A Project report on

IMAGE BASED DETECTION OF PLANT DISEASE USING MACHINE LEARNING & DEEP LEARNING WITH CNN

A report submitted in partial fulfillment of the Academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted By

Aravelli Abhinav (20H51A0505)

Mood Guru Sai Chawan (20H51A0517)

G.Naveen (20H51A0564)

Under the esteemed Guidance of

Mr.Saidulu

Assistant Professor

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution, Approved by AICTE, Affiliated to JNTUH, NAAC 'A+')

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD-501401.

2020-2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD-501401.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled "**IMAGE BASED DETECTION OF PLANT DISEASE USING MACHINE LEARNING & DEEP LEARNING WITH CNN"** is a bonafide work done by **Aravelli Abhinav(20H51A0505) , Mood Guru Sai Chawan(20H51A0517) , G.Naveen (20H51A0564)** in partial fulfillment for the award of **Bachelor of Technology** in **Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision

The results embodies in this project report have not been submit to any other University or Institute for the award of any Degree.

Mr.G.Saidulu
Assistant Professor
Dept.of CSE

Dr. S. Siva Skandha
Associate Professor , HOD
Dept.of CSE

ACKNOWLEDGEMENT

With great pleasure I want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr.Golla Saidulu** Assistant Professor, Dept of Computer Science and Engineering for his valuable suggestions and guidance during the execution of this project work.

We would like to thank **Dr. S.Siva Skandha**, Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academic, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana**, Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation

Finally I express my sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care. I sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

NAME	**ROLL NUMBER**
ARAVELLI ABHINAV	(20H51A0505)
MOOD GURU SAI CHAWAN	(20H51A0517)
G.NAVEEN	(20H51A0564)

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Agricultural productivity is something on which economy highly depends. This is the one of the reasons that disease detection in plants plays an important role in agriculture field, as having disease in plants are quite natural. If proper care is not taken in this area then it causes serious effects on plants and due to which respective product quality, quantity or productivity is affected. Detection of plant disease through some automatic technique is beneficial as it reduces a large work of monitoring in big farms of crops, and at very early stage itself it detects the symptoms of diseases i.e., when they appear on plant leaves. This project presents the most recent results in this field, and a comparison of deep learning approach with the classical machine learning algorithms. It also covers survey on different diseases classification techniques that can be used for plant leaf disease detection.

Human population steadily continues to grow, and along with it the need for food production increases. According to the UN projections, human population is expected to reach 9.7 billion in 2050, 2 billion more than today. Considering that most of the population growth is to occur in the least developed countries (around 80% increase in the next 30 years), where the food scarcity is the main problem, it is easy to conclude that minimizing food loss in those countries is a primary concern. It is estimated that the yield loss worldwide is between 20 and 40 percent, with many farms suffering a total loss. Easily spreadable diseases can have a strong negative impact on plant yields and even destroy whole crops. That is why early disease diagnosis and prevention are of very high importance.

CHAPTER – 1

INTRODUCTION

1.1 INTRODUCTION

The use of technology in the detection and analysis process increases the accuracy and reliability of these processes. For example, the people who use the latest technology to analyze the diseases that arise unexpectedly are at a higher chance of controlling them than those that do not. In the recent occurrence of coronavirus, the world relied on the latest technology to develop preventive measures that have helped reduce the rate at which the disease is transmitted. Crop diseases are a significant threat to human existence because they are likely to lead to droughts and famines. They also cause substantial losses in cases where farming is done for commercial purposes. The use of computer vision (CV) and machine learning (ML) could improve the detection and fighting of diseases. Computer vision is a form of artificial intelligence (AI) that involves using computers to understand and identify objects. It is primarily applied in testing drivers, parking, and driving of self-driven vehicles and now in medical processes to detect and analyze objects. Computer vision helps increase the accuracy of disease protection in plants, making it easy to have food security.

One of the areas that CV has helped most is the detection of the severity of the diseases. Deep learning (DL), a part of the CV, is useful and promising in determining the severity of diseases in plants and animals. It is also used to classify diseases and avoid the late detection of diseases. Plant diseases are slightly different from those that affect human beings. Many factors make diseases similar as well. However, the diseases that can be transmitted from humans to plants and vice versa are rare. The analysis of the data related to this field helps identify how the use of the latest technology can be improved. The images of leaves and other parts of the plants can be used to detect diseases in plants. The technology could be applied in analyzing images in human beings that also prove the presence of diseases and determine the extent of their destruction. This project study is aimed at analyzing the way image based technology can be used in detecting diseases in plants.

1.2 OVERVIEW

ML is the technology that allows machines to communicate with human beings and understand their needs. It also makes machines act like human beings and make the decision on behalf of humans. It is one of the areas that have grown fast over the past few years. ML helps in classifying plant diseases. The use of this technology is seen as a significant beginning and achievement in dealing with plant diseases. It has also increased productivity in the field of cultivation. Visualization techniques have also been included in this technology, and it has been improved over the last three years to the current improved levels. The challenges that face the world today, related to the diseases affecting plants can be reduced if the diseases are identified before they spread to vast areas. The use of ML is widespread in the world today. Diverse methods used in ML and DL help the experts to analyze the plant diseases and know their source in time. The detection of these diseases is affected mainly by several challenges that affect the effectiveness and accuracy of this technology.

The first challenge is the time complexity associated with the use of ML and DL, whereby some of the technologies used in the detection of these diseases are outdated or based on some information from the past. The other challenge is segmentation sensitivity. It means that the region of interest (RoI) requires a high level of accuracy and sensitivity to acquire the required usage and accuracy. The other challenge is that there is a language barrier that affects the way the technology is applied. Another challenge is the inadequate resources that are required to support the application of this technology. Most of the ML and DL activities need many resources to use and implement. Private and government entities usually fund the institutes that use this technology to detect diseases in plants, which could affect the success of the research and implementation of the technology.

The importance of plants in the world has increased over time. The discoveries about the critical roles that plants could play in medicine, energy production, and the recent concerns about the reduction of global warming have for long been a significant part of science and technology. A reduction in the plant cover in the world increases the risk of higher global warming and an increase in the related challenges. The need to build a state-ofthe-art convolutional system that supports the image detection technology and classification of plant diseases has led to many research programs to provide the scientists with the required knowledge.

Image detection could be applied when necessary to differentiate healthy leaves from those that are not healthy. The convolutional neural networks (CNNs) provide the differences among plant images that help determine the abnormalities that could exist in the plants in the natural environment. The background study shows that the scanning of the images that show the healthy and unhealthy plants forms a basis for comparison by the scientists in this field. DL can be used to detect abnormalities in plants. The pixel wise operations are used to analyze the leaves collected from sick plants, and this is used to classify the diseases according to their impact on the plants. The visible patterns in these leaves are used to decide the diseases that affect the plants and how they can be dealt with to prevent them from spreading. Research shows that the use of DL technology is up to 98.59% accurate. The field of plant pathology has contributed immensely to the control of diseases and reduced global warming. One of the essential background knowledge that guides the use of image detection technology is that the leaves of the infected plants are different from the healthy ones. The leaves are likely to have dark parts, and some may be dry along the edges. The dried parts are also likely to fold up, and this is easy to detect even with a bare eye. The use of ML is to detect these differences without human intervention.

CHAPTER-2

BACKGROUND WORK

Fungi usually cause diseases that affect the plants, and they typically attack the leaves. Viral and bacterial pathogens cause many others. Precision in agriculture has improved with the increased use of ML and its related features. The reduced production quantity in agriculture hurts many people and animals, which requires modern technology to solve. The extraction and detection of diseases are easier when the image-based detection system is used because of its high accuracy and reduced complications and duplication of data. In some plants like tomatoes, the use of the images to determine the diseases that affect them and the extent of the damages cannot be achieved unless there is a high accuracy rate. The survey on plant diseases shows that many diverse factors determine how technology-based image detection is applied. In other words, the diseases that cause visible dents and changes on the plants are the ones that can be detected using this technology as opposed to the ones that cause damages that cannot be detected from the plants images. The analysis in this research shows that plant diseases are usually detected when they start showing an impact on the physical appearance of the plants.

