GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Deemed to be university)

Bengaluru-561203

Project Report on

"Classification on Leaves Based on their Species using Machine Learning Technique"

Submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology in Computer Science and Engineering

Submitted by:

Chunduri Avinash	**321710306006**
Akash Sridhar	**321710306001**
K Sai Samarth	**321710306019**
M Gopi Chand	**321710306026**

Under the guidance of:

Dr. Dayanand Lal

Assistant Professor

Department of Computer Science and Engineering

GITAM School of Technology, Bengaluru Campus,

Nagadenahalli, Doddaballapur Taluk,

Bengaluru Rural District, Karnataka-562163

2020-2021

GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Deemed to be University)

Bengaluru- 561203

Department of Computer Science & Engineering

Certificate

This is to certify that the Mini-Project titled **“Classification on Leaves Based on their Species using Machine Learning Technique”** is the bonafide work carried out by ‘ChunduriAvinash (321710306006), AkashSridhar (321710306001), KSaiSamarth (321710306019), MGopiChand (321710306026)’with a student of B-Tech (CSE) of GITAM Deemed to be University, Bengaluru campus during the academic year 2020-21, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (Computer Science and Engineering) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

.

Signature of the Guide	Signature of HOD	Signature of Director
Dr. Dayanand Lal	**Prof. Dr BRAHMANANDA S.H**	**Prof. Dr DINESH S**
Assistant Professor	Head of the Department	Director, GST
Department of CSE, GST	Department of CSE, GST	

DECLARATION

We**, ChunduriAvinash , AkashSridhar , KSaiSamarth ,MGopiChand** students of 8th-semester B.Tech in Computer Science & Engineering from GITAM (Deemed to be University), Bangalore, hereby declare that the dissertation work entitled **“Classification on Leaves Based on their Species using Machine Learning Technique”** has been carried out under the guidance of **, Dr. Dayanand Lal** Assistant Professor Department of Computer Science Engineering, GITAM (Deemed to be University), Bangalore, in the partial fulfilment of the requirement of the degree of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE ENGINEERING OF GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM).** We declare that we have not submitted this dissertation either in part or in full to any other University for the award of any degree.

Signature of the Students

Place: Bengaluru

Date:

ACKNOWLEDGEMENT

We had been able to complete our project successfully. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to GITAM (Deemed to be University), Bangalore for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project.

Would like to express our gratitude towards Prof. Dr Dinesh S (Director of GST), Prof. Dr Brahmananda S.H (HOD of CSE, GST) and also to all the other supporting faculty and staff for their kind co-operation and encouragement which helped us in the completion of this project.

Chunduri Avinash	321710306006
Akash Sridhar	321710306001
K Sai Samarth	321710306019
M Gopi Chand	321710306026

Abstract

Recognition of plants has become an active area of research as most of the plant species are at the risk of extinction. This paper uses an efficient machine learning approach for the classification purpose. This proposed approach consists of three phases such as pre-processing, feature extraction and classification. The pre-processing phase involves a typical image processing steps such as transforming to gray scale and boundary enhancement. The feature extraction phase derives the common DMF from five fundamental features. The main contribution of this approach is the Support Vector Machine (SVM) classification for efficient leaf recognition. Different leaf features which are extracted and orthogonalized into many principal variables are given as input vector to the SVM. Classifier tested with flavia-dataset and a real dataset the proposed approach produces very good accuracy and takes very less execution time.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Plants play a vital role in the environment. There will be no existence of the earth's ecology without plants. However, recently, several species of plants are at the danger of extinction. In order to protect plants and to catalogue various species of flora diversities, a plant database becomes very essential. There is huge volume of plant species worldwide. In order to handle such volumes of information, development of a rapid and competent classification technique has become an active area of research [1]. Moreover, along with the conservation feature, recognition of plants has also become essential to exploit their medicinal properties and using them as sources of alternative energy sources like bio-fuel. There are various ways to recognize a plant, like flower, root, leaf, fruit etc. Recently, computer vision and pattern recognition techniques have been applied towards automated process of plant recognition [2]The classification of plant leaves is a vital mechanism in botany and in tea, cotton and other industries [3], [4]. Additionally, the morphological features of leaves are employed for plant classification or in the early diagnosis of certain plant diseases [5].

