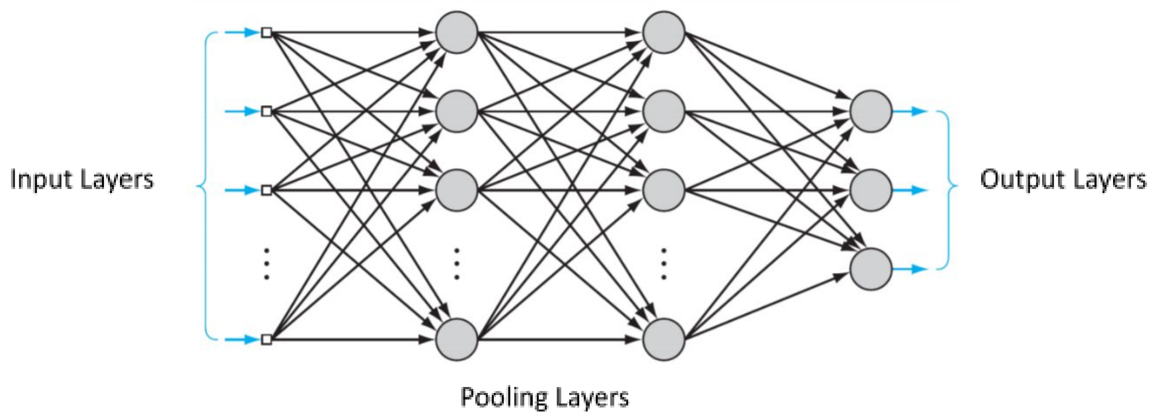


Figure 3 - Convolutional Neural Networks (CNN)

Convolution neural network is used in the process of feature reduction techniques. The number of features reduce as the layers increase with selective features selected for the output layer. The pooling layers consists of number of features in the algorithm. These features can be obtained by adding the filters between the pooling layers as per the necessity of the outcome of the algorithm. Specific features can be obtained through another technique called the Deconvolution method or the back-propagation algorithm that can be used as the feature selection process for the above network. Convolution neural network is one of the best methods for feature reduction technique process.

Literature Review

S. H. Lee, C. S. Chan, P. Wilkin and P. Remagnino, "Deep-plant: Plant identification with convolutional neural networks," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 452-456.

The author has explained how a pre-trained convolutional neural network (CNN) model as well as deconvolutional network (DN) model can be used in plant identification. Plant classification is performed by learning leaf features with the help of CNN and the learned features are visualized by using DN. The shape of the leaf is identified by using the curvature with the help of CNN model. Whereas in DN, the venation is performed in which structure of veins in a leaf is found. The author used a dataset consisting of 44 different plant species.

The CNN model the author used comprises of five convolutional layers of which output of each is layer is fed to next layer thus reducing kernels and size. A fully connected layer is obtained after performing convolution and pooling in the fifth layer which has 4096 neurons. In the DN model, back propagation mechanism is used to optimize the filter in each layer.

These optimized filters will extract unique and important features in leaf image. As a result, the author has obtained a new dataset named as MalayaKew (MK) leaf dataset with 44 classes. In order to enlarge the dataset, author rotated the leaf images in seven different orientations such as 45°, 90°, 135°, 180°, 225°, 270°, 315°. The author observed in a few cases the shape of leaves of two different plants will be similar and, in such case, it is not a good idea to identify the plant using the shape of the leaf. Finally, the author has concluded that the venation structure is important in plant identification than finding the shape of the leaves.