The main challenge affecting the field of agriculture is the reduction in production and poor-quality production in plants. The challenge is a result of the poor detection and management of the diseases that affect the plants. The challenge is also extended to affect human beings in several ways. The reduced plant cover due to plant diseases means that global warming, famine, and reduced air purification ensue. Hyper spectral imaging has become a reliable way of detecting crop diseases on time. It is hard to determine the factors that lead to the diseases unless they are detected on time. In other words, if a disease is detected on time, it is easy to relate it to the possible factors that lead to its occurrence. For example, scientists could determine if there was a change in weather or climate that could have led to the occurrence of the disease.

Further research shows an inadequate database that could be used to provide background knowledge for comparing the images taken. The other challenge is that the symptoms and characteristics of the diseases are diverse and could be similar to a certain degree. For example, many diseases could lead to the wilting of leaves. The challenge is yet to be resolved because more and new images are uploaded progressively by experts. The other challenge is the lack of suitable instruments for use in the work of image detection. Most of the experts in the field do not have the equipment they require to analyze the images they get from the field, and this makes it hard for them to acquire accurate data and identify the diseases. The other one is that there is a low rate of implementation in some areas due to the regulations put in place to ensure the credibility and reliability of the data from these analyses. For example, after the 4th and 6th International Conference on Machine Learning and Soft Computing, there have been many regulations that may derail the use of ML in some parts. The rules discourage some of the results from the ML functions from being applied in practice because they do not meet the required parameters.

The technology has been in existence for several years now. However, there are still many issues that have not been clarified about its application. The other challenge is related to this fact. Some of the important images that could help determine if disease exists have not been captured. The other one is that the future perspectives of the research are not clear, and this is because of the increased diversity in the diseases that affect plants. The application of image-based detection is also affected by the increased diversity in the way the diseases appear. Some of the diseases that used to affect the plants a few years ago have evolved into new forms, and they have different impacts and outcomes. It is difficult for the images to be used alone to conclude the dis- eases and choose a solution. Some of the solutions used in the past have also become ineffective, reducing the effectiveness of the technology.

2.1 The global burden of pathogens and pests on major food crops

Crop pathogens and pests reduce the yield and quality of agricultural production. They cause substantial economic losses and reduce food security at household, national and global levels. Quantitative, standardized information on crop losses is difficult to compile and compare across crops, agro ecosystems and regions. Here, we report on an expert-based assessment of crop health, and provide numerical estimates of yield losses on an individual pathogen and pest basis for five major crops globally and in food security

hotspots. Our results document losses associated with 137 pathogens and pests associated with wheat, rice, maize, potato and soybean worldwide. Our yield loss (range) estimates at a global level and per hotspot for wheat (21.5% (10.1–28.1%)), rice (30.0% (24.6–40.9%)), maize (22.5% (19.5–41.1%)), potato (17.2% (8.1–21.0%)) and soybean (21.4% (11.0–32.4%)) suggest that the highest losses are associated with food-deficit regions with fast-growing populations, and frequently with emerging or re-emerging pests and diseases. Our assessment highlights differences in impacts among crop pathogens and pests and among food security hotspots. This analysis contributes critical information to prioritize crop health management to improve the sustainability of agro ecosystems in delivering services to societies.

2.2 Using deep learning for image-based plant disease detection

Crop diseases are a major threat to food security, but their rapid identification remains difficult in many parts of the world due to the lack of the necessary infrastructure. The combination of increasing global smartphone penetration and recent advances in computer vision made possible by deep learning has paved the way for smartphone-assisted disease diagnosis. Using a public dataset of 54,306 images of diseased and healthy plant leaves collected under controlled conditions, we train a deep convolutional neural network to identify 14 crop species and 26 diseases (or absence thereof). The trained model achieves an accuracy of 99.35% on a held-out test set, demonstrating the feasibility of this approach. Overall, the approach of training deep learning models on increasingly large and publicly available image datasets presents a clear path toward smartphone-assisted crop disease diagnosis on a massive global scale.

2.3 Textural features for image classification

Texture is one of the important characteristics used in identifying objects or regions of interest in an image, whether the image be a photomicrograph, an aerial photograph, or a satellite image. This paper describes some easily computable textural features based on gray tone spatial dependencies, and illustrates their application in category identification tasks of three different kinds of image data: photomicrographs of five kinds of sandstones, 1:20 000 panchromatic aerial photographs of eight land-use categories, and Earth Resources Technology Satellite (ERTS) multi special imagery containing seven land-use categories. We

use two kinds of decision rules: one for which the decision regions are convex poly-hydra (a piecewise linear decision rule), and one for which the decision regions are rectangular parallel-piped (a min-max decision rule). In each experiment the data set was divided into two parts, a training set and a test set. Test set identification accuracy is 89 percent for the photomicrographs, 82 percent for the aerial photographic imagery, and 83 percent for the satellite imagery. These results indicate that the easily computable textural features probably have a general applicability for a wide variety of image-classification applications.

2.4 Support-vector networks

The support-vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data. High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

CHAPTER – 3

PROPOSED SYSTEM

3.1 PLANT DISEASE IMAGE DATASET

The data sets used in the research include the descriptions of the leaves before and after the diseases affect them. The data comprises tables and images of the leaves that are taken in the fields. The data is analyzed and classified in a way that is easy for the readers to understand. For example, fig 4.6 it show the leaves used to determine the soybean plants affected by the diseases. The data set shows healthy leaves and the ones that had dents due to the attack by septorial leaf blight, others by frogeye leaf spot, and those affected by downy mildew as shown in Figure 4.6.1. The images in Figure 4.6.1 show that there were visible differences between the leaves affected by the disease and those that had not. The data set was clear and easy to understand. Another form of data was the table that showed the number of leaves that were classified under each disease. The data set is clear and indicates the total number of leaves analyzed and classified into four categories. The other set of data is by which shows the graphical representation of the captured information in the tables.

The data set used in the project can be shown in tables, texts, graphs, and other forms. However, an essential aspect of all of them is the ease of analysis and ease of understanding. Some of the data are also grouped as per the required levels. For example, data can be grouped in terms of the dis- eases they represent, the time they were collected, or the analysis method. The data captured in some of the research outcomes also show the use of technology and its effective- ness. For instance, the data sets captured using technology allow for a controlled environment; the data sets show the type of control used and its effectiveness. For example, computer-aided diagnosis (CAD) systems were captured in the data analyzed .The data sets are provided to help understand the usage of this technology and its impact on the quality of the research. The same data shows the classification techniques used and gives reasons for the choice.

The other feature is that the data sets for leaves analysis are based on the primary data collected in the fields. The reliability of the data is high because it is based on the observable features of the leaves. The data sets are also divided into sections that are easy to understand. For example, the work of shows the divisions of the work in terms of the diseases like Rice Blast (RB), Bacterial leaf Blight (BLB), and Sheath Blight (SB).

The use of a “PlantVillage” data set was also applied in this project. The data set consists of 54,306 images of 14 different crops representing 26 plant diseases. The images that were included in the data set included leaves having different colors. Figure 4.6.2 shows some samples of the PlantVillage data set. The colors indicate the parts of the leaves affected by the diseases that were under investigation. The authors also used the augmented data set proposed by Geetharamani and Pandian. The Image Net data set was also used in the research, and this led to high-quality research outcomes because of the synergy of combining various methods.

The data sets used in the research studies are dependent on the type of information contained. For example, the project focused on the effects of uncontrolled pests in China and the impacts on the total food produced. It shows that poor control of pests in China leads to a loss of about 30% of the total foods produced. The data sets in the research are large and show the different production levels and how the pests affect them. The use of a public data set also made it possible to understand the different ways in which the research was done and verified using data that are available in the public domain. The Plant Village data set was also used where 14 different types of leaves of cucumber plant are analyzed for seven different diseases. The data sets were mainly combined to provide a good presentation of the data that was collected.

