Statistical Machine Learning for Medicinal Plant leaves Classification

Abstract

Medicinal plants are usually identified by practitioners based on years of experience through sensory or olfactory senses. The other method of recognizing these plants involves laboratory-based testing, which requires trained skills, data interpretation which is costly and time-intensive. Automatic ways to identify medicinal plants are useful especially those that are lacking experience in herbal recognition. There is no standard mechanism in identification of medicinal plants. Therefore, we introduce an automatic approach based on statistical machine learning to identify medicinal plants. The main objective is to develop an automatic algorithm to classify medicinal plants using medicinal plant leaves. Leaf images are considered as they contain large number of diverse set of features such as shape, veins, edge features, apices, etc that are useful in identifying medicinal plants. Furthermore, leaves are relatively easy to obtain without damaging the plants. A database of leaf images of medicinal plants in Sri Lanka is not yet available. Hence through this research, we establish a repository of medicinal plant images. This repository is made available to the public through an open-source R software MedLEA, available at https://CRAN.R-project.org/package=MedLEA for research reproducibility. Researchers usually struggle and spend a lot of time establishing a database by gathering many leaf samples as raw data. By sharing our database we produce a training/test database to other researchers to evaluate their algorithm. The images were taken on a white background, positioning center of the white paper. Furthermore, the images are obtained from a normal smartphone without flash light to remove the shadow. This is useful when converting images to binary images to capture the shape accurately. We used non-diseased leaves that have simple arrangement. Furthermore, we used the leaves without petiole. In addition to this we use four benchmark open-source datasets to evaluate our algorithm. They are (i) flavia 1907 images collected from China, (ii) swedish 975 images collected from Sweden, and (iii) kaggle 1584 images collected from UK. We refer to our medicinal plant classification algorithm as MEDIPI : **MEDI**icinal **P**lant **I**dentification. The MEDIPI is divided into offline phase and online phase. The classification algorithm is trained in the offline phase. In the online phase, the pre-trained classification model is used to real-time leaf image classification for general users. Our classification algorithm operates on the features extracted from the image leaves. The offline phase of the algorithm contains four main steps: i) Image processing, ii) Feature extraction, iii) Label images, and iv) Trained a algorithm. The purpose of image processing is to improve the leaf image by removing undesired distortion. The main image processing steps are i) Convert original image to RGB image, ii) Gray scaling, iii) Gaussian smoothing, iv) Binary thresholding, v) Remove stalk, vi) Closing holes, and vii) Resize image. Feeding RGB images with gray scaling, optimize the contrast and intensity of images by reducing dimensions and complexity. Smoothing techniques are applied to remove noise and make the image less clear or distinct. Furthermore, as the result of binary thresholding is used to separate foreground from its background. Removal of stalk and closing holes in foreground is important when capturing the shape of the leaf. The second stage is to extract features from plant leaf images. We introduced 52 computationally efficient interpretable features to classify plant species. These feature are mainly classified in to four groups as (i) shape, (ii) color, (iii) texture, and (iv) scagnostics. Length, width, area, mean of red values, texture correlation, and monotonocity are some of them. Next, we trained our algorithm using random forest, gradient boosting, and extreme gradient boosting. The model trained with random forest algorithm provides the highest accuracy. Our algorithm works as a hierarchical classification system. The hierarchy contains 3 levels. The first level classifies images according to the shape. The second level classifies according to the edge types. The bottom level classifies the plant species. Furthermore, we used high dimensional visualization approaches to visualize what is happening inside the trained algorithm and provides transparency to our black-box model. We compare the accuracy of our proposed algorithm against several benchmarks and other commonly used algorithms for medicinal plants classification. The MEDIPI algorithm yields accurate results to the state-of-the existing techniques in the field. The algorithm is developed based on Python.

1 Introduction

Located in the tropics, Sri Lanka has a collection of plant species with various medicinal properties that have been consumed by generations as herbal treatments for control of diseases and to cure various medical issues. Traditional medicine system which has more than 3000 years of tested and proven efficacy, is still in use (Waisundara and Watawana 2014). It consists of Ayurveda, Unani, and Deshiya Chikitsa (Gunawardana and Jayasuriya 2019). Some of the diseases with complicated etiologies such as diabetes, arthritis, and cancer (for which a permanent cure is not in sight at present) (Waisundara and Watawana 2014) have been known to be completely controlled or cured using the traditional medicinal treatments alone (Devalaraja, Jain, and Yadav 2011). Various plant origins are used to treat disease conditions (Gunawardana and Jayasuriya 2019) in the traditional medicine system (Goyal, Kapil, and Kumar 2018).

