1. INTRODUCTION

Image processing techniques can be applied to various types of images, including photographs, medical images, satellite images, and digital artwork. This project aims to use image processing techniques like adaptive thresholding, gaussian blur, morphological closing, noise reduction etc. on the dataset before feeding it to the Machine Learning Algorithms like SVM, and KNN.

Pre-processing allows us to eliminate unwanted distortions and improve specific qualities that are essential for the application we are working on. Those characteristics could change depending on the application. An image must be preprocessed in order for software to function correctly and produce the desired results.

Image processing filters can be of two types – Frequency Domain and Spatial Domain.

The Gaussian blur filter used here is a type of Frequency Domain low pass filter. Low-pass meaning it attenuates high frequency components in the image while allowing the low frequency components to pass through.

Morphological closing is a mathematical operation in image processing that is used to fill gaps and smooth out the boundaries of objects in an image. It is a dilation operation followed by an erosion operation. Morphological closing is commonly used to remove small holes and gaps in objects, to connect nearby objects that are almost touching, and to smooth out the boundaries of objects in an image. It can be applied to a variety of image types, including binary images (black and white), grayscale images, and color images.

Thresholding is a commonly used technique in image processing to create binary images from grayscale or color images. The main idea is to separate the foreground (object of interest) from the background by converting all pixel values above or below a certain threshold to a single value, usually white or black. There are different methods of thresholding, for example adaptive thresholding and Otsu's thresholding. In adaptive thresholding, different threshold values are calculated for different parts of the image based on the local characteristics of the image, such as variations in illumination. Otsu's thresholding is an automatic thresholding technique that determines the optimal threshold value by maximizing the variance between the two classes of pixels (foreground and background).

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."-Tom Mitchell.

Machine learning is concerned with creating models that can learn from data and make predictions or take actions based on that learning. Machine learning is categorized into four categories:

a. Supervised Learning
b. Unsupervised Learning
c. Semi-Supervised Learning
d. Reinforced Learning

Support Vector Machine (SVM) is a type of supervised machine learning algorithm. This means that it is an algorithm that requires labeled data to learn a mapping function from input variables (features) to output variables (labels or classes). SVM is a popular choice for supervised classification tasks due to its ability to handle complex data and its ability to work well with high-dimensional feature spaces. Its strength lies in its ability to find the optimal hyperplane that maximizes the margin between different classes of data points.

K-Nearest Neighbors (KNN) is a type of machine learning algorithm that also falls under the category of supervised learning. In the case of KNN, the algorithm predicts the class of a new data point based on the class of its nearest neighbors in the feature space. The value of k (the number of nearest neighbors to consider) is chosen by the user. It is popular due to its simplicity and effectiveness in certain scenarios since it is non-parametric, meaning it does not make assumptions about the underlying distribution of the data.

Convolutional Neural Networks (CNNs) are a specific type of machine learning algorithm used for image and video analysis, natural language processing, and other applications involving complex input data. CNNs are a type of neural network, a class of machine learning algorithms inspired by the structure and function of the human brain. Neural networks are designed to learn from data, and CNNs are specifically designed to learn from data that has a grid-like topology, such as images.

This project aims to use the aforementioned image processing techniques along with the machine learning algorithms to classify whether a plant is diseased or not. Some research works are done identifying plant disease but they have not covered the broader categories of plant diseases and image features for training purpose to obtain much accurate results for large set of images. The models for SVM, KNN and CNN will be developed and the results will be compared.

2. LITERATURE SURVEY

A computer vision algorithm designed to detect abnormalities in plants, specifically papaya leaves has been developed by the author. The algorithm first converts RGB images to grayscale to enable the calculation of shape descriptors and Haralick features. Then, to calculate the histogram, the image is converted to HSV. The algorithm distinguishes between healthy and diseased leaves using a Random Forest classifier [1]. The accuracy of the model achieved was 70%, and it can be improved by using additional local features. However, the accuracy of the models depended on the number and quality of images used for training. Therefore, further research is needed to improve the performance of these algorithms.