Plant recognition is an essential and challenging task. Leaf recognition plays an important role in plant classification and its key issue lies in whether the chosen features are constant and have good capability to discriminate various kinds of leaves. The recognition procedure is very time consuming. Computer aided plant recognition is still very challenging task in computer vision because of improper models and inefficient representation approaches. The main aim of plant recognition is to evaluate the leaf geometrical morphological and Fourier moment based features. This data is very vital in identifying the various classes of plants. Ji Xiang, Huang and Xiao Feng [6] carried out their investigation on recognizing the known plant species by salient features of the leaf such as physiological length, width, diameter, perimeter, area, smooth factor, aspect ratio and Fourier moments which could be employed to discriminate with each other. The extraction of leaf features from a plant is a key step in the plant recognition process [7, 8]. This feature extraction process creates a new challenge in the field of pattern recognition . The data acquisition from living plant automatically by the computer has not been implemented.

CHAPTER 2

LITURATURE SUVERY

Plant Leaf Recognition Using Zernike Moments and Histogram of Oriented Gradients by Dimitris G. TsolakidisDimitrios I. KosmopoulosGeorge Papadourakis

A method using Zernike Moments and Histogram of Oriented Gradients for classification of plant leaf images is proposed in this paper. After preprocessing, we compute the shape features of a leaf using Zernike Moments and texture features using Histogram of Oriented Gradients and then the Support Vector Machine classifier is used for plant leaf image classification and recognition. Experimental results show that using both Zernike Moments and Histogram of Oriented Gradients to classify and recognize plant leaf image yields accuracy that is comparable or better than the state of the art. The method has been validated on the *Flavia* and the *Swedish Leaves* datasets as well as on a combined dataset.

Leaf and Flower Recognition Using Preferential Image Segmentation Algorithm by N. ValliammalS. N. Geethalakshmi

Automatic plant classification systems are essential for a wide range of applications including environment protection, plant resource survey, as well as for education. With the aid of advanced information technology, image processing and machine learning techniques, automatic plant identification and classification will enhance such systems with more functionality, such as automatic labeling and flexible searching. Image segmentation and object recognition are two aspects of digital image processing which are being increasingly used in many applications including leaf recognition. In this paper, the Preferential Image Segmentation (PIS) method is used to segment an object of interest from the original image. A probabilistic curve evolution method with particle filters is used to measure the similarity between shapes during matching process. The experimental results prove that the preferential image segmentation can be successfully applied in leaf recognition and segmentation from a plant image.

An Effective Tea Leaf Recognition Algorithm for Plant Classification Using Radial Basis Function Machine by Arunpriya Antony Selvadoss Thanamani

A leaf is an organ of a vascular plant, as identified in botanical terms, and in particular in plant morphology. Naturally a leaf is a thin, flattened organ bear above ground and it is mainly used for photosynthesis. Recognition of plants has become an active area of research as most of the plant species are at the risk of extinction. Most of the leaves cannot be recognized easily since some are not flat (e.g. succulent leaves and conifers), some does not grow above ground (e.g. bulb scales), and some does not undergo photosynthetic function (e.g. cataphylls, spines, and cotyledons).In this paper, we mainly focused on tea leaves to identify the leaf type for improving tea leaf classification. Tea leaf images are loaded from digital cameras or scanners in the system. This proposed approach consists of three phases such as preprocessing, feature extraction and classification to process the loaded image. The tea leaf images can be identified accurately in the preprocessing phase by fuzzy denoising using Dual Tree Discrete Wavelet Transform (DT-DWT) in order to remove the noisy features and boundary enhancement to obtain the shape of leaf accurately. In the feature extraction phase, Digital Morphological Features (DMFs) are derived to improve the classification accuracy. Radial Basis Function (RBF) is used for efficient classification. The RBF is trained by 60 tea leaves to classify them into 6 types. Experimental results proved that the proposed method classifies the tea leaves with more accuracy in less time. Thus, the proposed method achieves more accuracy in retrieving the leaf type