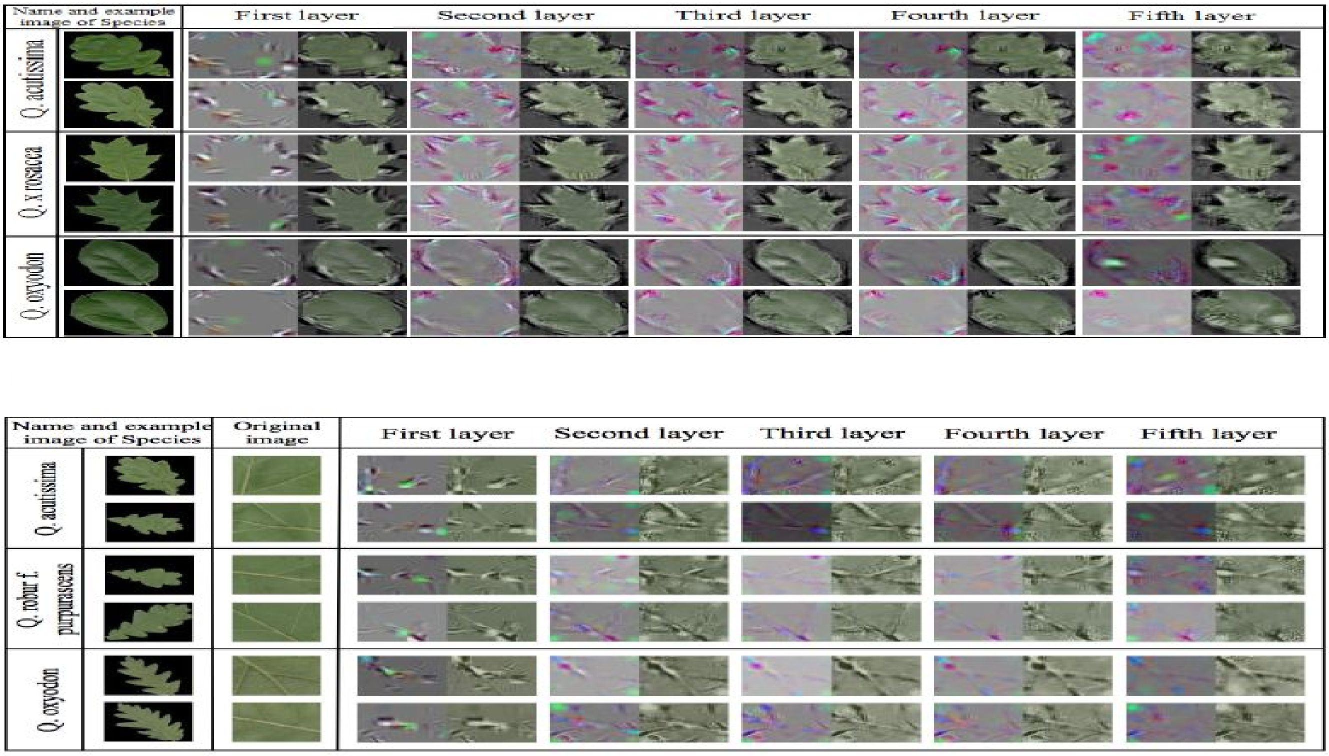


Figure 4 - Feature Visualization using Deconvolutional Network (DN)

$$Y_i^{(l-1)} = \sum_{j=1}^{m_1^{(l)}} \left(K_{j,i}^{(i)}\right)^T * Y_j^{(l)}$$

Equation 1 – Back Propagation

Where, 'l' indicates convolutional layer

'K' are the filters

'n' is feature map

The deconvolution process begins from (l-1) layer for all filters from j to i.

Ma, J., Du, K., Zhang, L., Zheng, F., Chu, J. and Sun, Z., 2017. A segmentation method for greenhouse vegetable foliar disease spots images using colour information and region growing. *Computers and Electronics in Agriculture*, *142*, pp.110-117.

The author has proposed a method for image processing with the help of colour details as images captured in real field conditions will have uneven illumination and high noises in background which will rise difficulty in disease spots identification. This method comprises two steps of which the first step involves selecting a comprehensive colour feature and the second step involves achieving disease spot identification by an interactive growing method from clutter background. The author used cucumber downy mildew images to assess the proposed algorithm and achieved high accuracy. The proposed method groups ExR (Excess Red Index), b* component of L*a*b* colour space, and H component of HSV color space. An ExR parameter is endorsed in the range of 0 and 1 to reduce the effect of illuminations. A CCF (Comprehensive Colour Feature) ratio is proposed which gives closeness of CCF value of normal leaf region and light influenced leaf region. If the CCF value is small, it means that there is a small amount of illumination of light on the leaf.

CCF ratio formula is given as

$$R = \frac{\frac{1}{M}\sum_{i=1}^{M} CCF(x,y)}{\frac{1}{N}\sum_{j=1}^{N} CCF(x,y)}$$

Equation 2 – CCF Ratio

Where **R** is CCF ratio,

M is the number of pixels in the light influenced leaf region,

N is the number of pixels in the normal leaf region,

CCF (x, y) is the value of the corresponding image pixel.

Janwale, Asaram. (2017). Plant Leaves Image Segmentation Techniques: A Review. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 5. 147-150.

The author mentioned the different methods of segmentation techniques that are used to divide a digital image to a number of parts of interest. The Edge detection technique is used to locate the edges and boundaries of the leaf in the image. The threshold technique converts the image to the binary image and the comparison can be done based on the pixel value less than the threshold value. Segmentation can be classified based on splitting, merging and growing. K-means algorithm on the textured images is used in the histograms through clustering segmentation. Watershed segmentation works on the Topological analysis which relates to the gradient of the image as the topographic surface. Fuzzy numerical morphology is used to calculate the edge of the image by increasing the contrast of the edges. One of the algorithms to activate segmentation is to remove similar features, make a histogram using the non-similar features, divide the image into 8-bit grayscale images and finally contrast grey scale images. The colour image segmentation has worked on the two-way process. One of the processes is to pick up the pinnacle level histogram values from the image and combine the features and another way is to combine the hypothesis and few pictures spoken in various shading spaces to get a last dependable and precise division result. The shading spaces is obtained by utilizing the blend administer and choice. K-means clustering algorithm plays out the division assignment of wheat leaf scab image which helps in automatic disease detection. Thus, the author talks about how the segmentation can be done to the image using texture, intensity, and colour.

Pujari, D., Yakkundimath, R. and Byadgi, A.S., 2016. SVM and ANN based classification of plant diseases using feature reduction technique. *IJIMAI*, *3*(7), pp.6-14.