According to the IUCN (International Union for Conservation of Nature) and the World Wildlife Fund, there are 550 medicinal plants in Sri Lanka. Furthermore, the distribution of medicinal plants is not uniform across the world and Sri Lanka is in the top 15 (see Figure 1).

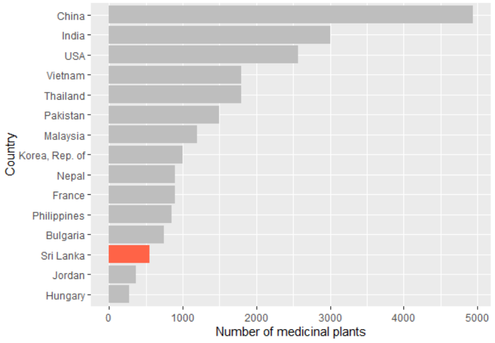


Figure 1: Top 15 countries in the world by distribution of medicinal plants (Source: [@inbook])

As shown in Figure 1, Asian countries like China, India, Nepal, Philippines, Malaysia, Thailand and North American countries like United States (USA) have a large collection of medicinal plants when compare with Sri Lanka. Even so, the percentages of medicinal plants of China, India, Nepal, Philippines, Malaysia, Thailand and United States (USA) are lower than Sri Lanka (see Figure 2). Furthermore, Sri Lanka is in the top 7 (see Figure 2).

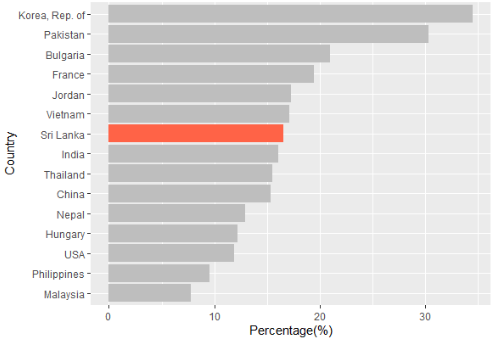


Figure 2: Top 15 countries in the world by percentage of medicinal plants (Source: [@inbook])

In past 10 years, Sri Lanka had a high demand for exporting medicinal plants and the value around 32 USD one hundred million (see Figure 3) which is an evidence that Sri Lanka has a good market in exporting medicinal plants around the world.

Furthermore, not only Asian countries but also European and North American countries have an interest of buying medicinal plants in Sri Lanka (See Figure 4). This is a proof that how much valuable and popular

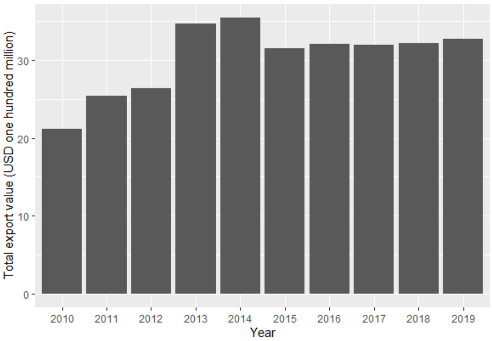


Figure 3: Distribution of export value of medicinal plants in Sri Lanka on last 10 years (Source: 2020, Trade Map - Trade statistics for international business development, https://www.trademap.org/Index.aspx)

of Sri Lankan medicinal plants.

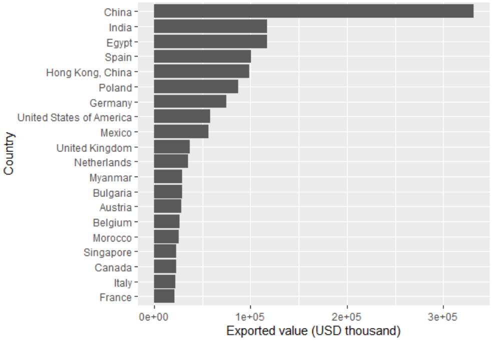


Figure 4: Top 20 exporters of medicinal plants in Sri Lanka on 2020 (Source: 2020, Trade Map - Trade statistics for International Business Development, https://www.trademap.org/Index.aspx)

Even though medicinal plants have a high demand around the world there is no standard mechanism in identifying medicinal plants.

Most algorithms use images as inputs to train the model. Hence, the quality of the image has a direct impact on its performance. Therefore researchers use built-in cameras of a mobile device (Waldchen and Mader 2018) or a special camera or a scanner to take photographs. Also most of the researchers use secondary datasets such as Flavia, Swedish etc (Goyal, Kapil, and Kumar 2018; Waldchen and Mader 2018). Due to less availability (Goyal, Kapil, and Kumar 2018) of adequate databases and the datasets contain few number of plant species (Goyal, Kapil, and Kumar 2018; Waldchen and Mader 2018), researchers tend to collect their own image datasets. There are restrictions while capturing the plant images. Single leaf, light illumination, shadow effect, and line of sight angle are few of them (Goyal, Kapil, and Kumar 2018).