Various processes such as image capture, denoising, enhancement, segmentation, feature extraction, classification, with detection using different algorithms such as PSO SVM, BPNN, and random forest were used by the author [2]. With BPNN having the highest accuracy, which was 98.6%, followed by PSO SVM at 96.7% and Random Forrest at 88.2%. The review suggested that the PSO SVM algorithm produces better accuracy in classification and detection of grape leaf diseases compared to other algorithms.

A comparative study between a support vector machine and traditional method of neural networks has been done by the author. The study concluded that the SVM model had a far higher accuracy of 89.66% for the five patterns, compared to the neural networks' 41.38% [3]. The SVM model employs Principal Component Analysis to extract the eigenvalue from the gray information of the images to build a decision surface that classifies the images based on the Inner-Product Kernel. The SVM model proved to be more effective than the traditional neural network method, as demonstrated through extensive experiments. However, the study's limitations, such as the small sample size and the absence of comparative analysis with other machine learning models, should be considered.

A novel approach to plant species classification using convolutional neural networks (CNNs) by the author. The CNN architecture was designed to classify images of sixteen different types of plants. The results showed that the CNN-based approach outperforms the SVM-based classifier, and the accuracy achieved was 96.7%. The proposed method was compared to a traditional support vector machine (SVM) classifier that uses local binary pattern (LBP) and GIST features [4]. The experiments were conducted on experimental data acquired under natural outdoor illumination provided by TARBIL Agro-informatics Research Center of ITU. The authors suggested that future work should focus on building different architectures with various activation functions and experimenting with pre-processing methods to improve classification performance.

A framework for detecting and classifying tomato crop diseases using image processing techniques has been developed by the author[5]. The approach used images of plant leaves that exhibit visual symptoms of particular diseases, and extracted useful features from the images for disease classification. The proposed system had achieved a maximum average accuracy of 98.3% during experimentation. The approach only used texture features. The study highlighted the potential of image processing techniques for more effective and efficient

detection and classification of crop diseases, which can be of great benefit to farmers. However, more research was needed to explore the use of different features and approaches for better accuracy and reliability.

MNIST dataset of hand-written digits were classified using Convolutional Neural Networks(CNN) by the author. The CNN was trained on small grayscale images and accuracy achieved was 98% [6]. The author noted that processing such images can be computationally expensive, and suggested that future work could involve stacking more layers and training on larger datasets using clusters of GPUs to improve accuracy. The author also suggested that the CNN approach could be extended to classify larger, colored images for image segmentation purposes. Overall, the paper demonstrated the effectiveness of CNNs for image classification and suggested potential areas for further research and improvement.

Convolutional Neural Networks (CNNs) in image recognition, particularly for their reduced computational complexity and improved computing precision has been used by the author. The fault tolerance of CNNs allowed for the use of incomplete or fuzzy background images, thereby enhancing the precision of image recognition[7]. In the study, two feature extraction methods and three classifiers were compared in their abilities to identify seven tea leaf diseases. The results revealed that LeafNet yielded the highest accuracies with an average classification accuracy of 90.16%, while that of the SVM algorithm was 60.62% and that of the MLP algorithm was 70.77%.