According to author most of the plant diseases are caused by a fungus, bacteria, viral and nematodes. The author focuses on the automatic disease diagnosis using an algorithm by feature reduction techniques. Some of the common plant diseases are stan borer, brown stripes, and dowmy mildew. The disease type classification is performed using the kNN or the nearest neighbour algorithm and adaptive Bayes classifier using the Gaussian mixture models. It is predicted that the adaptive Bayes classifier performs better than its latter. To recognize the diseases such as melanose, greasy spot, and scales, infected areas of the leaf are identified using segmentation techniques. The manual threshold is calculated based on the HSI model colour and is compared with the centroid location method. The centroid location method performs better compared to the manual threshold. Once plant classification takes place through image acquisition, pre-processing of the data or the image takes place to cancel out the shade correction, artefacts removal and formatting. The feature selection can give a quantitative or morphological description of the pattern based on the RGB (Red, Green, Blue) Colour Space along with HSI (Hue, Saturation, Intensity) and the texture features based on the GLCM (Grey-level co-occurrence matrix). The H of the HIS is responsible for the pure colour of yellow, orange and red. The S of the HIS is responsible for the pure colour and the white light and the I of the HIS is responsible for the Intensity of the Grey level. The first level of the feature reduction occurs through the colour feature which is reduced to 8 and the texture feature is the next level of feature reduction that is reduced to 5. Delta is the value between two features in each of the feature reduction. From the author's finding it is clear that SVM classifies the diseases in plants with better classification accuracy and a minimum number of features in the minimum compilation time. However, there is a lot of scope of improvement in the classification of plant diseases through SVM.

$$\mu \frac{1}{N} \sum_{z=1}^{N} x_z = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Equation 3 - RGB pixel value

Where, $x_N = N$ th pixel value

$x_z = z$ th pixel value

'N' is the total Number of pixels

When the RGB component is extracted we can count the number of total pixels in the image using which any required z^{th} pixel can be calculated using the formula mentioned above.

Standard Deviation

$$\sigma = \frac{1}{N} \sum_{z=1}^{N} \sqrt{(x_z - \mu)^2}$$

Equation 4 - Standard Deviation

Where, 'N' is the total number of pixels

'x_z' is the z^{th} pixel value

Variance = $\sigma * \sigma$

Where, 'σ' is standard deviation

HSI components

Hue (H) =

$$\{2 - \left\{ \frac{[(R-G)+(R-B)]}{\sqrt[2]{(R-G)^2+(R-G)(G-B)}} \right\}, B > G \quad \left\{ \frac{[(R-G)+(R-B)]}{(R-G)^2+(R-G)(G-B)} \right\}, \; B \le G$$