(a) **(b)**

(c)

(d)

(e)

Fig 3.1: Data set used for classification

The use of the PlantVillage data set was used to show how the collected information was helpful in understanding plant diseases. The data sets used in most references were from the data collected by different researchers and combined in one set. The reliability of the project could be compromised if there is no control over what is to be included in the data sets. The nature of the research also determines how the data sets are used. For example, the collection of leaves and combining them in a table usually involves using the PlantVillage data set.

The use of coffee leaf data sets in the project is backed by the need to show the diseases that affect the coffee leaves and tomatoes and how they can be detected using image-based detection methods. In the project we used an information-rich color data set and large numerical data sets to display the data collected. The machine learning process also employs the training data set to predict and analyze unseen data. The data sets also include the expected data, making it easy to rely on the data in the research and

determine if the data is valid. The use of data sets related to the nature of the research helps achieve the goals of the data collection and analysis. The aforementioned data sets are affected by some challenges and limitations that reduce their applicability.

Another well-known data set used by the research com- munity is Northern Leaf Blight (NLB), which contains infected maize leaves. Some sample images from the NLB data set are shown in Figure 3. NLB consists of 1787 images having 7669 lesions. The images were obtained from maize plants in the field while using a handheld camera. The images in NLB were captured in uncontrolled conditions as opposed to the PlantVillage data set.

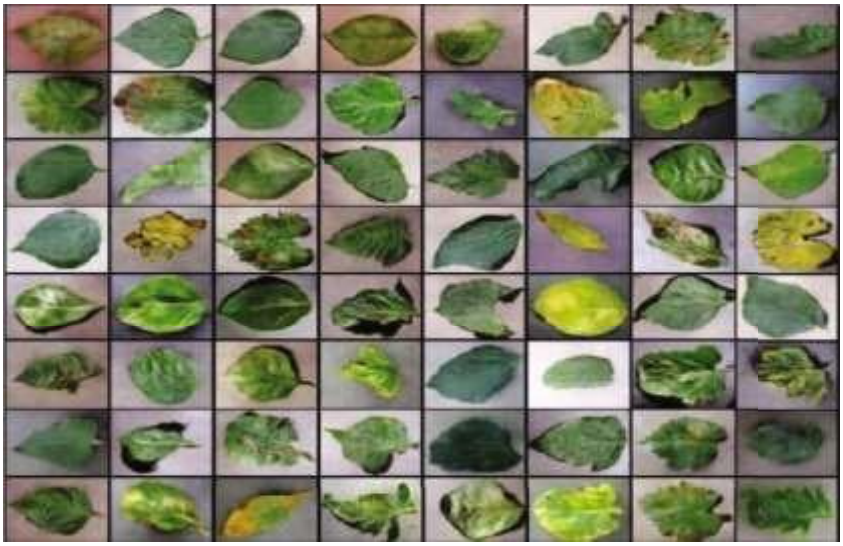

Fig 3.2: PlantVillage data set: examples of different phenotypes of various plants

3.1.1FEATURE EXTRACTION FOR DISEASE IDENTIFICATION

The images of the plants have three key features, namely, color, shape, and texture. Compared to color and texture, the shape feature cannot help find the plant's diseases. Hlaing and Zaw classified tomato plant disease using a combination of texture and color features. They used the Scale Invariant Feature Transform (SIFT) to find the texture information, containing details about the shape, location, and scale. Similarly, they gathered the color details from the RGB channel.

Dandawate and Kokare developed an approach for the automatic detection of diseases in soybean plants. They converted the image from RGB to HSV (Hue Saturation Value) color space. Color and cluster-

based methods were employed for segmentation. The SIFT method was used to detect the type of plant, based on the shape of the leaf.

Pydipati et al identified the citrus disease using color texture features along with discriminant analysis. They also employed the color co-occurrence method (CCM) to determine if hue, saturation, and intensity (HSI) color features and the statistical classification algorithms could help identify the diseased leaves. Their method achieved an accuracy of more than 0.95.Al-bayati and Üstündağ extracted only the area of the leaf affected by the disease. Furthermore, they used feature fusion which helped in feature reduction.

Image-based detection requires many resources, and the authorities should ensure they are available so that the activities are smooth. DL in general and CNN in particular have been developed to analyze multidimensional data such as images. The underlying model is based on the multilayer ANN. Nevertheless, a convolutional layer performs kernel operations over various areas of the provided image. The obtained representation is independent of the operations such as translation or rotation. These kinds of features have been proved to work better as compared to the traditional features earlier used in the detection of plant diseases**.**

3.2 REQUIREMENTS:

The main purpose of our project "**IMAGE BASED DETECTION OF PLANT DISEASE USING MACHINE LEARNING & DEEP LEARNING WITH CNN**" is to predict whether the plant is healthy or diseased.

3.2.1 FUNCTIONAL REQUIREMENTS

Graphical User interface with the User.

3.2.3 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements describe aspects of the system that are not directly related to the functional behaviour of the system.

Security: This system provides a login for all the users to access the software. So the chances of the software getting intruded are very less.

Reliability: Reliability is the ability of a system component to perform its required functions under stated conditions for a specified period of time.

Performance: Ease with which the software is doing the work it is supposed to do.

Usability: Ease with which the software can be used by specified users to achieve specified goals.

3.2.4 SOFTWARE REQUIREMENTS

Programming Language	: Python
Database	: kaggle.com
Workbench	: Jupyter Note Book, Visual studio code
Packages	: Django, Python required packages
Operating System	: Windows or Linux
Designing	: HTML, CSS, Java Script

3.2.5 HARDWARE REQUIREMENTS

Hard Disk	:	20GB
Floppy Driver	:	1.44 Mb
Processor	:	Intel core i5
RAM	:	1GB
Mouse	:	Optical Mouse
Monitor	:	14’ Colour Monitor

3.3 SYSTEM IMPLEMENTATION:

The purpose of system implementation can be summarized as follows:

Making the new system available to the prepared set users (the deployment), and positioning on-going support and maintenance of the system within the performing organization. Transitioning the system support responsibilities involves changing from a system development to the system and maintenance mode of operation, with ownership of the new system moving from the project team to the performing Organization.

ALGORITHM

In our application we have used three algorithms:

- Machine Learning Algorithms
  - Linear SVM
  - KNN
- Deep Learning Algorithms
  - FCNN

3.4 Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good

separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher dimensional space, presumably making the separation easier in that space.

3.5 k- Nearest Neighbor Algorithm

"K-Nearest Neighbour (KNN)" is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It also stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. This algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. This algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data

3.6 Fully Convolutional Neural Network

It is a type of feed-forward artificial network where the connectivity pattern between its neurons is inspired by the organization of the animal **visual cortex**. Convolutional neural network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to automatically and adaptively learn spatial hierarchies of features through a back propagation algorithm. Familiarity with the concepts and advantages, as well as limitations, of convolutional neural network is essential to leverage its potential to improve radiologist performance and, eventually, patient care. It is one of the technique to do image classification and image recognition in neural networks. It is designed to process the data by multiple layers of arrays. This type of neural network is used in applications like image recognition or face recognition. The primary difference between CNN and other

neural network is that CNN takes input as a two-dimensional array. And it operates directly on the images rather than focusing on feature extraction which other neural networks do.

CNN takes an image as input, which is classified and process under a certain category such as dog, cat, lion, tiger, etc. The computer sees an image as an array of pixels and depends on the resolution of the image. Based on image resolution, it will see as **h * w * d**, where h= height w= width and d= dimension. For example, An RGB image is **6 * 6 * 3** array of the matrix, and the grayscale image is **4 * 4 * 1** array of the matrix.

