CHAPTER 1

INTRODUCTION

Globally, it has been found that there are more than 1.7 million species of living organisms (human beings, plants and algae) on Earth, out of which, plants species plays a vital role in human life. Plants are an essential resource for human well-being and can exist everywhere. Most of the plants carry significant information for the development of human society and are considered as essential resource for human well-being. Plants are of plenty of use as they form the base for food chain and a lot of medicines are derived from plants. Plants are also vitally important for environmental protection.

Even after several innovative advancements made in the field of botany, there are still a huge number of plants that are yet to be discovered, identified and used. It is a well-known fact that unknown plants are treasures waiting to be found. Today's ethno-botanists are combining regions of the world, looking for future medicines and agricultural products.

The functional characteristics and the association of plants within ecosystems are explored by them in order to understand the need for diversity to manage the plant resources. Scientists of 21st century are exploring how genetic diversity and ecological sensitivity are necessary in solving problems such as feeding the population and fighting disease. Two main plant aspects of plant taxonomy that play a vital role in these endeavours are the identification and classification of plants.

- **Plant Identification** is the determination of the identity of an unknown plant in comparison with previously collected specimen. The process of recognition connects the specimen with a botanical name. Once this connection is established, related details like name and other properties of the plant can be easily obtained.

- **Plant Classification** is the placing of known plants into groups or categories to show some relationship. They use features that can be used to group plants into a known hierarchy.

This research focuses on the automation of plant identification through leaf recognition. Apart from using the whole plant, the automation of plant identification can be performed using various parts of a plant anatomy like stem, flower, petal, seed and leaf.

This study uses the leaf part of the plant to identify a plant. The continued interest in biodiversity along with the ease of creating digital images, increased the need for processing power of computers and economical methods. Inorder to gather the information, plant identification using computers has become an interesting subject of research. Global shortage of expert taxonomists has further increased the demand for automated tools that would allow non-botanical persons to carry out valuable field work of identifying and characterizing plants. These tools are of importance in several fields including agriculture, forestry and pharmacological science (Cotton Incorporated USA, 2009; National Institute for Agricultural Botany, 2005). The first step during the design and development of such tools starts with leaf recognition. Compared with other methods, such as cell and molecule biology methods, identification of plants based on leaf image is the most successful and proven method (Wu *et al.*, 2007). Sampling leaves and obtaining a photograph of them is convenient and viable, due to the availability of low cost digital cameras.

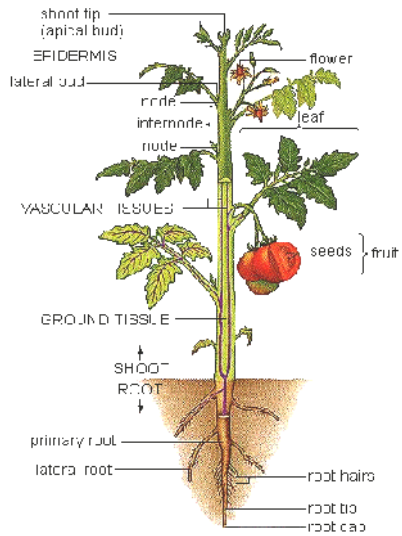
Currently, plant identification through leaf recognition involves finding information about a plant that most matches the species name (key) that has to be known in advance. Though identifying plants using such key is a time consuming task, correct utilization of key plays a direct role in the success of the plant search. The alternative method of allowing users to provide a leaf image is very convenient, user friendly and eliminates the need for key.

The task combines the challenges of different fields like image processing, machine learning and pattern recognition. Identifying the most favourable algorithms and techniques from these fields, for plant identification through leaf recognition, is the main focus of this study. The rest of the chapter presents the introductory materials related to this topic.