Equation 5 - Hue

Saturation (S) = $1 - \left[\frac{3.min(R,G,B)}{R+G+B} \right]$

Equation 6 - Saturation

Intensity (I) = $\frac{R+B+G}{3}$

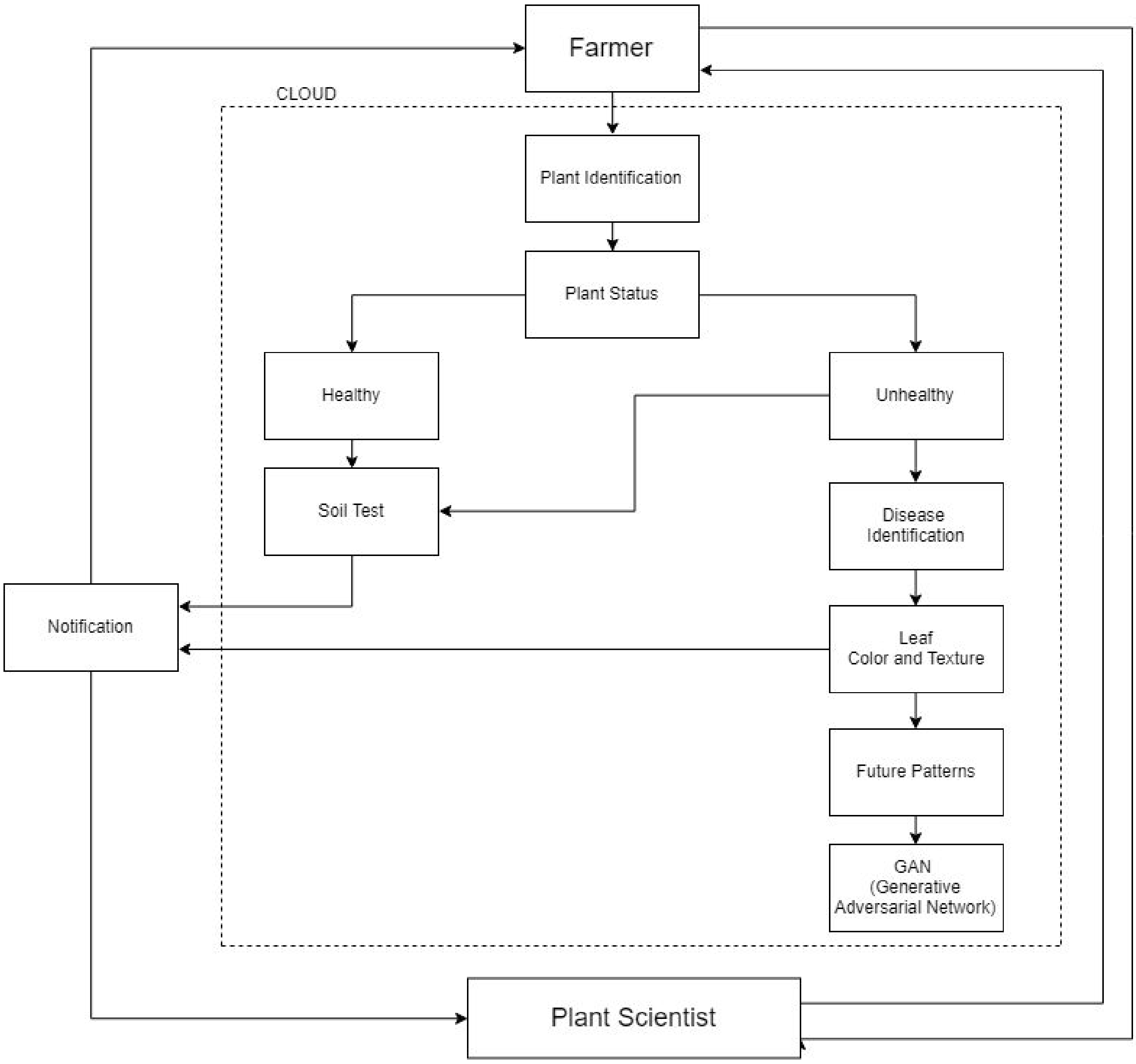
Equation 7 - Intensity

Application

Data Preparation

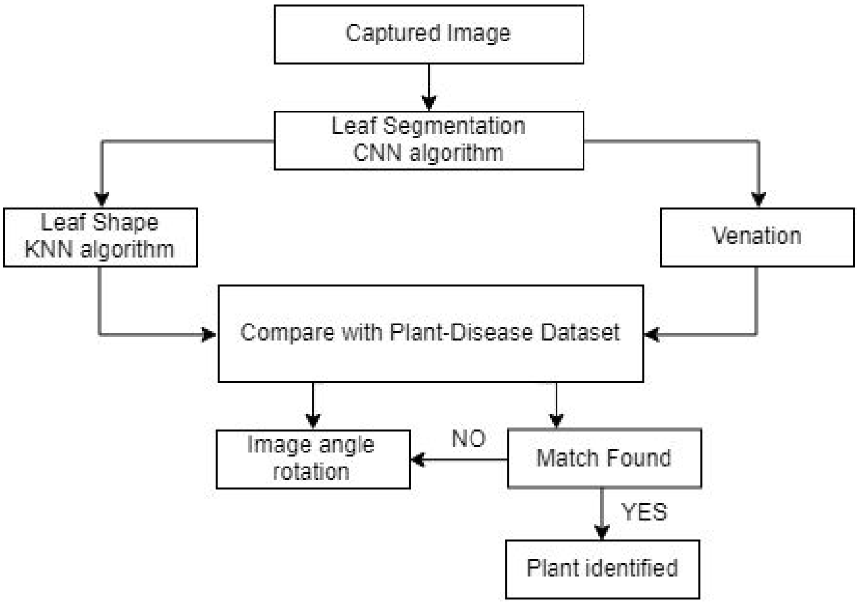
We have used the Plant-Village dataset which was available free of cost in (https://github.com/

spMohanty/PlantVillage-Dataset). This dataset is very large and consists of both healthy and diseased leaf images. This dataset contains 54,306 images. Labelled data consists of pictures of many plants such as Scab Angular, Powdery Mildew, Downy Mildew, and Anthracnose. Labelled data is trained in GAN (Generative Adversarial Network) and a trained model is generated. Then the unlabelled data is segregated and again filtered for labelled data and trained again.



Flowchart 1 - Overall process

From flowchart 1, The farmer clicks the photograph of his farming land on a regular basis. This image is sent to the cloud through the smartphone-based application where the farmer and plant scientist have direct contact with each other[2]. As the image is sent to the cloud, the plant identification process takes place with the help of the internal properties of the leaf of the species. The plant status is helpful to categorize the healthy or unhealthy plant by the SVM (Support vector machine) which distinguishes the features of the species through its algorithm. Once the plant is found to be in healthy condition, soil test takes place to check the presence of oil molecules that are considered to a silent factor to disturb the growth of the plant and the findings of which will be notified to the farmer as well as the plant scientist. In case, the plant is found to be in an unhealthy condition, the colour and the leaf structure of the leaf is sent to the farmer and the plant scientist for the close examination of the plant disease or the forthcoming disaster to the farming land. From the images obtained the future patterns of the dataset is predicted when the outliers are observed from the dataset that is given to the Generative adversarial network (GAN) is used to classify the unlabelled features of the dataset to produce real samples to the merged dataset along with the classification of discriminator network. As the amount of unlabelled dataset increases, the GAN keeps on working.