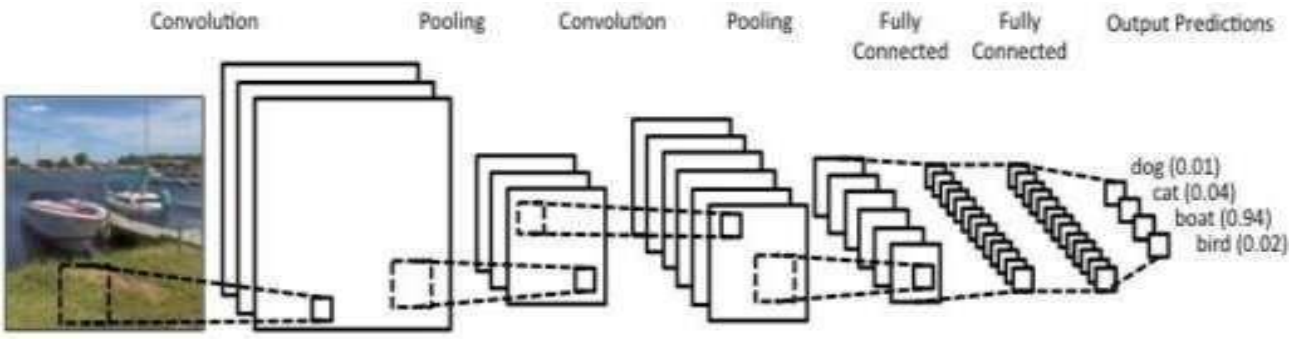

Fig 3.6.1 CNN Process

CHAPTER-4

DESIGNING

SYSTEM DESIGN

System design is the process or art of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development.

We look the design process from two distinct perspectives:

- Logical Design
- Physical Design

Logical Design:

The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system. This is often conducted via modelling, using an overabstract model of the actual system. In the context of system design are included. Logical design includes ER Diagrams i.e Entity Relationship Diagrams.

Physical Design:

The physical design relates to the actual input and output processes of the system. This is laid down in terms of how data is input in to a system, how it is verified/authenticated, how it is processed, and how it is displayed as output. In physical design, following requirements about the system are decided.

4.1 SYSTEM ANALYSIS

System analysis is the survey and planning of the system and planning of the system and project, it is the study and analysis of the existing and information system and requirements and priorities for a new or improved system.

One of the most important factors in the system analysis is to understand the system and its problems. A good understanding of the system enables designers to identify and correct problems based on the drawbacks of the existing system the system is being planned.
So the total definition of the problem has been analyzed.

4.1.1 EXISTING SYSTEM

Plant disease detection has become an important topic to ensure health of the plants and taking necessary measures to prevent it from getting deteriorated and causing heavy losses to the farmers. There should be solutions for detecting and classifying the diseases to get some knowledge which will later help in improving the quality of plants. So, patterns on the plant's leaves will help in identifying what problem it has. Various techniques of image processing and pattern recognition have been developed for detection of diseases occurring on plant leaves, stems, lesion etc. by the researchers. The earlier a disease appears on the leaf, the earlier it should be detected, identified and corresponding measures should be taken to avoid loss. Hence a fast, accurate and less expensive system should be developed.

4.1.2 DISADVANTAGES OF EXISTING SYSTEM

- Data Collection Problem
- It searches from a large sampling of the cost surface.

4.1.3 PROPOSED SYSTEM

Traditional methods for detecting diseases require manual inspection of plants by experts. This process needs to be continuous, and can be very expensive in large farms, or even completely unavailable to many small farm holders living in rural areas. "the Plant Village" Dataset is used . It consists of images of plant leaves taken in a controlled environment. In total, there are 54 306 images of 14 different plant species, distributed in 38 distinct classes given as species/disease pair. Classical methods rely on image pre-processing and the extraction of features which are then fed into one of the ML algorithms. Popular algorithm choices are Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Fully Connected Neural Networks (FCNN), Decision Trees, Random Forests etc.

4.2 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The three major areas consider while determining the feasibility of our project are:

- Economic Feasibility
- Technical Feasibility social Feasibility

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it.

4.3 INPUT AND OUTPUT DESIGN

4.3.1 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

- Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
- It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
- When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

4.3.2 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard

copy output. It is the most important and direct source information to the user.

Efficient and intelligent output design improves the system's relationship to help user decision-making.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

OBJECTIVES

- Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
- Select methods for presenting information.
- Create document, report, or other formats that contain information produced by the system.

4.4 SYSTEM ARCHITECTURE

The definition and modelling of an architecture dedicated to the activities of analysis of big data, as the ones produced by social networks as Twitter, is currently still at an early stage of its development and consolidation. Unlike traditional data warehouse or business intelligence systems, whose architecture is designed for structured data, systems dedicated to big data work instead with semi-structured data, or so called "raw data", i.e. without a particular structure. It should also be pointed out that such systems should be able to allow processing and analysis of data not only in batch mode, but also in a real-time fashion.

Nowadays a huge amount of data, daily produced by social networks, can be processed and analyzed for different purposes. These data are provided with several features, among which: dimension, peculiarities, source, reliability.

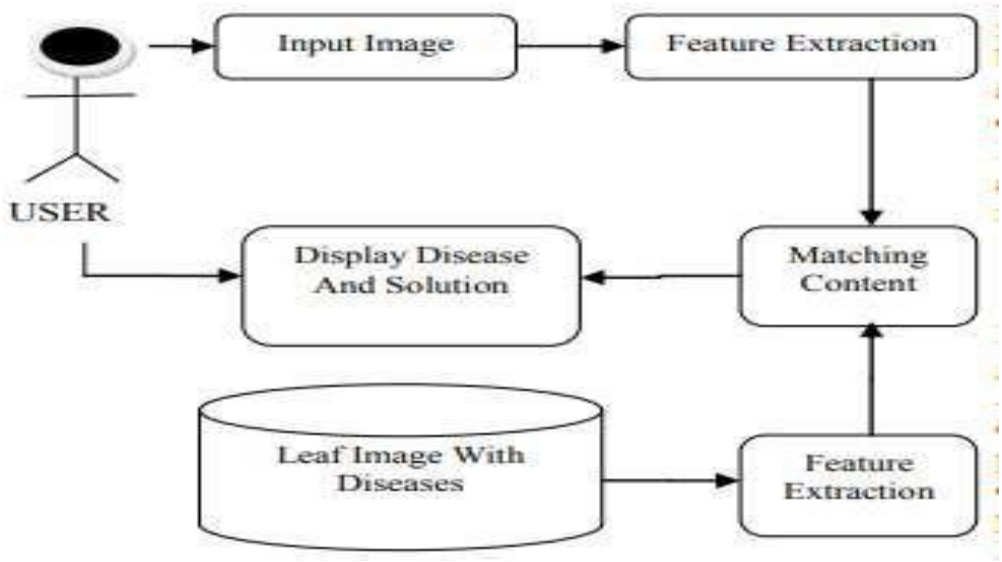

Fig 4.4: System Architecture

4.4.1 DATA FLOW DIAGRAM

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated.

Process: A process takes data as input, execute some steps and produce data as output.

External Entity: Objects outside the system being modelled, and interact with processes in system.

Data Store: Files or storage of data that store data input and output from process.

Data Flow: The flow data from process to process.