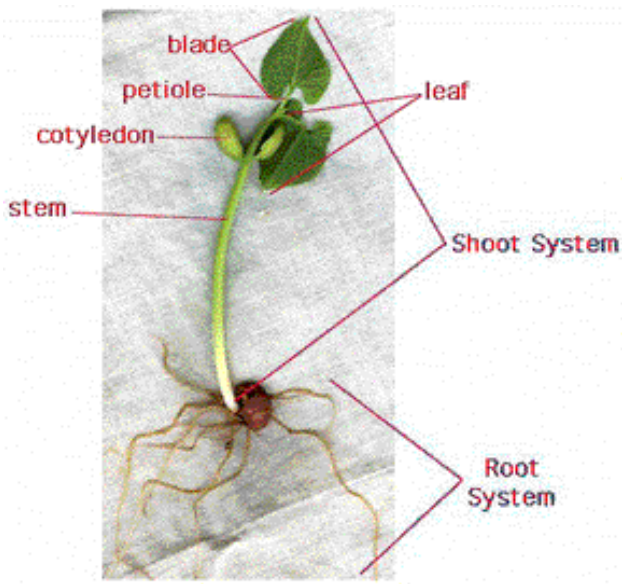
1.1. OVERVIEW OF PLANT KINGDOM

Plants are living organisms belonging to the vegetable kingdom that can live on land or in water reference. There are more than 300,000 species of plants. Plants are the backbone of all life on Earth and an essential resource for human well-being. Life on earth depends on plants. They are responsible for the presence of oxygen, which is vital for human beings. They are the base of the human food chain and humans directly or indirectly take their food from plants. Plants regulate the water cycle: they help distribute and purify the planet's water. Plants store carbon and have helped keep much of the carbon dioxide produced out of the atmosphere. Plants are used to prevent soil erosion and they are also used for providing building materials. They play a vital role in the field of medicines, where more than one-quarter of all prescribed drugs come directly from derivatives of plants. (http:\\www.botanical-online.com/ theimportanceofplants.htm).

Plant anatomy, a term referring to the structure of plants, provides a description of the physical and external forms of plant structure and function (Raven *et al.*, 2005; Evert, 2006). A typical plant body consists of different parts as illustrated in Figure 1.1a. The plant body consists of two major organ systems, namely, shoot system and root system (Figure 1.1b). The shoot systems exist above the ground and include organs like buds, leaves, fruits, flowers and seeds.



(a)



(b)

Figure 1.1 : Parts of Plant Body

Source : (a) http://www.uic.edu/classes/bios/bios100/labs/plantanatomy.htm
(b) http://www.emc.maricopa.edu/faculty/farabee/biobk/biobookplantanat.html

Plants are enormously important to human welfare. Plants are a source of food, housing materials, clothing, dyes and medicines. They replenish the air through which animals and human being breathe . They also help to cool the atmosphere and fix soil in place.

During plant identification using leaves, botanists place importance on various external leaf characteristics like size, shape, color, texture and veins. These characteristics make leaves determinant, as they grow and achieve a specific pattern and shape, after which the changes in leaf structure stops. This makes leaf as a promising candidate for identification of plants.

1.2. THE LEAF – IMPORTANT ORGAN OF PLANT

A leaf is an organ of a vascular plant, as defined in botanical terms, and in particular in plant morphology. Foliage is a mass noun that refers to leaves as a feature of plants. Typically ,a leaf is a thin, flattened organ borne above ground and specialized for photosynthesis (James, 2009). The various useful functions of leaves are

- Photosynthesis process for manufacturing Food Respiration process (gas/air exchange),
- Security to vegetative and floral buds,
- Transpiration (medium for water transport) and
- Store food during germination.

The main component of a plant system is the leaves and it is a non-rigid object, thus leading to various types of deformation. They are readily apparent structures on plants that are immediately available for examination and analysis without any additional experimentation. Leaf recognition consists of tasks that involve analysis of various intra and inter-class variations like size, shape, color and vein structure. Figure 1.2 presents the various external leaf parts and a description of each of these parts is given below.