Flowchart 2 - Plant Identification

From flowchart 2, The captured image is then processed for the leaf segmentation. The purpose of the segmentation is to pick out a particular part of the image and clearing out the unnecessary noise or background features of the segmented image. The segmentation also has the capacity to cancel out the shadow features present in the image[3]. Once this process takes place, the Convolution neural networks (CNN) algorithm is applied to the above procedure to extract the leaf shape and venation of the image which is nothing but the leaf of the plant.

$$P = \sum_{i=1}^{n} f_1 + f_2 + f_3 + \ldots\ldots + f_n$$

Equation 8 - CNN equation

Where, 'P' is the pooling layer

'n' is number of filters

'f' is the filter

$$FM = \sum_{i=0}^{n} p_1 f_1 + p_2 f_2 + \ldots\ldots p_n f_n$$

Equation 9 - Feature Map calculations

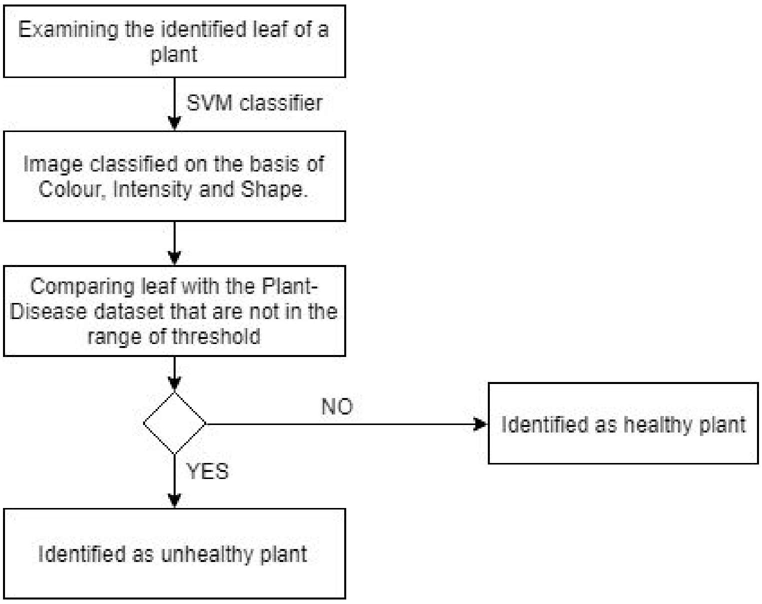
Where, 'FM' is the Feature Map

'n' refers to number of pooling layers and filters,

'p' is the pooling layer,

'f' is the filter.

The shape of the leaf is found through the kNN algorithm which is nothing but the nearest neighbour that is based on the curvature of the shape of the leaf. However, the venation of the leaf is always different from each other. These datasets from the image are compared with the plant dataset. If the conditions match or satisfy, the plant is identified or else the image rotates through the angles 45°, 90°, 135°, 180°, 225°, 270° and 315° degrees to pass the test cases for the plant identification which is compared with the plant dataset.



Flowchart 3 - Plant Status

From flowchart 3, The identified leaf of the plant is examined closely. Now, using the Support Vector Machine (SVM) the image is classified based on the colour, intensity, and shape. After the classification, the comparison of the leaf with plant disease dataset that is not in the range of threshold are found out[4]. Once such images are obtained, we can come to a decision that these can be classified as an unhealthy plant. If none of the datasets of the leaf is out of range from the plant dataset, then the plant can be classified as the healthy plant[5]. The role of the Support Vector Machine is to classify the features of the plant according to various factors such as colour, intensity, shape or the threshold of the pixel values of the leaf image[6]. This gives the status of the plant.