S no.	Authors	Paper and Publication Details	Findings	Relevance to the project
1.	Maniyath, S. R., P V, V., M, N., R, P., N, P. B., N, S., & Hebbar, R.	Plant Disease Detection Using Machine Learning. International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C), 25 April 2018	Accuracy achieved using random forest classifier was70%. Accuracy can be improved by using additional local features	Usage of Haralick features. Converting RGB images to grayscale for pre-processing.
2.	Arshiya S. Ansari, Malik Jawarneh, Mahyudin Ritonga, Pragti Jamwal.	Improved Support Vector Machine and Image Processing Enabled Methodology for Detection and Classification of Grape Leaf Disease. Hindawi Journal of Food Quality, 9 July 2022	Highest accuracy was of BPNN 98.6%, SVM and Random Forest were at 96.7% and 88.2% respectively.	Use of the PSO SVM model.
3.	Xiaowu Sun, Lizhen Liu, Hanshi Wang, Wei Song, & Jingli Lu.	Image classification via support vector machine. 4th International Conference on Computer Science and Network Technology (ICCSNT), 19 December 2015	SVM had the highest accuracy of 89.6% and Neural Networks had 41.4%. Uses Principal Component Analysis for feature extraction.	Information regarding the kernel being used in SVM. Feature extraction methods used.
4.	Yalcin, H., & Razavi, S.	Plant classification using convolutional neural networks. 5th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 18 July 2016	CNN had the highest accuracy of 96.7%, compared to SVM. Varied and large dataset was used.	Ample information regarding the CNN based approach.
5.	Muhammad Zaka-Ud-Din, Wakeel Ahmad, Sumair Aziz.	Classification of Disease in Tomato Plants' Leaf Using Image Segmentation and SVM. Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan, 5 August 2018.	The SVM model used along with GLCM feature extraction had achieved a high accuracy of 98.6%.	Usage of SVM model along with pre-processing techniques such as RGB extraction and grayscale conversion.
6.	Muthukrishnan Ramprasath, M.Vijay Anand, Shanmugasundaram Hariharan.	Image Classification using Convolutional Neural Networks. International Journal of Pure and Applied Mathematics. Volume 119 No. 17 2018, 1307-1319.	The CNN model achieved an accuracy of 98%. The training was computationally expensive.	Usage of a basic CNN model, but with grayscale images.
7.	Jing Chen, Lingwang Gao	Visual Tea Leaf Disease Recognition Using a Convolutional Neural Network Model. Yearly journal, China Agricultural University, Beijing 100193, China.	LeafNet, a CNN based model achieved an accuracy of 90.1%. While that of SVM was 60.2%.	The fault tolerance level of CNN is quite high and it can work relatively well with incomplete or fuzzy background images.

Table no. 1: Literature survey

3. PROBLEM DEFINITION

Due to change in climate change in climatic conditions and certain environmental factors, plants acquire various diseases. Detection of these diseases in their early stages is and has been a very challenging task for farmers.
Given that they come in a broad variety of colors and sizes, identifying plant illnesses and diseases may be a very taxing task. Furthermore, some of the said infections only occur in particular families in certain specific conditions, hence making the acquisition of data in those cases extremely difficult. Being able to recognize a disease without the requirement of a botanist would be an extremely helpful tool for farmers and agriculturalists alike. Shape of the infection is the one of the most easily recognizable highlights in an infection but there are hundreds of different conditions with similar shapes. Even when the shape of the ailment is considered, the human eye can make mistakes and recognize it to be another one. Also, there may arise situations where an expert may not be available at the time of need, which can have varying consequences.

4. SOLUTION STRATERGY

The plant species include Corn, Wheat, Tea and Potato. The images collected are for various plant species and diseases. In total, we have 18 classes, with 1 healthy class and 2 unhealthy classes for each plant, with the only exception being Tea, which consists of 7 unhealthy classes. For Image pre-processing, the image is first converted from BGR color space to RGB color space .The RGB image is then converted to grayscale.

We then apply a Gaussian blur filter to the grayscale image with a kernel size of 25x25.

Otsu's thresholding technique is applied to the blurred grayscale image. This thresholding technique automatically calculates the optimal threshold value for the image based on its histogram. The resulting image is a binary image with white pixels indicating the object of interest and black pixels indicating the background. A flag inverts the binary image, so the object of interest is now black.

After that, we move onto with a simple which creates a 50x50 matrix of ones.
Next a morphological operation to the binary image using the kernel created in step 5. This particular operation is a closing operation, which fills in small holes and gaps in the object of interest and smooths its edges. The resulting image is stored in the variable *"closing"*.
The images are then converted into a array and stored in a "*csv" format*.