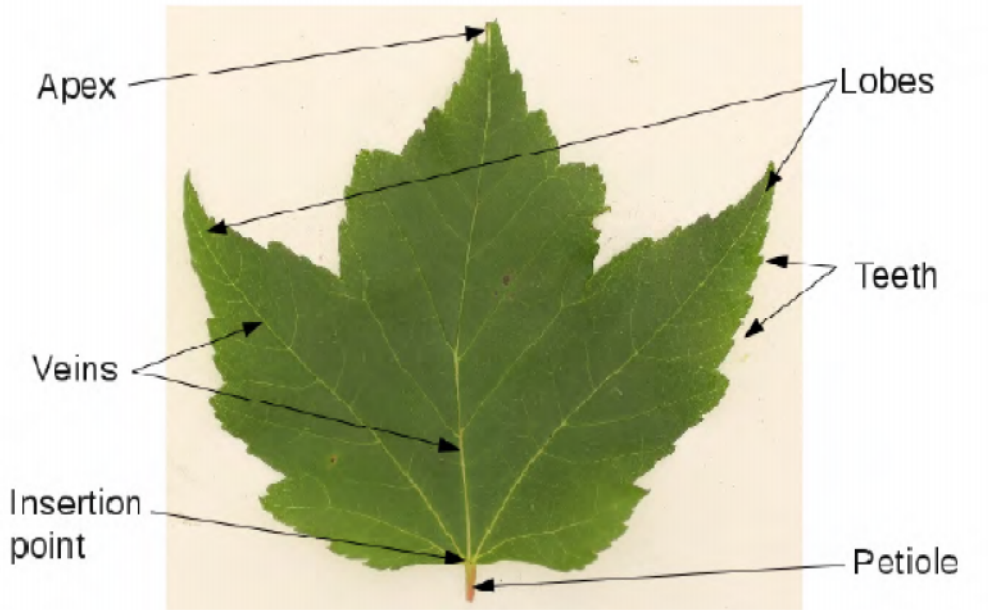


Figure 1.2 : Features of Leaf

Source : Cope *et al.* (2012).

1. Apex — Leaf tip or the outer end of the leaf. It is the point farthest from the point of attachment.
2. Base — Leaf part that connects to the stem (petiole)
3. Teeth or Margin — Edges of leaves
4. Vein — Vascular tissue of the leaf, located in the spongy layer of the mesophyll.
5. Venation — The pattern of the veins is called venation, and is typically characterized by hierarchical structures with abundant closed loops.
6. Lobe — Part of a leaf, often rounded, formed by incisions to about halfway to the midrib.
7. Petiole — The stalk of a leaf.

Leaves are important organs of a plant. They are planar and can be easily given as input to computers. Natural history museums have recently provided on-line access to hundreds of thousands of images of leaf specimens to help in identifying rare plants. In spite of large amount of leaf image databases available, the task of finding a plant, given a leaf specimen, currently require the species name to be known in advance. This problem can be solved by allowing the user to search through this data using algorithms that match unknown plant's leaf image with previously discovered images.

1.3. NEED FOR AUTOMATED PLANT IDENTIFICATION

Plant identification is an important task because of concerns about climate change and the resultant changes in geographic distribution along with abundance of species. Development of new crops often depends on the incorporation of genes from wild relatives of existing crops and hence it is important to keep track of the distribution of all plant taxonomy (Cope *et al.*, 2012).

Automated identification of plant species using leaf images is a worthwhile goal because of the current combination of rapidly dwindling biodiversity and the shortage of suitably qualified taxonomists. This is particularly important in geographic locations,

- that currently have a huge number of species and
- those with the largest number of species restricted to that geographic area.

The species to which an organism belongs is often regarded as its most significant taxonomic rank.

Identification of a plant to its class allows access to historical or existing knowledge and the mapping of a plant to a class currently depends on a specific name. As a variety of hybrids exists, an automatic method that depends on plant features instead of its name is very much favoured.

In addition, with the deterioration of environments, even though many of the rare plant species are already dead, still many more of the rare plant species are at the margin of extinction. So, the investigation of plant recognition can contribute to environmental protection (Du *et al.*, 2007). The plant world is in constant flux, due to human and other factors, the possibility of extinction for many plants and animals can be envisaged. Just as importantly, the need to understand ecological systems which preserve biodiversity is also realized. Today's scientists are exploring how genetic diversity and ecological sensitivity are necessary in solving such problems as feeding the population and fighting human diseases. The recognition of plant leaves is a vital process in botany and in tea, cotton and other industries and is also used during early diagnosis of plant detects like diseases.