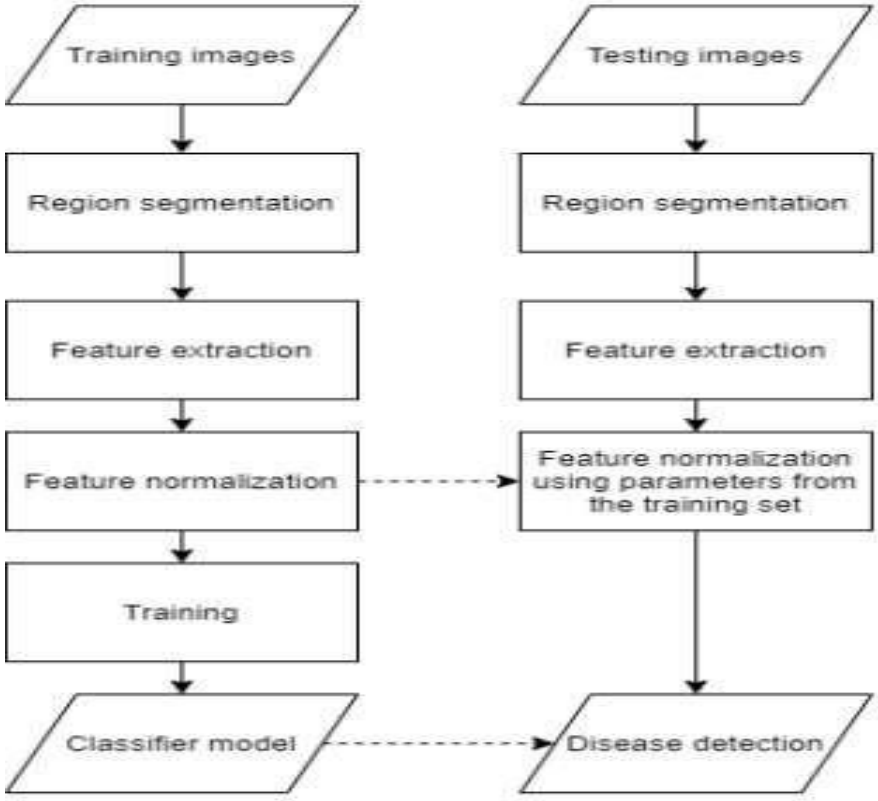

Fig 4.4.1: Data Flow Diagram

4.5 UML DIAGRAMS

The UML is a visual modelling language that enables system builders to create blue prints that capture their visions in a standard, easy-to-understand way, and provides a mechanism to effectively share and communicate these visions with others.

Design is the first step in the development phase for an engineered product or system. It is the place where quality is focused in software development. It is the only way we can accurately translate the user's requirements into a finished software product or a system. Software design serves as the foundation for all the engineers and software maintenance steps that follow. Without design we risk building an understandable design one that will fail when small changes are made, one that may be difficult to test, and one whose quantity cannot be accessed until late in the software engineering process.

Any real-world system is used by different users. The users can be developers, testers, business people, analysts, and many more. Hence, before designing a system, the architecture is made with different perspectives in mind. The most important part is to visualize the system from the perspective of different viewers. The better we understand the better we can build the system.

UML plays an important role in defining different perspectives of a system. These perspectives are:

- Design
- Implementation
- Process

Design of a system consists of classes, interfaces, and collaboration. UML provides class diagram, object diagram to support this.

Implementation defines the components assembled together to make a complete physical system. UML component diagram is used to support the implementation perspective.

Process defines the flow of the system. Hence, the same elements as used in Design are also used to support this perspective.

Diagrams in the UML

A diagram is the graphical presentation of a set of elements, most often rendered as a connected graph of vertices (things) and arcs (relationships). In theory, a diagram may contain any combination of things and relationships. In practice, however, a small number of common combinations arise, which are consistent with the five most useful views that comprise the architecture of software – intensive system. For this reason, the UML includes six such diagrams.

- Use Case Diagram
- Sequence Diagram
- Class Diagram
- Activity Diagram
- Component Diagram
- Entity Relationship Diagram

4.5.1 USE CASE DIAGRAM

A use case diagram shows a set of use cases and actors and their relationships. Use case diagrams address the static use case view of a system. These diagrams are especially important in organizing and modelling the behaviours of a system. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

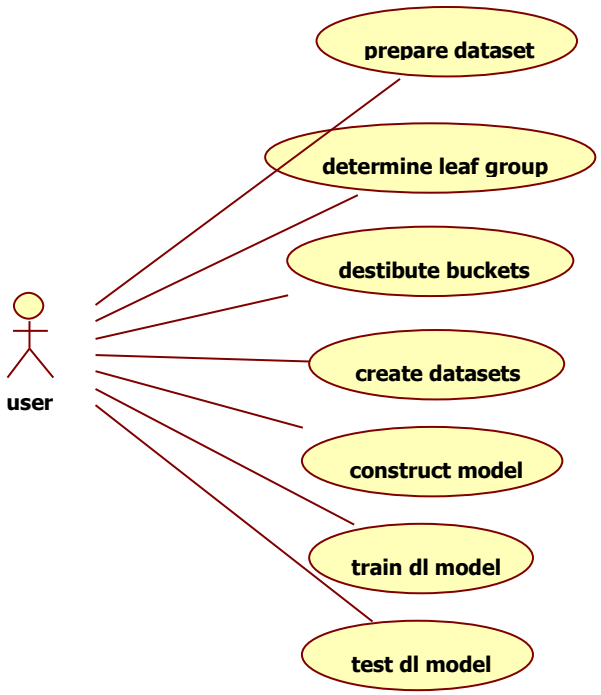

Fig 4.5.1: Use case Diagram

4.5.2 SEQUENCE DIAGRAM

A sequence diagram is an interaction diagram that emphasizes the time ordering of messages. It's consisting of a set of objects and their relationships, including the messages that may be dispatched among them. Sequence diagrams address the dynamic view of a system.