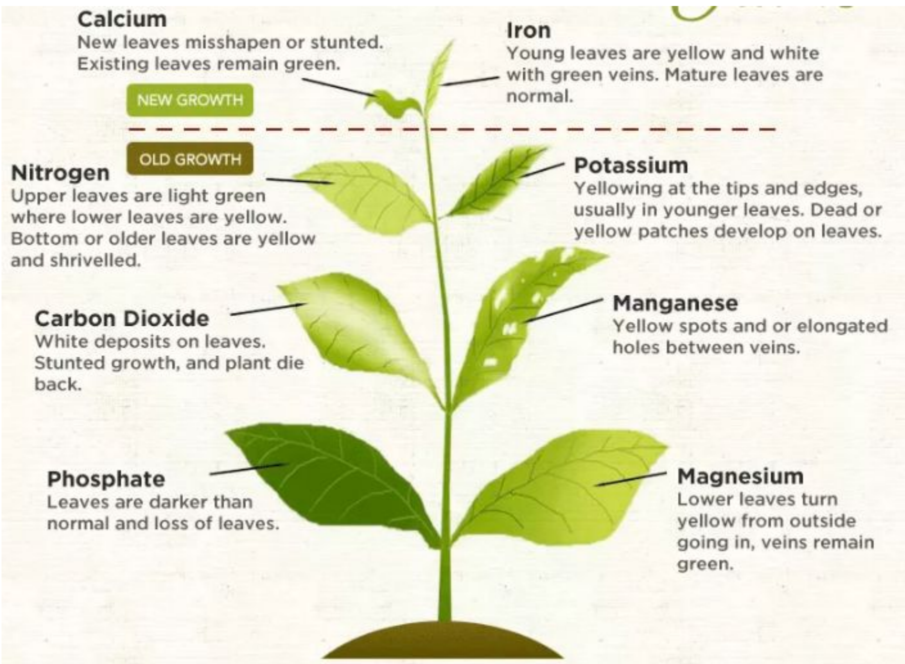
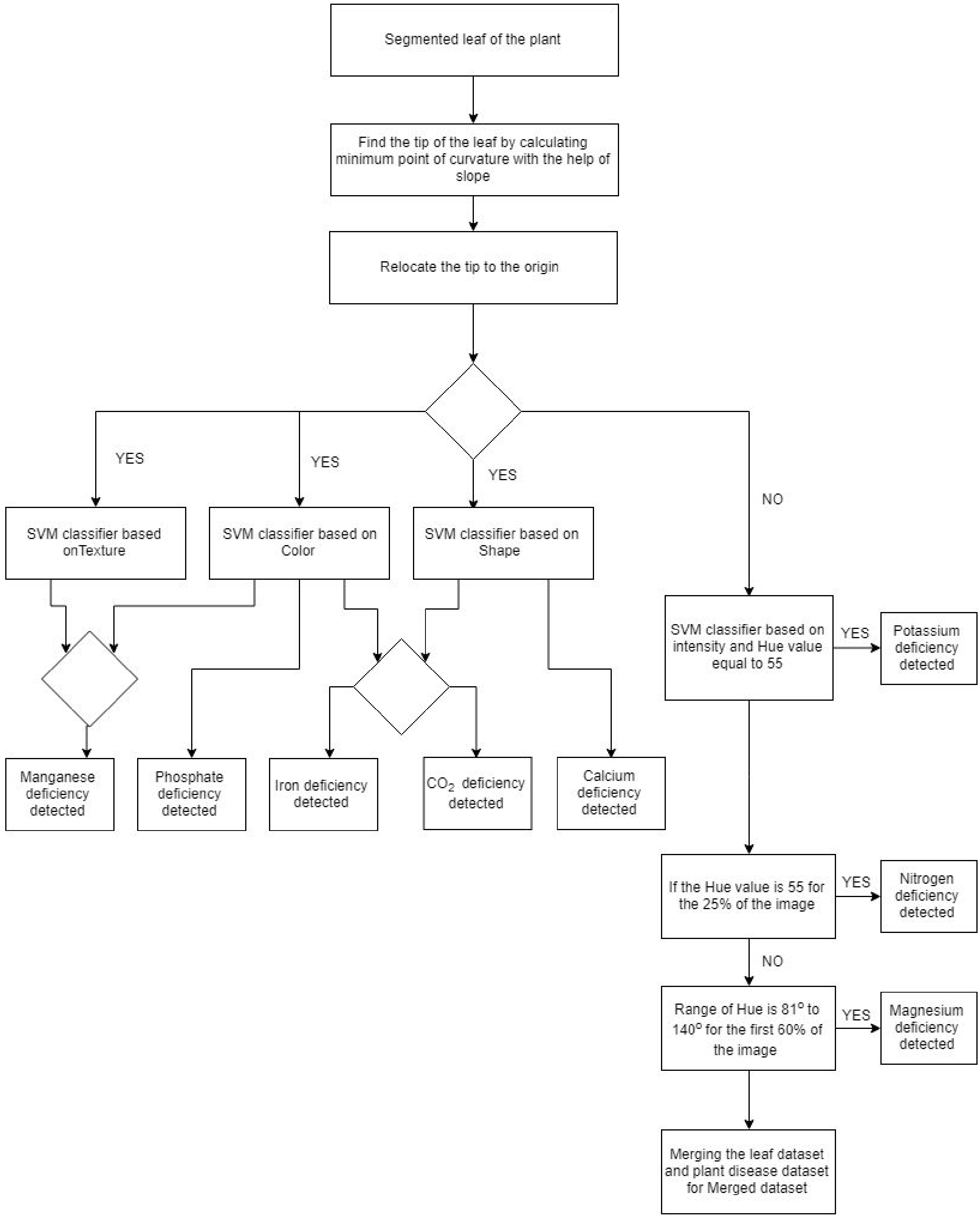
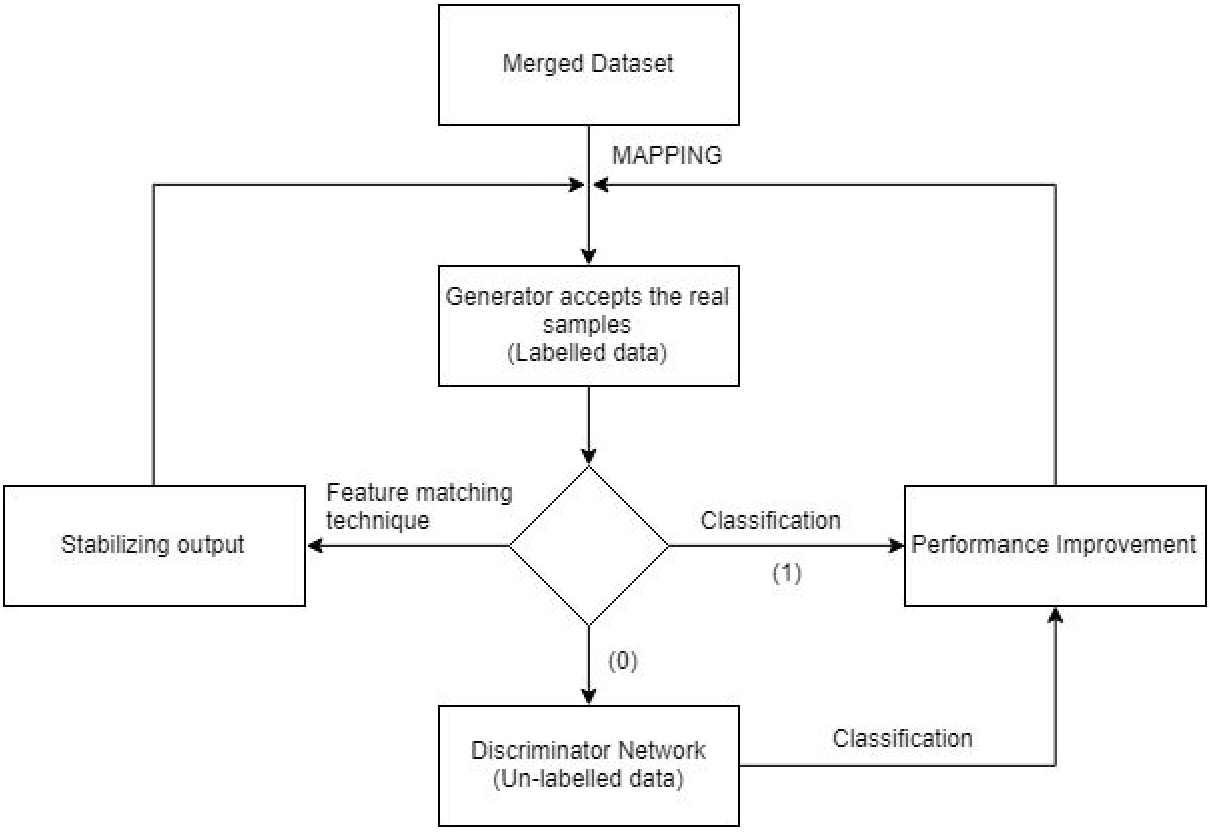