Thus, automatic plant classification is vital to these endeavours and is considered by this research work.

1.4. CHALLENGES OF AUTOMATIC PLANT RECOGNITION

The design and development of automatic recognition and identification systems for plants is important and has numerous usage. This field is considered as a challenging field, which motivates more and more researches to be conducted worldwide. A number of systems that aim to recognize plant species from the shapes of their leaves have been developed. The main challenge of automatic leaf recognition using leaf images is to develop computational methods which learn to distinguish among a number of classes from examples. This ability is instrumental in building the next-generation artificial systems, which can cope up with novel situations and aims to achieve general goals as opposed to specific and which integrate capabilities normally associated with people. The following are some of the issues that are faced by the existing plant identification systems.

- **Image quality:** Quality of the leaf image captured plays a very important role in the accuracy of plant identification. In general, the quality of leaf

images is affected by three factors, namely, contrast, blur and noise. The presence of these degradation factors has a direct impact on the performance of the automation process.

- **Leaf variation**: In general scenario, a leaf image can take a great number of biological variations.These variations produce more than one representation of the same leaf. An example of variation of leaves for a single specimen of *Quercus nigra* is shown in Figure 1.3.

Figure 1.3 : Example of Leaf Variation

Source : Cope *et al.* (2012)

The analysis and identification process has to handle this scenario carefully. Accurate and efficient feature extraction techniques that best distinguish these similar leaves are required for successful design of automated system. Further, the availability of huge number of leaf features and selecting a subset that best enhances the process of identification is challenging.

- **Lack of standard leaf datasets :** The design and implementation of a consistent automatic plant species identification system from leaf images requires a representative database that can be used by the machine learning algorithms to identify plants accurately. There is a lack of standard leaf image database that can be used for plant classification and therefore, the database is normally constructed by the researchers. Assembling such a database is time consuming and complex.

The demand for automated systems has led to the development of several techniques which have revolutionized the area of automatic plant

classification. This increase in the number of techniques has given rise to the dilemma of deciding which of these methods possesses the best properties and potentials for effective recognition. This problem is of particular importance in the botany field where the distortion of information may lead to inaccurate diagnosis.

1.5. GENERAL PLANT RECOGNITION SYSTEM

Automated plant identification through leaf recognition is to teach a computer how to recognize leaves, and then use the obtained results to identify the plant species. Compared with other methods, such as cell and molecule biology methods, identification of plants based on leaf image is the most successful and proven method (Wu *et al.*, 2007). Sampling leaves and obtaining a photograph of them is convenient and viable due to the availability of digital cameras. The general process for plant classification through leaf recognition is given in Figure 1.4 and is explained below.

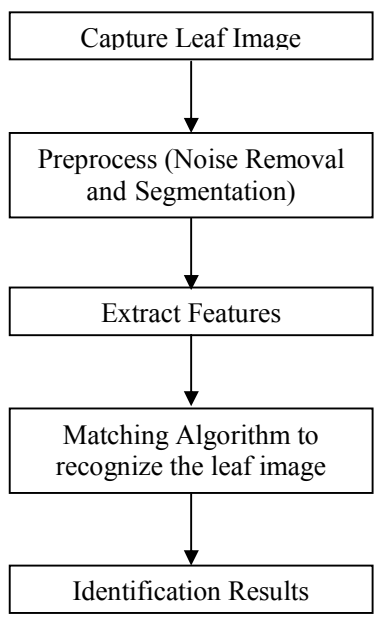


Figure 1.4 : General Architecture of Plant Identification System

Automatic plant classification process begins with the capturing of the leaf image and then performing enhancement of the image captured, extracting

the important features from the image and matching the extracted features with the leaf image database to identify the plant species to which the leaf belongs to. All the steps involved are considered very important for the accuracy and efficiency of the classifier and are explained in the subsequent sections.

1.5.1. Acquisition of Leaf Image

The first stage in any plant identification, through leaf recognition system is leaf database collection. The leaf images are acquired by scanners or digital cameras. As classification is a complex operation that needs high memory usage, a lossy compression format like JPEG / BMP / GIF is normally used.