Computer-Aided Plant Species Identification (CAPSI) Based on Leaf Shape Matching Technique by Ji-Xiang Du, De-Shuang Huang, Xiao-Feng Wang,

In this paper, an efficient computer-aided plant species identification (CAPSI) approach is proposed, which is based on plant leaf images using a shape matching technique. Firstly, a Douglas - Peucker approximation algorithm is adopted to the original leaf shapes and a new shape representation is used to form the sequence of invariant attributes. Then a modified dynamic programming (MDP) algorithm for shape matching is proposed for the plant leaf recognition. Finally, the superiority of our proposed method over traditional approaches to plant species identification is demonstrated by experiment. The experimental result showed that our proposed algorithm for leaf shape matching is very suitable for the recognition of not only intact but also partial, distorted and overlapped plant leaves due to its robustness .

CHAPTER 3

SYSTEM REQUIREMENT SPECIFICATION

3.1 System Analysis

Prediction of terrorism activities is an important area of concern for researchers. The large number of events makes it difficult to predict terrorist group responsible for some terrorist activity.

The current research is focused on finding out the correlation between terrorism and its causal factors. Existing efforts have not been good enough for prediction. Machine learning approaches can ad in predicting the likelihood of a terrorist attack, given the required data. The results of this work can help the security agencies and policy makers to eradicate terrorism by taking relevant and effective measures.

Hence there is an approach to analyzing terrorism region and country with the machine learning techniques and terrorism specific knowledge to fetch conclusions about terrorist behavior patterns.

3.2 Functional Requirement

The particular necessities are user interfaces. The outside clients are the customers.

Every one of the customers can utilize this product for ordering and looking.

- Hardware Interfaces: The outside equipment interface utilized for ordering and looking is PCs of the customers. The PC's might be portable PCs with remote LAN as the web association gave will be remote.
- Software Interfaces: The working Frameworks can be any rendition of windows.
- Performance Prerequisites: The PC's utilized must be atleast pentium 4 machine with the goal that they can give ideal execution of the item.

3.3 Non-Functional Requirements

Non utilitarian necessities are the capacities offered by the framework. It incorporates time imperative and requirement on the advancement procedure and models. The non useful prerequisites are as per the following:

- Speed: The framework ought to prepare the given contribution to yield inside fitting time.
- Ease of utilization: The product tought to be easy to understand. At that point the clients can utilize effortlessly, so it doesn't require much preparing time.
- Reliability: The rate of disappointments ought to be less then just the framework is more solid.
- Portability: It thought to be anything but difficult to actualize in any framework

H/W System Configuration:

Processor	Dual Core.
Speed	1.1 G Hz.
RAM	1GB.
Hard Disk	500MB.

S/W System Configuration:

Operating System	Windows 10.
Technology	Machine Learning.
Front End	GUI-tkinter.
IDLE	Python 3.7 or higher.

Hardware requirements

The most widely recognized arrangement of prerequisites characterized by any working framework or programming application is the physical PC assets, otherwise called equipment, An equipment necessities list is frequently joined by an equipment similarity list, particularly if there should be an occurrence of working frameworks. A HCL records tried, perfect, and now and then incongruent equipment gadgets for a specific working framework or application. The accompanying sub-segments examine the different parts of equipment prerequisites.

All PC working frameworks are intended for a specific PC design. Most programming applications are restricted to specific working frameworks running on specific structures. In spite of the fact that engineering free working frameworks and applications exist, most should be recompiled to keep running on another design.

The energy of the focal preparing unit (CPU) is a central framework necessity for any product. Most programming running on x86 engineering characterize preparing power as the model and the clock speed of the CPU. Numerous different highlights of a CPU that impact its speed and power, similar to transport speed, store, and MIPS are frequently overlooked. This meaning of energy is regularly wrong, as AMD Intel Pentium CPUs at comparative clock speed frequently have distinctive throughput speeds.