After performing all the operations mentioned above on the whole dataset, it is then fed to the machine learning algorithms SVM and KNN. CNN takes images as input. Since SVM cannot perform multiclass classification at once, we will have to one-versus-rest and Gaussian radial bias function.

5. DESIGN STRATERGY

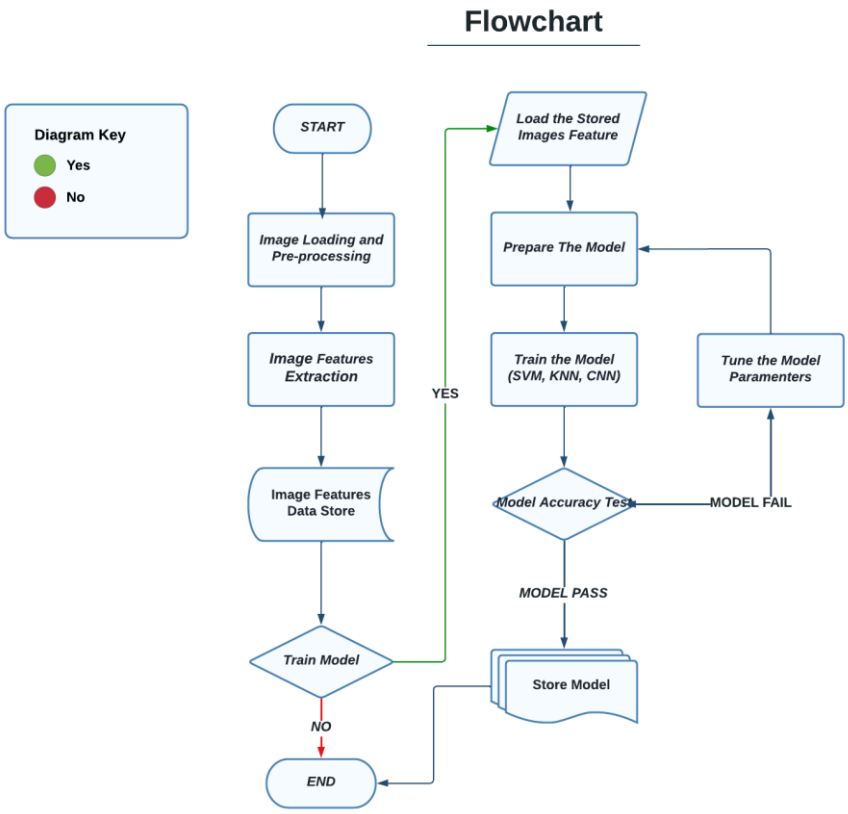

Fig1. Image feature extraction and model preparation

A. Image Acquisition

The images were collected for various plants species and diseases. The images were placed in jpg format. The source images were taken from the Kaggle plant village dataset, corn leaf disease dataset contributed by Samarnjit Ghose, wheat leaf dataset by Olyadgetch and identifying disease in tea leaves dataset by Shastwatwork. The train (folder containing images for training purpose of the models) and valid (folder containing images for validation purpose of the model) folder consists of images in ratio of 80 to 20 for training and testing purpose respectively for seventeen different categories which is shown in Table no. 2.

B. Image Feature Extraction

A total of 10 features are extracted, including color and textures. Firstly, the mean and standard deviation of each color channel (red, green, and blue) of the input image are calculated. Next, the image is converted to grayscale and a Gaussian blur filter is applied to reduce noise. Texture features are then extracted using the Haralick algorithm, which calculates contrast, correlation, inverse difference moments, and entropy using the gray-level

co-occurrence matrix (GLCM). The GLCM represents the relative frequencies of a pair of grey levels present at a certain distance and angle. The values of the texture features are then used to form a vector, which is appended to a data frame. The features can be used for further image analysis or machine learning tasks. The mathematical equations of the features used are given below (1)-(5):

$$G = \begin{bmatrix} P(1,1) & \cdots & P(1,Dg) \\ \vdots & \ddots & \vdots \\ P(Dg,1) & \cdots & P(Dg,Dg) \end{bmatrix} \quad \text{(1)[ResearchGate, 24/03/2023]}$$

Contrast = $\sum_{n-0}^{Dg-1} x^2 \{\sum_{i=1}^{Dg}\sum_{j=1}^{Dg} P(i,j)\},\ |i=j| = n$ (2)[ResearchGate, 24/03/23]

where Dg is numbers of gray levels that can be represented by a matrix G having dimension Dg as shown in Equation (1) with any pixel point (i,j) and P(i,j) represents the probability of presence of pixel pairs at certain distance d at angle θ in GLCM image.