1.5.2. Preprocessing

The preprocessing step involves two main tasks, namely, enhancement and segmentation. Real world input data always contains some amount of noise and therefore, enhancement techniques that reduce its effect are always desirable. Noise is defined as anything that hinders the identification and recognition system from fulfilling its respective task. Enhancement techniques also include operations that can improve leaf image properties which help to increase the overall performance of the identification system.

Image enhancement is an art and an algorithmic challenge, which uses image processing algorithms to improve the quality of an image. The goal of leaf image preprocessing is to increase the quality of the image and interpretability of the image data, so as to improve segmentation, feature extraction and classification processes.

The second part of preprocessing is segmentation, which subdivides the input leaf image into various parts of meaningful entities. Segmentation techniques use the various features extracted like gray or color features or texture features to separate the various regions of the input image. All these methods work on a common objective, that is, to provide a solution for

efficient automatic leaf image segmentation. Apart from these methods several researches were carried out in the past decades to find an optimal segmentation solution. But still the area is considered immature because of the various complex and changing features of the images. The main disadvantages of the existing segmentation techniques are

- The performance degrades when the image size is huge,
- Most of the techniques are resolution or context based,
- Accuracy of segmentation degrades when provided with noisy or degraded images and
- Most of the existing segmentation techniques do not meet the standard speed required by real time classification and recognition systems.

1.5.3. Feature Extraction

A successful leaf identification system requires a set of leaf features that best describe the leaf image and which can provide maximum discrimination between different leaves. The quality of a feature vector is related to its ability to discriminate examples from different classes (Figure 1.5). In the figure, it shows that examples from the same class should have similar feature values while examples from different classes have different feature values.

Figure 1.5 : Features - Distinction between Good and Poor Features

The feature extraction task consists of three sub-tasks as listed below.

(i) Feature construction,

(ii) Feature selection and

(iii) Dimensionality reduction.

Feature construction is one of the key steps in the data analysis process, largely conditioning the success of the subsequent machine learning endeavour. In particular, care should be taken not to lose important information at the feature construction stage. This task also combines the various existing features of a leaf image to form a feature vector.

Feature selection is the process of selecting relevant and informative features from the constructed features. Feature selection provides various advances like reducing the amount of storage space required and thus reducing the time complexity of the algorithm. A reduced feature set saves costly resources, improves identification accuracy and improves data understanding.

Dimensionality reduction is another form of feature selection that focuses on reducing the number of variables required to represent a leaf image. These techniques are included to handle the 'curse of dimensionality'. As the dimensionality of feature space increases, the accuracy increases but leads to sparseness of training data, which decrease the performance of classification. Thus, the design should seek a balance between the dimensionality and number of training vector to improve the performance of the classification algorithm used for leaf recognition.

1.5.4. Matching Algorithm

Together with feature extraction, the most crucial phase in the process of leaf recognition is classification. All the preceding stages should be designed and tuned for improving the success of this phase. The operation of the classification phase can be simplified as being a transform of quantitative input data to qualitative output information. The output of the classifier may either be a discrete selection of one of the predefined classes, or a real-valued vector expressing the likelihood values for the assumption that the pattern originated from the corresponding class.

Classification, also known as pattern recognition, discrimination, supervised learning or prediction, is a task that involves construction of a procedure that maps data into one of several predefined classes (Montejo-Raez, 2005). It applies a rule, a boundary or a function to the sample's attributes, in order to identify the classes. A classifier works to partition the feature space into decision regions that are identified, using pre-defined labels. An efficient classifier should be able to differentiate these partitions with precise decision boundaries (borders between decision regions) as shown in Figure (1.6).

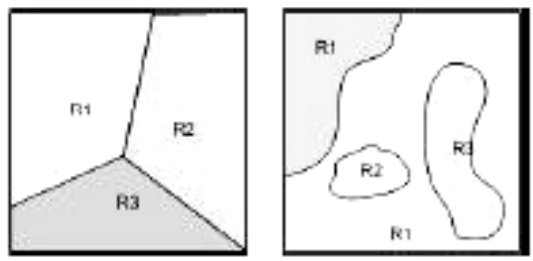


Figure 1.6 : Classifier and Decision Boundaries

The efficiency of a classification technique depends on various factors such as

(i) Whether learning method is a supervised or unsupervised method
(ii) Type of label output (binary or multiple) and
(iii) Whether they are statistical or non-statistical in nature.