- 10GB HDD(min)
- 128 MB RAM(min)
- Pentium P4 Processor 2.8Ghz(min)

Software requirements

Programming necessities manage characterizing programming asset necessities and requirements that should be introduced on a PC to give ideal working of an application.

These necessities or requirements are for the most part excluded in the product establishment bundle and should be introduced independently before the product is introduced.

- Python 3.7 or higher
- Pycharm
- opencv

3.4 Tools and Technology details

Tool: IDLE is Python's Integrated Development and Learning Environment. It allows programmers to easily write Python code. Just like Python Shell, IDLE can be used to execute a single statement and create, modify, and execute Python scripts.

Technology:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: *The ability to learn*. Machine learning is actively being used today, perhaps in many more places than one would expect.

Data Preprocessing in Machine learning

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

Why do we need Data Preprocessing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
It involves below steps:

- **Getting the dataset**
- **Importing libraries**
- **Importing datasets**
- **Finding Missing Data**
- **Encoding Categorical Data**
- **Splitting dataset into training and test set**
- **Feature scaling**

CHAPTER 4

SYSTEM DESIGN

System design is the process 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. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development "blends the perspective of marketing, design, and manufacturing into a single approach to product development," then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

4.1 Overall system architecture

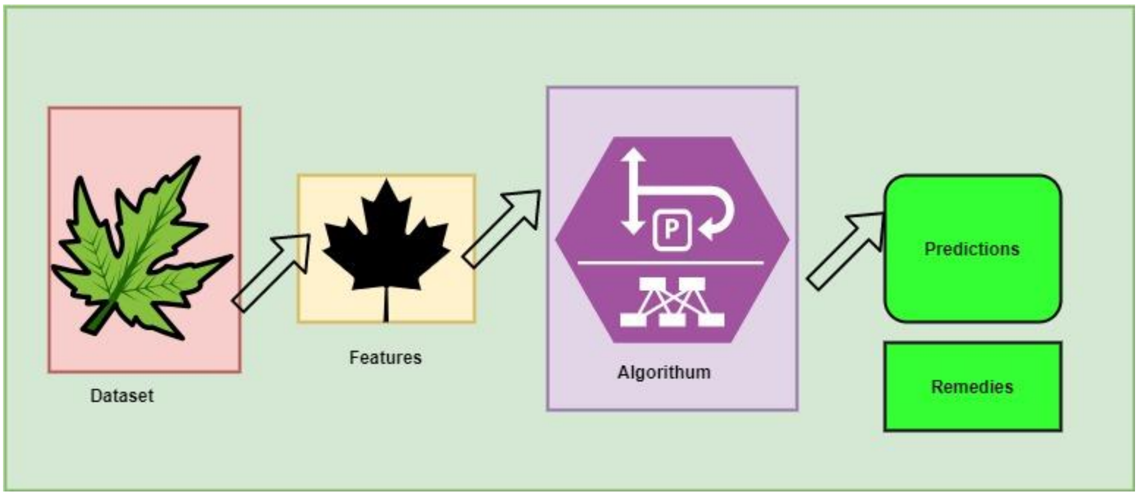

Fig 1 Architecture

4.2 Activity Diagram

We use Activity Diagrams to illustrate the flow of control in a system and refer to the steps involved in the execution of a use case. We model sequential and concurrent activities using activity diagrams. So, we basically depict workflows visually using an activity diagram. An activity diagram focuses on condition of flow and the sequence in which it happens. We describe or depict what causes a particular event using an activity diagram