Figure 5 - Types of deficiencies in plants



Flowchart 4 - Disease identification

From flowchart 4, the disease identification with leaf and colour texture can be found out[7]. The Segmented leaf of the plant from figure 2 obtained and the point of minimum curvature is found out using the slope of the image using the expression $\frac{d_y}{d_x}$ where x represents the x-axis and y represents the y-axis[2]. Once, the slope of the curvature gets to the zero, then the minimum point of the curvature is found out. This point is the tip of the segmented leaf. Now, this point is set to the origin of the axis to place it in the centre of the viewpoint[8]. Now, the features of the deficiencies can be obtained through the SVM classifiers based on colour, texture, and shape. From recent studies, new plant diseases and their symptoms are being identified and updated on a regular basis. So, let's focus on the colour, texture and shape features on the leaf to find out the changes in the plant growth as the deficiencies leading up to various diseases and upcoming disasters to the farming land. Phosphate deficiency is detected based on the colour feature. The carbon-di-oxide deficiency in the plant is obtained through the features of shape and colour. There is white sediment formed on the surface of the leaf to detect the deficiency of the plant. The calcium deficiency is found using the shape classifier as smaller the size of the leaf, it has higher chances of spreading the infections to the other parts of the plants. The manganese deficiency is found through the colour and the texture features of the classifiers and the iron deficiency is found through the colour and the shape of the classifier. Apart from this, SVM classifier can also be categorized based on the intensity of the leaf. When the value of the hue is equal to 55, i.e. when there is the presence of yellow or the yellow-green colour in the texture of the image, the Potassium deficiency is detected in the dataset. If the value of the hue is 55 for the top 25 percent of the image, then nitrogen deficiency is found out. When the range of the hue is ranging from 81 to 140 degrees for the first 60 percent of the image, the magnesium deficiency is detected from the specimen of the image. Any other disturbances or changes in the texture of the leaf image should have taken place due to the insect attack on the plant species. Finally, the merging of the leaf dataset and the plant disease dataset produces a new merged dataset in this process.



Flowchart 5 - Generative Adversarial Network (GAN)

From the flowchart 5, The generative adversarial network plays a major role in merged dataset. From the merged dataset, the mapping process takes place to accept the real samples and place it in the generator which is nothing but the labelled data. This data is then rained for the performance improvement through the classification techniques. The dataset that is sent for the performance improvement is noted with the Boolean value 1 and the remaining dataset is noted with the Boolean value 0 to the network discriminator which is nothing but the unlabelled data which is further classified through the performance improvement system to add in more of the labelled data to the merged dataset. Meanwhile, the feature matching technique is used to stabilize the output and improve the feature selection techniques for the available labelled dataset. This process keeps on going as the classification works on both labelled and unlabelled dataset simultaneously. However, the number of unlabelled datasets is usually on the higher side.

Soil Test

Soil test has been one of the important features to examine the quality of soil used in the farming land. This is in fact, one of the underestimated factors that can destroy the farming land in the limited period as no such symptoms of soil degradation can be seen with the naked eye of the human beings. Oil spills that occur in the sea due to the leakage in the oil containers of the ship not only damage the aquatic life as oil spilled water can be transferred to the lakes or smaller water bodies which are close to the farming land. These polluted waters have higher chances of entering the farming land thus destroying the growth of the healthy crops[9]. The pores in the roots of the block that suck the water in the ground for the growth of the crops get blocked due to the higher density of the oil molecules. This blockage does not let the water enter the pores of the crop and the length of root under the soil gradually starts to degrade thereby, leading the loss of food crops. Another factor to test the quality of soil is to check the level of sodium content. There are higher chances that the water used in the farming land can have excess sodium contents in it that can degrade the nature of the soil[10]. The drawback of sodium is that it can suck or absorb large amounts of water thereby leading to the moisture loss in the crops in case of drought-prone areas of the country. Another symptom of soil degradation is that the soil loses its capacity to absorb the water on the land. Hence, the smartphone application must have access to convert the image of the soil to the thermal image to detect the oil contents in the land to find the symptoms of soil degradation on the onset. Water cannot be detected in RGB colour space whereas oil can be detected using RGB colour space due to the high density of the oil molecules. Orange colour indicates oil spills in crop land in the below images.