Examples include Artificial Neural Network (Basheera and Haimeer, 2000), Decision Tree Classifiers (Jenhani *et al.*, 2008), Support Vector Machines (Steinwart, 2002), Naïve Bayes Classifiers and Rule-Based Classifiers, (Mencar *et al.*, 2011).

Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should satisfy two conditions as follows.

(i) It should fit the input data well and

(ii) It should correctly predict the class labels of records it has never seen before.

The primary goal of the learning algorithm is to build models with good generalization capability so as to accurately predict the class labels of previously unknown records. A basic classification model is shown in Figure 1.7.

Figure 1.7 : Basic Classification model

Supervised learning (machine learning) takes a known set of input data and known responses to the data and seeks to build a model that generates predictions for the response to new data. During the process of recognition, the known data (training features) has to be first collected. As mentioned previously, by selecting the appropriate set of features, the performance of the classifier can be improved. These training feature data are then used by the learning algorithm, to mimic the operation of the human brain. The learned knowledge is then applied to a new data (test features) for identification and recognition.

The input data for a classification task is a collection of feature records F (f_1, f_2 .., f_k where k is the number of dimensions) arranged as a vector (F, t). where F is the feature attribute and t is a special attribute called category or target attribute, designated as the class label. Thus, classification is defined as the task of learning a target function or classification model 'f(F, t)' that maps each feature set of F to one of the predefined category labels C (c_1, c_2, …, c_n where n is number of categories).

1.6. MOTIVATION

Regardless of its vast usefulness, many important species of plants are facing extinction. It is estimated that around 27,000 species become extinct each year, which amounts to an alarming number of about 3 per hour. Botanists need to identify near extinct plants, before climate change and development erase their living record. Reasons for extinction include global warming, introduction of exotic and hybrid species, habitat loss, disease, pollution, lack of knowledge and over exploitation. Out of these reasons, lack of knowledge contribute more, which necessitates the need for automated tools that can help botanists, researchers and students to identify plants from leaf images.

Manual process of identification requires botanists to have in depth knowledge of world's herbaria (Belhumeur *et al.*, 2008). This process is time consuming and in most of the cases, it can only be performed only by botanists with deep knowledge and who has specialized knowledge in plant taxonomy. Thus, automated tools that help to identify and classify plants are the current urgent need of the botanical field. Environment protection has triggered awareness on saving these plants, and botanists are discovering methods to protect them. Out of these methods, using leaves for identifying and recognition of plants has become the most important method.

India is an agriculture based country wherein seventy percent of the population depends on agriculture. Production of crops decreases, due to various types of pests and diseases affecting these agricultural products. Treatment of disease relies more on the type of plant, and therefore, recognition of plants from its own parts like leaves, stems and flowers have become vital. Presently, in most of the cases, farmers approach experts who are not readily available at all time.

Plant recognition is an area of research work that exists for the past few decades. Only now the use of computers and pattern recognition techniques to automatically recognize the plants from leaf images has been taken up.

Currently, most of the existing systems, used for this purpose, are applicable to certain species and requires human (botanist) intervention to define terms for feature extraction and pre-processing. This emphasizes the fact that the classification system should focus not only on the classification algorithm but equal attention should also be given to the other phases like pre-processing, feature extraction, etc. Further,the limited literature available pin points to the fact that the field is still raw and needs to follow a line of investigation to increase the accuracy and speed of classification.

According to Tzionas *et al.* (2005), designing a convenient and automatic recognition system of plants is important to facilitate fast classification of plants. Plant identification through leaf recognition is the area of research which is gaining maximum importance recently (Gwo *et al.*, 2013; Tilneac and Dolga 2010 and Cope,2012). When given a leaf image, these systems try to identify or classify the leaf, to a plant category. Automatic extraction and transferring of leaf features automatically to computer is a task that is both challenging and complex. Research in this field is an on-going process, where the goal is to identify and select only those characteristics of leaf that can best represent a plant and which can increase the accuracy of plant recognition.