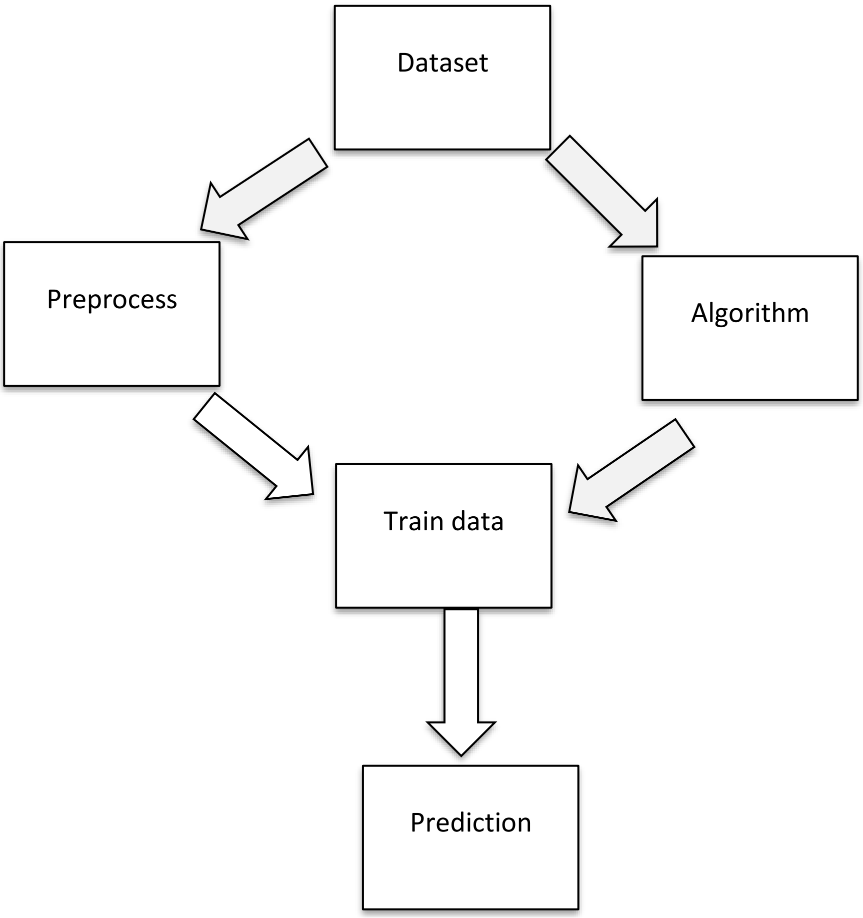

Fig 2 Activity Diagram

4.3 Data Flow Diagram(DFD)

A data flow diagram is a graphical representation of the "flow" of data through an information system, modeling its process aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design). The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

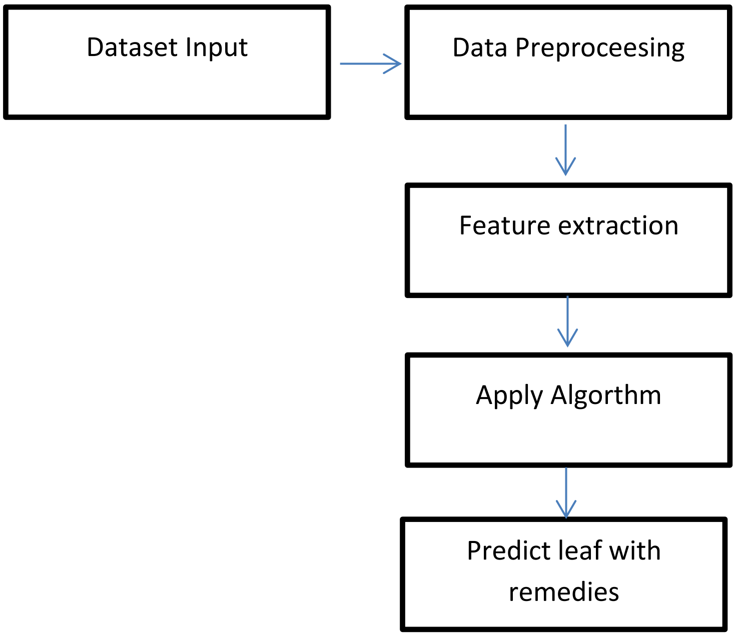

Fig 3 Data Flow Diagram

CHAPTER 5

IMPLEMENTATION

There are 3 steps for implementation;