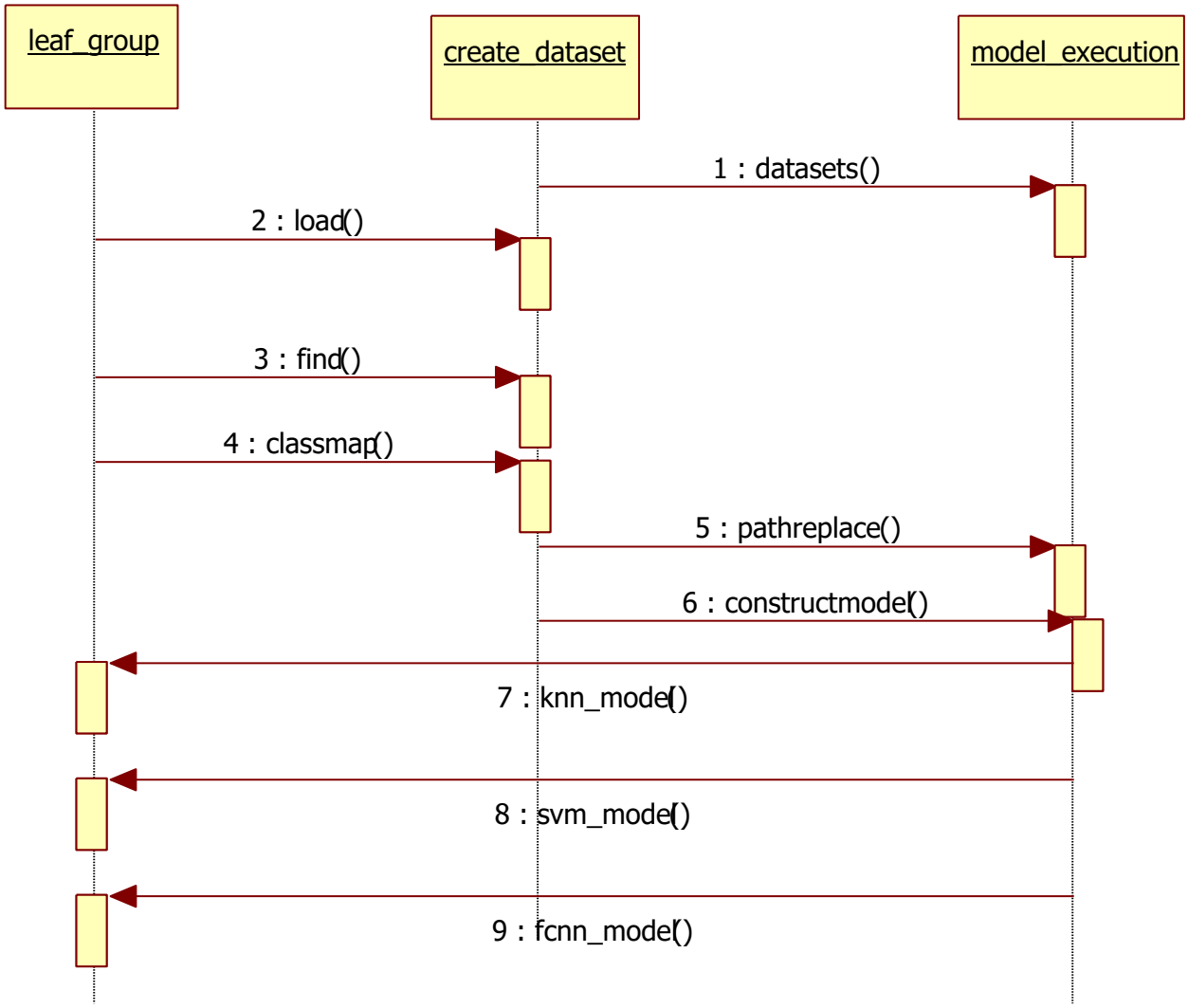

Fig 4.5.2: Sequence Diagram

4.5.3 CLASS DIAGRAM

A Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. It describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

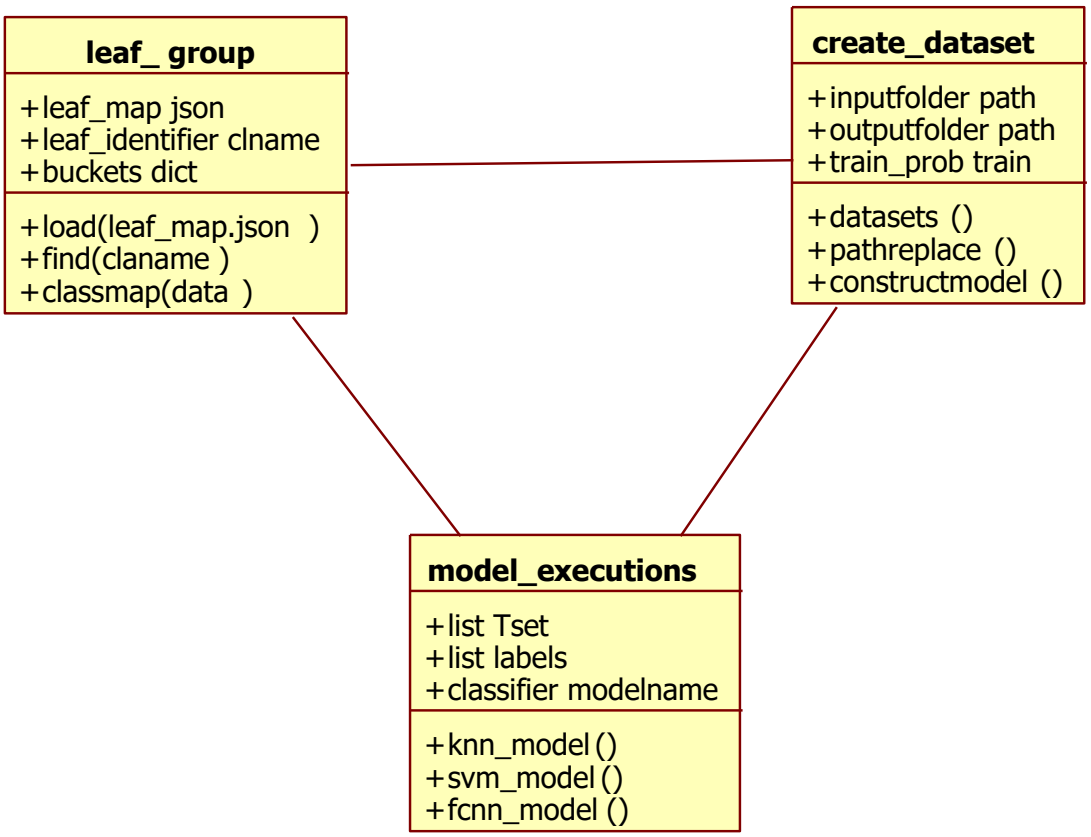

Fig 4.5.3: Class Diagram

4.5.4 ACTIVITY DIAGRAM

An activity diagram is a special kind of a state chart diagram that shows the flow from activity to activity within a system. Activity diagrams address the dynamic view of a system. They are especially important in modelling the function the function of a system and emphasize the flow of control among objects.

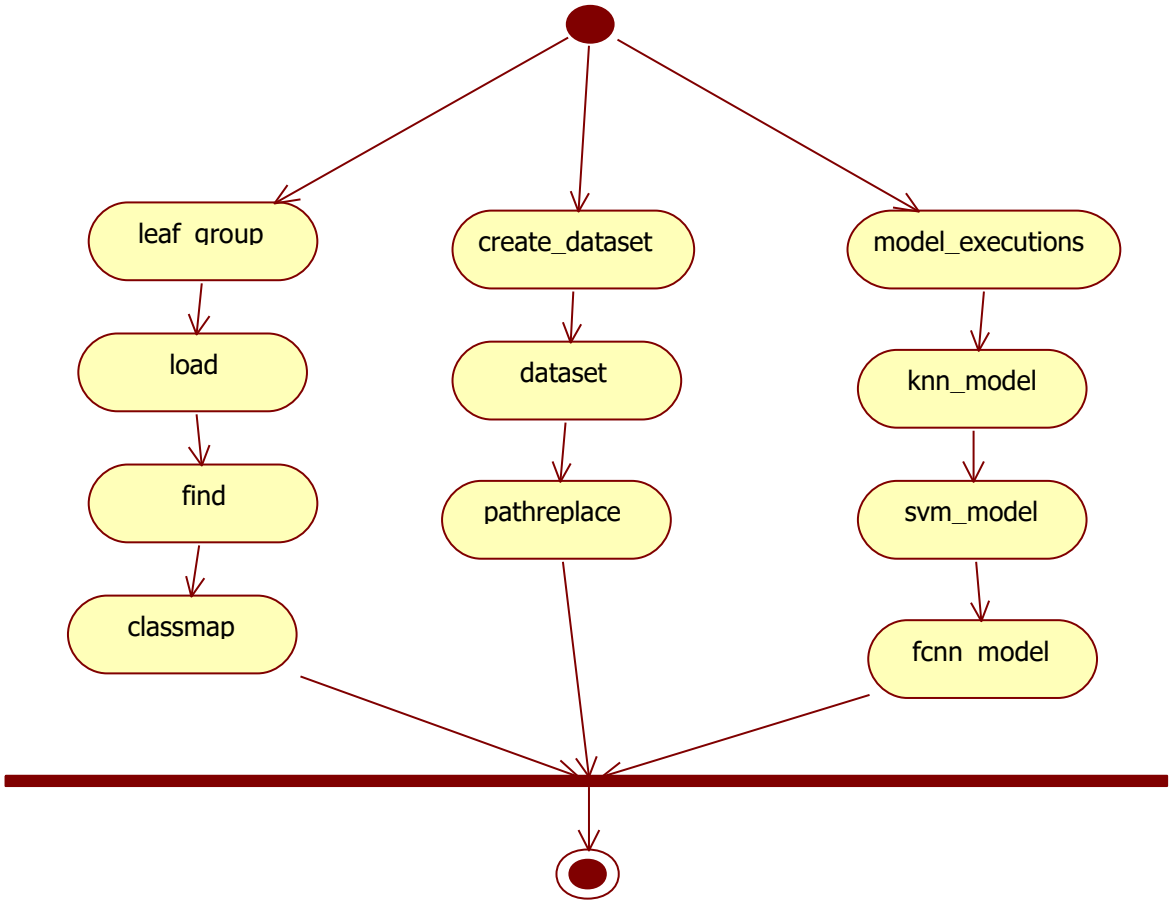

Fig 4.5.4 : Activity Diagram

CHAPTER-5
RESULT AND DISCUSSIONS

Our current project work covers state-of-the-art plant disease recognition using AI. We summarize a series of observations that emerge from this work in the following paragraph:

- **Available Databases and Size Issue:**

It is difficult to obtain leaf images for specific plant infections. Due to this fact, the sizes of the available plant data sets are very small. Only limited works have reported thousands of images for research purposes. Due to the small database size problem, a large portion of the data set is used for the training phase in most of the deep learning methods. However, very few exceptions are there. Furthermore, the available database images are collected in very constrained environmental conditions. We believe that images must be gathered in real-world conditions to make the algorithms more practical. Efficient image acquisition of leaf images is the need of the hour. If these images are captured in real-time scenarios, such databases would be warmly welcomed in the research community. In most of the recently reported works, the images captured with smart mobile devices are gaining popularity. Some single click image systems are also introduced, but much more is supposed to be done by the researchers to automate plant disease identification algorithms. The transition of imagecapturing systems to smart devices may help overcome serious issues related to database size.