Correlation $= \frac{\sum_{i=1}^{Dg}\sum_{j=1}^{Dg}(i,j)P(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$ (3) [Source: SciElo, 24/03/2023]

where μx μy are means and σx σy are standard deviations of Px and Py the partial derivative function.

Inverse Difference Moments = $\sum_{i=1}^{Dg}\sum_{j=1}^{Dg}\frac{1}{1+(1-j)^2}P(i,j)$ (4) [Source: IJSTE, 24/03/2023]

Entropy = - $\sum_{i=1}^{Dg}\sum_{j=1}^{Dg} P(i,j)\log[P(i,j)]$ (5) [Source: SciElo, 24/03/2023]

C. Model Training and Testing

There exist several techniques for solving Multi-class classification problems using Support Vector Machine (SVM), such as One Against-One (OAO), One-Against-All (OAA), Binary Tree (BT), and Directed Acyclic Graph (DAG) classifiers. The primary goal of SVM is to construct an optimal hyperplane that acts as a decision surface using input samples to maximize the margin between two sides. To perform multi-class classification with SVM, the one-versus-rest method and Gaussian radial basis function will be utilized. The RBF kernel is a positive parameter used to regulate the radius and is given by Equation (6) . Since SVM cannot perform multi-class classification simultaneously, it uses the one-versus-rest method to perform binary operations on each dataset before making the final multi-class classification.

$$K(xi,xj) = \exp\left(\frac{-||xi-xj||^2}{2\sigma^2}\right) \quad \text{(6) [Source: DataFlair, 24/03/2023]}$$

where k is the kernel function , xi = (xi1, xi2, . . . , xiN) corresponds to the attribute set for the ith sample in each sample tuple represented by (xi, xj) in N training data of a binary classification.

K Nearest Neighbor (KNN) is a versatile technique that can be used for both classification and regression tasks. It is a non-parametric algorithm that is widely used for pattern recognition, where it selects the k nearest neighbors for classification or regression purposes. KNN is a classification method that determines the class of a given data point by looking at the K nearest neighbors and selecting the most frequent class based on the similarity of those neighbors, which is computed using distance metrics. Some of the common distance functions are Euclidean, Manhattan, Minkowski and Hamming distance.

In this study, Convolutional Neural Network (CNN) was chosen as the preferred deep learning method. CNN has the ability to identify and classify objects with minimal pre-processing, making it successful in analyzing visual images and easily extracting necessary features with its multi-layered structure. The key layers in CNN include convolutional layer, pooling layer, activation function layer, and fully connected layer. The scikit-learn Python library will be utilized.

6. WORK DONE

S. no	Plant Categories as sub folders in root folders Train and Test	Images count in respective sub folders of root folder (Train)	Images count in respective sub folders of root folder. (Test)
1.	Corn_Blight	200	50
2.	Corn_Gray_Leaf_Spot	80	20
3.	Corn_Healthy	280	67
4.	Potato_Early_blight	88	22
5.	Potato_Late_blight	88	22
6.	Potato_Healthy	90	22
7.	Tea_algal_leaf	90	22
9.	Tea_Anthracnose	80	20
10.	Tea_brown_blight	80	20
11.	Tea_gray_blight	80	20
12.	Tea_brown_blight	90	22
13.	Tea_red_leaf_spot	95	23
14.	Tea_white_spot	112	28
15.	Tea_Healthy	74	30
16.	Wheat_septoria	77	19
17.	Wheat_stripe_rust	129	32
18.	Wheat_Healthy	80	22

Table no. 2: Dataset Description

The dataset used in this study was sourced from multiple sources as specified earlier. After importing the dataset into the Jupyter notebook, it was partitioned into two parts: one for training and the other for testing, with a ratio of 80:20 respectively. The 80% portion was designated for training purposes, while the remaining 20% was reserved for testing.