1. Dataset collection
2. Data Feature Extraction
3. Machine learning algorithm apply
4. Prediction

5.1 Dataset collection:

We collect the leaf flavia dataset it includes total 32 leaf categories

5.2 Data Feature Extraction

There are many features extracted from the individual leaf with opencv mechanisms.,

Features
['area','perimeter','pysiological_length','pysiological_width','aspect_ratio','rectangularity'
,'circularity,'mean_r','mean_g','mean_b','stddev_r','stddev_g','stddev_b',
'contrast','correlation','inverse_difference_moments','entropy']

Then this will go to the svm algorithm

5.3 Machine learning algorithm apply

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Let’s consider two independent variables x1, x2 and one dependent variable which is either a blue circle or a red circle.

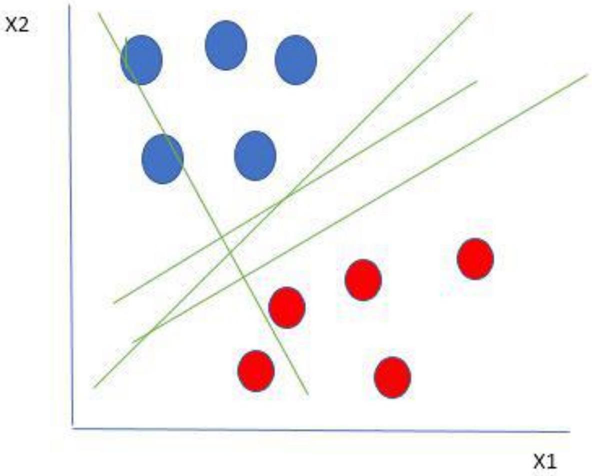

Fig 4 Identify the right hyper-plane

From the figure above its very clear that there are multiple lines (our hyperplane here is a line because we are considering only two input features x1, x2) that segregates our data points or does a classification between red and blue circles. So how do we choose the best line or in general the best hyperplane that segregates our data points.

Select One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes .choose the best hyper-plane:

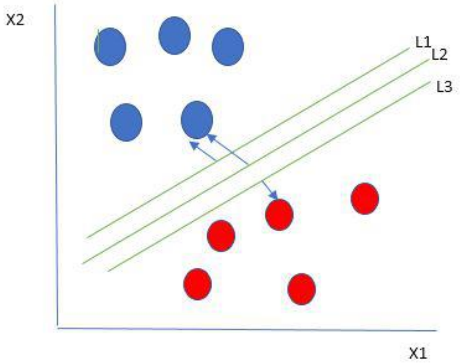

Fig 5 Finding best hyperplane

So we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. So from the above figure, we choose L2.

Here we have one blue ball in the boundary of the red ball. So how does SVM classify the data? It’s simple! The blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers

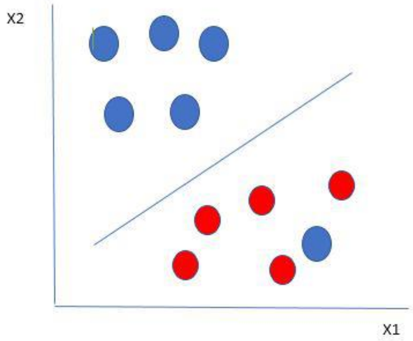

Fig 6 Ignoring outlier data

So in this type of data points what SVM does is, it finds maximum margin as done with previous data sets along with that it adds a penalty each time a point crosses the margin. So the margins in these type of cases are called soft margin. When there is a soft margin to the data set, the SVM tries to minimize (1/margin+∧(∑penalty)). Hinge loss is a commonly used penalty. If no violations no hinge loss. If violations hinge loss proportional to the distance of violation.