- **Issues with Available Feature Extraction Methods:**

Performing the tasks of preprocessing, feature extraction, and segmentation plays a key role in developing a machine learning-based algorithm. Selecting the most suitable method for preprocessing and segmentation further depends on the nature of the data set. Among many techniques, one that is most suitable for a specific acquisition usually serves the purpose. We observe variability span in the reported algorithms so far under different modules. We observe somehow similar observations for various feature extraction techniques.In a nutshell the standardization of the report methods is yet to be fixed and achieved

- **Difficulties in Classification Module:**

Plant disease automation and detection is an active area of research for a long time. Considering very few images for training and testing, highly acceptable results are reported by researchers. Many classifiers are explored by researchers in this domain. This study concluded that back propagation neural net- work, SVM, and discriminant analysis (particularly linear) perform much better than others. These are then followed by Naïve Bayes, random forest, k- nearest neighbour, and multilayer perceptron. How- ever, stateof-the-art results are much improved with recently introduced optimized deep neural net- works. More proper utilization of deep convolutional neural networks can help in improving the results for large data sets.

- **Limitations of Available Systems:**

We argue that image analysis methods are comparatively better than the techniques that visually rate the severity of a particular disease. However, systems which are designed using these imaging techniques are not perfect. The performance of a system highly depends on the quality of the training data. In plant disease automation, it is the training images and certain extracted features, which significantly affect the performance of a system. A system trained with good quality data is trained well. However, most of the existing systems have a specific set of requirements needed to be fulfilled for a system to perform accurately. If some of these constraints are not fulfilled, the system may give inaccurate results, ultimately leading to wrong disease detection. For example, most of the DL-based methods particularly and conventional machine learning methods generally are faced with the problem of over fitting. Researchers must think of adaptive systems which are designed with more flexible requirements. Additionally, some generalized methods should be adapted which work in heterogeneous environments. For improving efficiency, in-depth knowledge of the methods and proper usage of the tools are also necessary.

- **Evaluation Measures**:

Many measures are available to compare different models for disease classification. These measures are based on four statuses: true-positive (TP), indicating the number of infected samples correctly identified; true-negative (TN), describing the correctly identified healthy images. Similarly, false-positives (FPs) showing the number of healthy samples that have been incorrectly classified as infected ones. Lastly, false-negative (FN) rep- resents the infected samples wrongly categorized as healthy ones. Accuracy is the ratio of the correct classification (TP) the total number of classifications (TP + FP + TN + FN). Precision represents the ratio of the correctly identified samples as infected (TP) to the total samples identified as infected (sum of TP and FP). Similarly, recall is the ratio of TP to the actual number of infected samples (sum of TP and FN). Lastly, F-measure represents the harmonic mean of precision and recall.

- **Comparison of Results.** State-of-the-art results for plants disease detection are compared and summarized for various data sets and methods