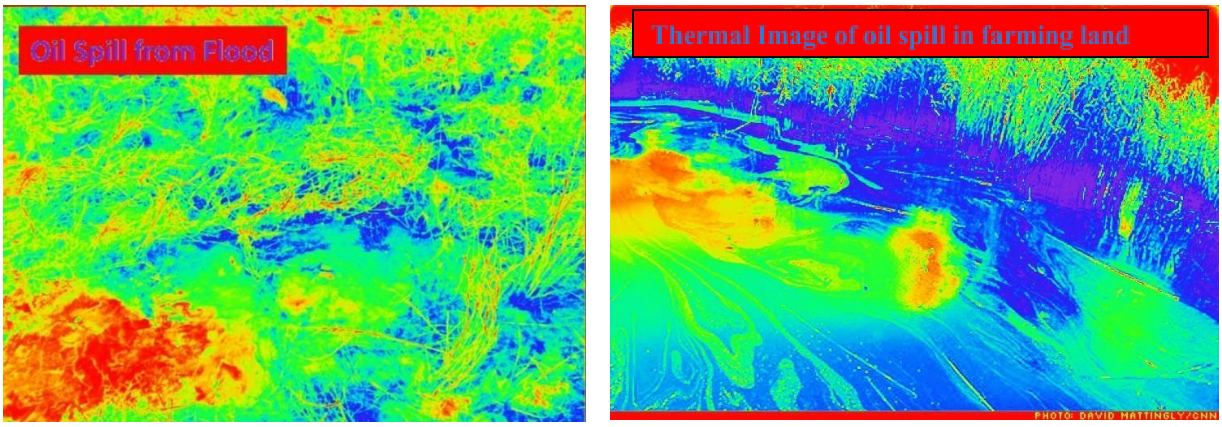


Figure 6 - Thermal image of oil spills

Conclusion

Crop protection is one of the important concerns for society. Usage of machine learning algorithms helps in the prediction of diseases that attack crops in various possible situations. Convolution Neural Networks (CNN) has been one of the important solutions for the reduced feature selection technique that filters out the necessary features using the set of filters. However, the deconvolution method helps in backpropagation algorithm where feature selection process takes place by creating many features through the layers. Another algorithm that works quite efficient for the feature identification is the Support Vector Machine (SVM) where the features of the image are extracted on the basis of intensity, colour, and texture of the image. Hence, a threshold is formed and values closer to it are formed as a feature. Many such classifications techniques are under research for SVM classification for producing results in a shorter period of time. Hence, the right mix of both the algorithms under the specific purpose helps in bringing a faster cloud performance for the farmers and the plant scientists to solve the new issues in this field. Apart from this, the machine learning community needs to work on quality of soil used in farming land to protect the soil quality from oil spills and save the crops for healthy food production.

References

1 Chapelle, O., Scholkopf, B., and Zien, A.: 'Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]', IEEE Transactions on Neural Networks, 2009, 20, (3), pp. 542-542

2 Lee, S.H., Chan, C.S., Wilkin, P., and Remagnino, P.: 'Deep-plant: Plant identification with convolutional neural networks', in Editor (Ed.)^(Eds.): 'Book Deep-plant: Plant identification with convolutional neural networks' (IEEE, 2015, edn.), pp. 452-456

3 Lomte, S., and Janwale, A.: 'Plant Leaves Image Segmentation Techniques: A Review', 2017

4 Pujari, D., Yakkundimath, R., and Byadgi, A.S.: 'SVM and ANN based classification of plant diseases using feature reduction technique', IJIMAI, 2016, 3, (7), pp. 6-14

5 Tóth, B.P., Tóth, M.J., Papp, D., and Szücs, G.: 'Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval', in Editor (Ed.)^(Eds.): 'Book Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval' (2016, edn.), pp. 569-578

6 Šulc, M., and Matas, J.: 'Fine-grained recognition of plants from images', Plant methods, 2017, 13, (1), pp. 115

7 Ma, J., Du, K., Zhang, L., Zheng, F., Chu, J., and Sun, Z.: 'A segmentation method for greenhouse vegetable foliar disease spots images using color information and region growing', Computers and Electronics in Agriculture, 2017, 142, pp. 110-117

8 Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., and Huang, T.: 'Large-scale image classification: fast feature extraction and svm training', in Editor (Ed.)^(Eds.): 'Book Large-scale image classification: fast feature extraction and svm training' (IEEE, 2011, edn.), pp. 1689-1696

9 Ofualagba, G.: 'Detection and Classification of Land Crude Oil Spills Using Color Segmentation and Texture Analysis', Journal of Imaging, 2017, 3, (4), pp. 47

10 Mira, D., Gil, P., Alacid, B., and Torres, F.: 'Oil Spill Detection using Segmentation based Approaches', in Editor (Ed.)^(Eds.): 'Book Oil Spill Detection using Segmentation based Approaches' (2017, edn.), pp. 442-447