Before the training process commenced, the data in the training partition was subjected to a normalization process. Normalization refers to the process of rescaling the features of a dataset to a standard scale. This ensures that all features have an equal contribution during the training process. The normalization process that was employed in this study involved transforming the features into z-scores. This means that the features were scaled in such a way that their mean was equal to zero, and their standard deviation was equal to one.

Subsequently, the normalized training dataset was subjected to several preprocessing algorithms to extract various features. These algorithms involved the extraction of color features such as the red, green, and blue values of the images, and the calculation of their standard deviation. Texture features such as contrast, correlation, inverse difference moments, and entropy were also extracted. These features were computed to help the model differentiate between the different classes in the dataset.

After feature extraction, the resulting feature values were stored in CSV format for easy access and manipulation.

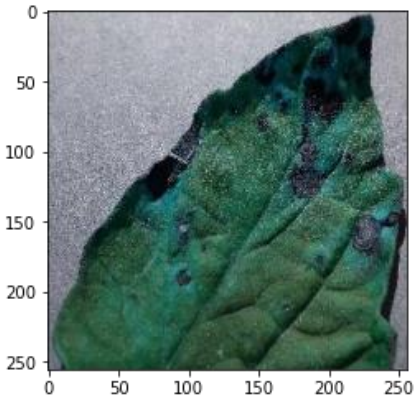

Fig 2. Original Image

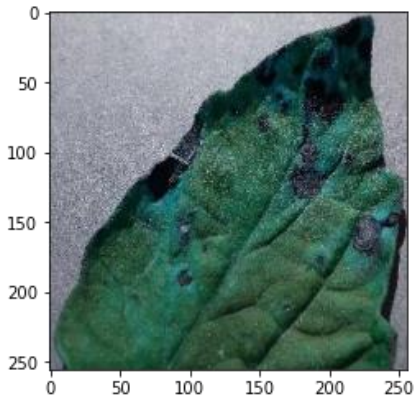

Fig 3. Gaussian Blur

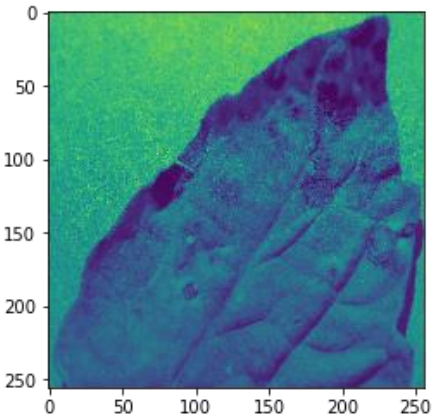

Fig 4. Grayscale

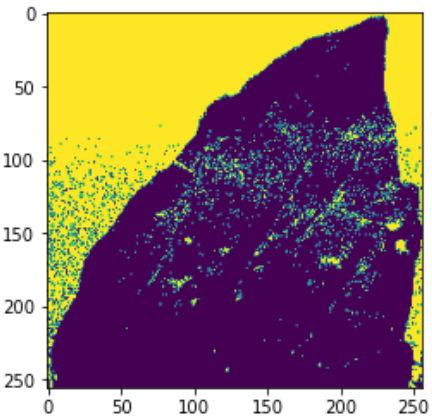

Fig 5. Otsu's Thresholding

A. Challenges Faced

During the course of working on a project, we often come across several challenges that need to be addressed to ensure the successful completion of the project. In our case, we faced a number of challenges related to data acquisition, data cleaning, handling outliers, and selecting the appropriate preprocessing algorithms for our machine learning model.