OUTPUT SCREEN SHOTS

- **HEALTHY PLANT SAMPLE PREDICTIONS**

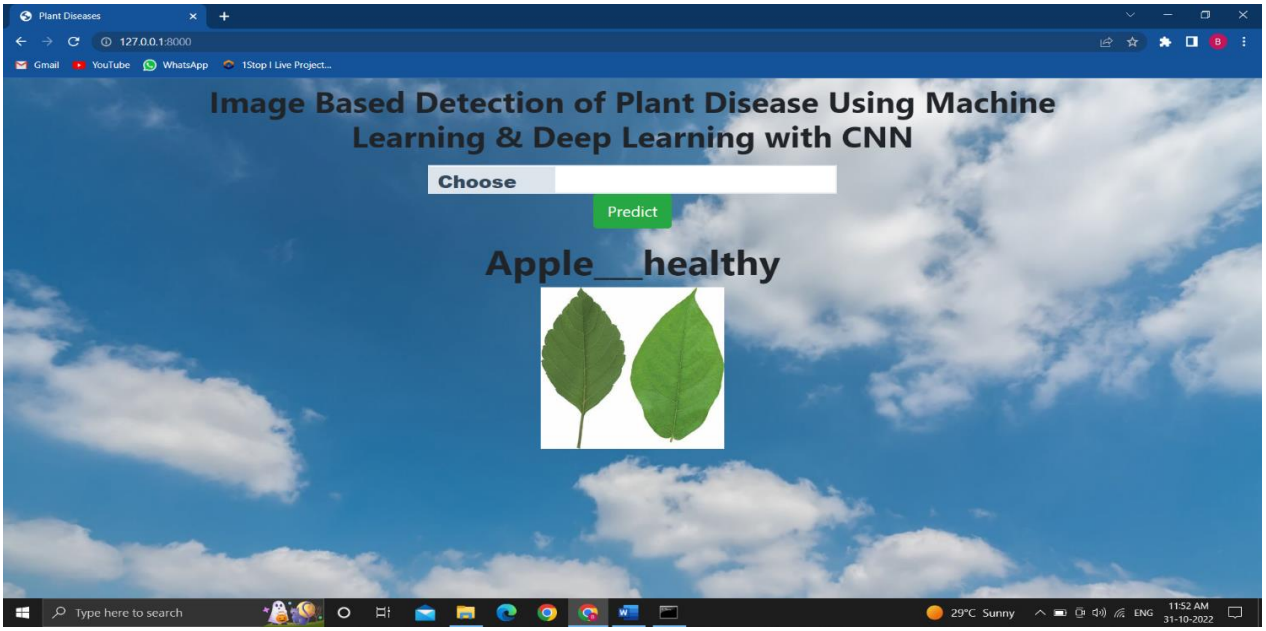

Fig 5.1.1 Healthy Image taken from dataset

Fig 5.1.2 Healthy Image taken from web

➢ DISEASED PLANT SAMPLE PREDICTIONS

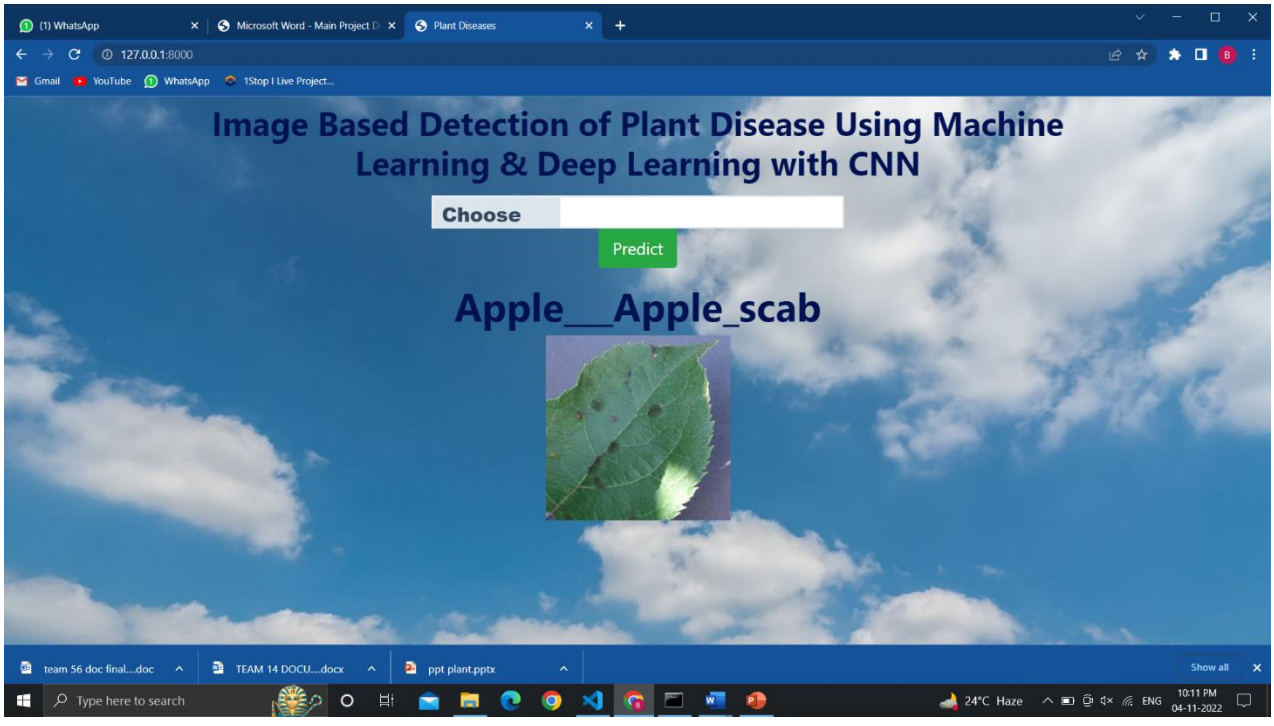

Fig 5.1.3 Diseased Image taken from dataset

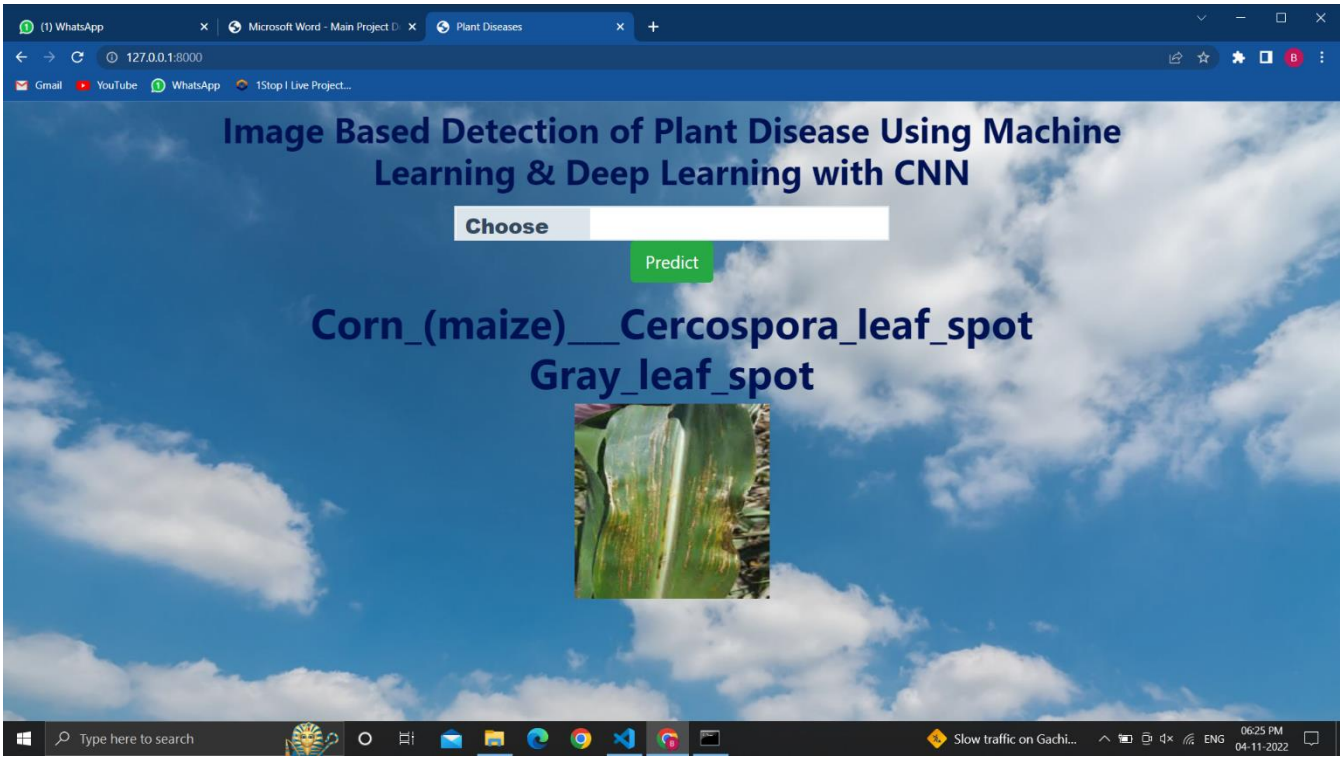

Fig5.1.4

Diseased Image taken from web

CHAPTER-6

CONCLUSION

AND

FUTUREWORK

6.1 CONCLUSION

This project presents the dominance of the DL method over the classical ML algorithms. Both the simplicity of the approach and the achieved accuracy confirm that the DL is the way to follow for image classification problems with relatively large datasets. As the achieved accuracy of the DL method is already very high, trying to improve its results on the same dataset would be of little benefit. Further work with the DL model could be done by expanding the dataset with more diverse images, collected from multiple sources, in order to allow it to generalize better. The considered ML algorithms achieved relatively high accuracy, but with error rates still an order of magnitude higher than the DL model. Further work in improving accuracy of the classical approach can be done by experimenting with other algorithms and by improving the features, as most likely they are the limiting factor of this approach.

FUTURE ENHANCEMENT

This project presents the dominance of the DL method over the classical ML algorithms. Both the simplicity of the approach and the achieved accuracy confirm that the DL is the way to follow for image classification problems with relatively large datasets.

As the achieved accuracy of the DL method is already very high, trying to improve its results on the same dataset would be of little benefit. Further work with the DL model could be done by expanding the dataset with more diverse images, collected from multiple sources, in order to allow it to generalize better.

The considered ML algorithms achieved relatively high accuracy, but with error rates still an order of magnitude higher than the DL model. Further work in improving accuracy of the classical approach can be done by experimenting with other algorithms and by improving the features, as most likely they are the limiting factor of this approach. Different image processing methods like Hue-based Segmentation, Morphological Analysis (i.e. erosion, dilation etc.), Blob Detection, Largest Connected Component, Color co-occurrence methodology, Texture Analysis etc. are implemented and applied.

CHAPTER-7
REFERENCES

REFERENCES

[1] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019: Highlights (ST/ESA/SER.A/423).

[2] Savary, Serge, et al. "The global burden of pathogens and pests on major food crops." Nature ecology & evolution 3.3 (2019): 430.

[3] Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection." Frontiers in plant science 7 (2016): 1419.

[4] Fujita, E., et al. "A practical plant diagnosis system for field leaf images and feature visualization." International Journal of Engineering & Technology 7.4.11 (2018): 49-54.

[5] Haralick, Robert M., Karthikeyan Shanmugam, and Its' Hak Dinstein. "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics 6 (1973): 610621.

[6] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[7] Cunningham, Padraig, and Sarah Jane Delany. "k-Nearest neighbour classifiers." Multiple Classifier Systems 34.8 (2007): 1-17.

[8] Haykin, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

[9] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[10] Duan, Kai-Bo, and S. Sathiya Keerthi. "Which is the best multiclass SVM method? An empirical study." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2005.