1.7. PROBLEM STATEMENT AND OBJECTIVES

In order to solve the issues identified, the research problem is formulated as follows :

> **"***Given an input leaf image L_i and a template leaf database $L = \{L_1, L_2, \ldots, L_N\}$ having a set of leaves images belonging to different plant species, $P = \{p_1, p_2, \ldots, p_m\}$, the research problem is to use enhanced image processing and machine learning algorithms to find a match of L_i in L so as to recognize the corresponding plant species in P in an automated fashion that*

satisfies the three requirements, namely, high recognition rate, low error rate and high speed".

To solve the above problem statement, the primary objective of the research work is to design and develop a Computer-Aided Plant Identification through Leaf Recognition (CAP-LR) system that helps botanical industry to map a leaf image to a plant category using enhanced image processing and machine learning algorithms. To achieve this primary objective, the following sub-goals were formulated.

- To improve the quality of the leaf images using wavelet based denoising and edge enhancement algorithm,
- To perform texture based color segmentation of leaf images from its background using wavelet coefficients,
- To perform feature extraction techniques to identify and extract characteristics that best represent a leaf image and then use feature selection techniques to obtain a reduced feature sub-space and
- To identify a plant using two-level recognition algorithm to perform a matching process that maps a leaf image to a plant category.

1.8. CHAPTER FORMULATION

This chapter (**Chapter 1, Introduction**) presented an introduction to plant kingdom, automatic plant identification task along with research objectives. The rest of thesis is arranged as follows.

Literature review is a critical look at the existing research that is significant to the work and is carried out. A literature study of the various works proposed related to the research topics is presented in **Chapter 2, Review of Literature**.

As mentioned previously, CAP-LR consists of five steps, namely, noise removal, extraction of leaf from its background, feature extraction, dimensionality reduction and plant identification through leaf recognition. The overall methodology used in the design of CAP-LR along with a brief description of the various techniques used by each step of CAP-LR is presented in **Chapter 3** titled **Methodology**.

Chapter 4 (Design of Noise Removal Technique) presents leaf image enhancement technique that aims to improve the quality of the input leaf image. **Chapter 5 (Design of Segmentation Technique)** presents the techniques used to extract the leaf image from its background. Identification and recognition heavily depend on the feature extraction and classification methods used. C**hapter 6, Design of Feature Extraction and Selection** analyses the various methods used to enhance these processes.**Chapter 7, Design of Leaf Image Classification Models for Plant Identification**, describes the proposed classification models for leaf recognition and plant identification.

The CAP-LR system was evaluated using two template databases with different performance metrics with the aim of identifying efficient algorithms that enhance the overall process of plant identification through successful leaf recognition. The results of these experiments are presented in **Chapter 8, Results and Discussion**.

The research work is summarized and concluded with future research directions in **Chapter 9, Summary and Conclusion**. The work of several researchers are quoted and used as evidence to support the concepts explained in this thesis. All such evidences used are listed in the Reference Section of the thesis.

Sample leaf images of both standard and real dataset used during recognition are shown in **Appendix A**. The Preprocessed results of both standard and real datasets are shown in **Appendix B**. The segmentation results

of both standard and real dataset are shown in **Appendix C**. The feature set used during classification and recognition of standard and real datasets are shown in **Appendix D** respectively.

1.9. CHAPTER SUMMARY

Plant management consists of two tasks, "Know your plants and know how to manipulate them". To transform unsuitable habitat into suitable habitat is a power-packed phrase. With growing eco-friendly projects and medical requirements, it is necessary to obtain knowledge on plants correctly, and plant identification systems are much sought after techniques in these situations. The focal point of this research work is to identify plants through leaf recognition. This field amalgamates techniques from various image processing, pattern recognition and machine learning areas. In spite of researchers and industrialists concentrating on improving the performance of plant identification and leaf recognition tasks, it is still considered to be in its infant stage. In order to design an efficient CAP-LR system, it is important to understand the current status of research in this field and the information of such a literature study is presented in the next chapter, **Review of Literature**.