One of the main challenges we encountered during the data acquisition process was the lack of reliable and trustworthy sources of data on the internet. We had to rely on a few sources, including Kaggle, to obtain the data we needed for our project. However, even with these sources, we found that two of the datasets we acquired, namely the tea leaf and wheat leaf datasets, were not up to the standard we required. For instance, the tea leaf dataset was completely unlabeled, and we had to use Bulk Renaming Utility(4) to rename each image. On the other hand, the wheat leaf dataset had images with poor quality that required pre-processing to improve their quality.

Another challenge we faced was the presence of outliers in the dataset. While most of the outliers were found in the tea and wheat leaf datasets, we also found some outliers in other classes of plants. We had to go through each image carefully and remove the outliers by either normalizing the image or deleting it entirely.

Deciding on which preprocessing algorithm to use was another significant challenge we faced. We had to carefully review various research papers and articles to understand the different preprocessing algorithms used for training and testing machine learning models. We spent a significant amount of time researching and comparing different algorithms before deciding on the best method to use for our dataset.

In summary, we faced several challenges during the development of our machine learning model. We encountered issues with data acquisition, cleaning, handling outliers, and selecting appropriate preprocessing algorithms. However, by carefully addressing each of these challenges and utilizing our research skills, we were able to overcome these obstacles and proceed further with the project.

7. GANTT CHART

ACTIVITY	TIME FRAME			
	January 2023	February 2023	March 2023	April 2023
Literature Survey				
Problem Definition				
Identification and Preprocessing of Dataset				
Development of Model				
Testing and Validation				
Documentation				

Proposed Activity | Achieved Activity | Ongoing Activity

Figure no 3: Gantt chart for Plant disease recognition using convolutional neural networks

8. REFERENCES

[1] Maniyath, S. R., P V, V., M, N., R, P., N, P. B., N, S., & Hebbar, R. (2018). Plant Disease Detection Using Machine Learning. 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C).

[2] Arshiya S. Ansari, Malik Jawarneh, Mahyudin Ritonga, Pragti Jamwal. Improved Support Vector Machine and Image Processing Enabled Methodology for Detection and Classification of Grape Leaf Disease. Hindawi Journal of Food Quality.

[3] Xiaowu Sun, Lizhen Liu, Hanshi Wang, Wei Song, & Jingli Lu. (2015). Image classification via support vector machine. 2015 4th International Conference on Computer Science and Network Technology (ICCSNT).

[4] Yalcin, H., & Razavi, S. (2016). Plant classification using convolutional neural networks. 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics).

[5] Muhammad Zaka-Ud-Din, Wakeel Ahmad, Sumair Aziz. Classification of Disease in Tomato Plants' Leaf Using Image Segmentation and SVM. Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan.

[6] Muthukrishnan Ramprasath, M.Vijay Anand, Shanmugasundaram Hariharan. Image Classification using Convolutional Neural Networks. International Journal of Pure and Applied Mathematics. Volume 119 No. 17 2018, 1307-1319.

[7] Jing Chen, Lingwang Gao .Visual Tea Leaf Disease Recognition Using a Convolutional Neural Network Model. Yearly journal, College of Plant Protection, China Agricultural University, Beijing 100193, China.

Equations Referenced

(1) https://www.researchgate.net/publication/342149372

(2) https://www.researchgate.net/publication/342149372

(3) https://www.scielo.br/j/aabc/a/VCt3WLtyDNCVbg9S3nXzSnt/?lang=en

(4) https://www.scielo.br/j/aabc/a/VCt3WLtyDNCVbg9S3nXzSnt/?lang=en

(5) http://www.ijste.org/articles/IJSTEV2I11015.pdf

(6) https://data-flair.training/blogs/svm-kernel-functions

Websites Referred

1. https://www.kaggle.com/datasets/olyadgetch/wheat-leaf-dataset

2. https://www.kaggle.com/datasets/soumiknafiul/plantvillage-dataset-labeled

3. https://cropwatch.unl.edu/soybean-management/plant-disease

4. https://www.bulkrenameutility.co.uk/