5.4 Prediction:

Then trained model with sample input to test which category also their remedies

CHAPTER 6

TESTING

Testing is a critical element which assures quality and effectiveness of the proposed system in (satisfying) meeting its objectives. Testing is done at various stages in the System designing and implementation process with an objective of developing an transparent, flexible and secured system. Testing is an integral part of software development. Testing process, in a way certifies, whether the product, that is developed, complies with the standards, that it was designed to. Testing process involves building of test cases, against which, the product has to be tested.

Test objectives

- Testing is a process of executing a program with the intent of finding an error.
- A good case is one that has a high probability of finding an undiscovered error.
- A successful test is one that uncovers a yet undiscovered error. If testing is conducted successfully (according to the objectives) it will uncover errors in the software. Testing can't show the absences of defects are present. It can only show that software defects are present.

6.1 Testing principles

Before applying methods to design effective test cases, a software engineer must understand the basic principle that guides software testing. All the tests should be traceable to customer requirements.

Unit Testing

The first level of testing is called unit testing. Unit testing verifies on the smallest unit of software designs-the module. The unit test is always white box oriented. In this, different modules are tested against the specifications produced during design for the modules. Unit testing is essentially for verification of the code produced during the coding phase, and hence the goal is to test the internal logic of the modules. It is typically done by the programmer of the module. Due to its close association with coding, the coding phase is frequently called "coding and unit testing." The unit test can be conducted in parallel for multiple modules.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Table X: Functional Testing items

Valid Input	Identified classes of valid input must be accepted.
Invalid Input	Identified classes of invalid input must be rejected.
Functions	Identified functions must be exercised.
Output	Identified classes of application outputs must be exercised.

Systems/Procedures: Interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

CHAPTER 7

CONCLUSION

A new approach of plant classification based on leaves recognition is proposed in this paper. An efficient machine learning approach for plant leaf recognition is presented in this research. The approach consisted of three phases namely the pre-processing phase, feature extraction phase and the classification phase. The computer can automatically classify 32 kinds of plants via the leaf images loaded from digital cameras or scanners. 12 commonly used Digital Morphological Features (DMFs) obtained from 5 basic features are extracted in the feature extraction phase. SVM classifier is adopted for the classification approach as it has better accuracy, fast training speed and simple structure. 12 features are extracted and processed by PCA to form the input vector of SVM. The performance of the proposed approach is evaluated based on the accuracy and execution time. For Further research by incorporating efficient kernel functions the performance of the classifier can be improved.

References

[1] Jyotismita Chaki, and Ranjan Parekh, “Plant Leaf Recognition using Shape based Features and Neural Network classifiers,” International Journal of Advanced Computer Science and Applications (IJACSA), 2011, vol. 2, no. 10.

[2] J. Pan, and Y. He,”Recognition of plants by leaves digital image and neural network,” International Conference on Computer Science and Software Engineering, 2008, vol. 4, pp. 906 – 910.

[3] Cotton Incorporated USA, The classification of Cotton, 2005, http://www.cottoninc.com/ClassificationofCotton.

[4] National Institute for Agricultural Botany, Chrysanthemum Leaf Classification. Cambridge, 2005.

[5] N. Kumar, S. Pandey, A. Bhattacharya, and P.S. Ahuja, “Do leaf surface characteristics affect agro bacterium infection in tea,” J. Biosci., vol. 29, no. 3, 2004, pp. 309–317.

[6] Ji-Xiang Du,De-Shuang Huang, Xiao-Feng Wang, and Xiao Gu, “Computer-aided plant species identification (capsi) based on leaf shape matching technique,” Transactions of the Institute of Measurement and Control, vol. 28, 2006, pp. 275-284.

[7] Y. Li, Q. Zhu, Y. Cao, and C. Wang, “A Leaf Vein Extraction Method based on Snakes Technique,” Proceedings of IEEE International Conference on Neural Networks and Brain, 2005.

[8] H. Fu, and Z. Chi, “Combined thresholding and Neural Network Approach for Vein Pattern Extraction from Leaf Images,” IEEE Proceedings-Vision, Image and Signal Processing, 2006, vol. 153, no. 6.