Images of various parts as leaf, flower, bark, and fruit (Waldchen and Mader 2018) of the plant species use to train the model. Since leaf contains significant features, most of the researchers use to identify and classify the plant species in developing. Furthermore most of the researchers were focused on shape features (Goyal, Kapil, and Kumar 2018). But it is not sufficient to train reliable model properly. Therefore researchers more concern to find what are the most important features to classify plant species.

Furthermore, the existing algorithms mostly developed based on CNN, ANN, PNN, KNN, etc. These models require a large number of memories and become computationally prohibitive and hence, its usefulness can be limited. In addition to that while these methods can deliver good predictions their interpretability and transparency of the model is limited. We address these research gaps by proposing a image feature-based statistical machine learning algorithm.

Normally medicinal plants are grown in the backyards of houses and very little nurturing effort is required for their growth. They also have high growth rates. Therefore sometimes medicinal plants are considered as weeds (Waisundara and Watawana 2014). Most Sri Lankans are familiar with the traditional medicinal system and are even able to identify or administer the medicinal plants growing within their area of residence. Therefore, the locals can be observed consuming these medicinal plants to control a disease without the advice of a traditional medicinal practitioner, as they are familiar with the usage of these herbs because of the traditional knowledge, which has been passed down by their ancestors (Saslis-Lagoudakis et al. 2014) substantial botanical expertise is required by the manual identification process and it is also costly and time-consuming. This identification process is a very challenging task for the general public. There is also no standard mechanism in identification of medicinal plants.

Therefore by addressing the issues above, our main objective is to develop an automatic algorithm to classify medicinal plants by using statistical machine learning approach. To accomplish this main objective, we seek to achieve some other objectives.

A database of leaf images of medicinal plants in Sri Lanka is not yet available. Hence through this research, we establish a repository of medicinal plant images. Researchers usually struggle and spend a lot of time establishing a database by gathering many leaf samples as raw data. By sharing our database we produce a training/test database to other researchers to evaluate their algorithm. Leaf images are considered as they contain large number of diverse set of features such as shape, veins, edge features, apices, etc. Therefore through this research we identify features that are useful in classifying medicinal plants based on leaves images. Another objective is to develop an algorithm to extract and quantify leaf features. Furthermore, we used high dimensional visualization approaches to visualize what is happening inside the trained algorithm. We develop the proposed algorithm through an open-source software to identify medicinal plants in Sri Lanka by using leaf images.

The significance of this research is to avoid misidentifying medicinal plants in Sri Lanka. This is beneficial in conservation and ecological efforts. Researchers define that endangered medicinal plants as the plants which are facing a high risk of becoming extinct because they are either few in numbers, or threatened by changing environmental parameters (Nilaweera 2010). The International Union for Conservation of Nature (IUCN) has defined Threatened Herbal plants in three schemes as Critically Endangered, Endangered, and vulnerable. In the world, nearly 15,000 species of medicinal plants are now threatened. In Sri Lanka 280 plant species are threatened. According to the recent surveys (Nilaweera 2010), there are 1432 medicinal plant species in Sri Lanka, and out of the 100-200 species are threatened. Abarema begimena, Ashoka tree(Saraca Asoka), Beautiful Leaf(Calopyllum trapezifolium), Aglaia apiocarpa are few of them.

The algorithm developed by us is based on the leaf images. Since leaves are relatively easy to obtain without damaging the plants, there is no harm for the plants because of the development of algorithm. Our algorithm works as a hierarchical classification system. Therefore even though we don't know the exact species name, we can follow the first 2 levels. As the result of that misidentification rate and computation time will be decreased.

Outline should be written.

2 Methodology

2.1 Overview of the Algorithm

The aim of this chapter is to provide a general overview of the methodology used to develop our classification algorithm. The classification algorithm we introduce contains two main phases (i) The offline phase, and (ii) The online phase.

As shown in Figure 5, the workflow of the offline phase of the algorithm contains main 4 steps as:

1. Image Acquisition

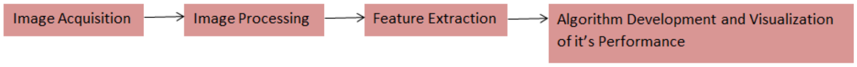


Figure 5: Workflow of the offline phase of the algorithm

2. Image Processing

3. Feature Extraction

4. Algorithm Development and Visualization of Algorithm Performance

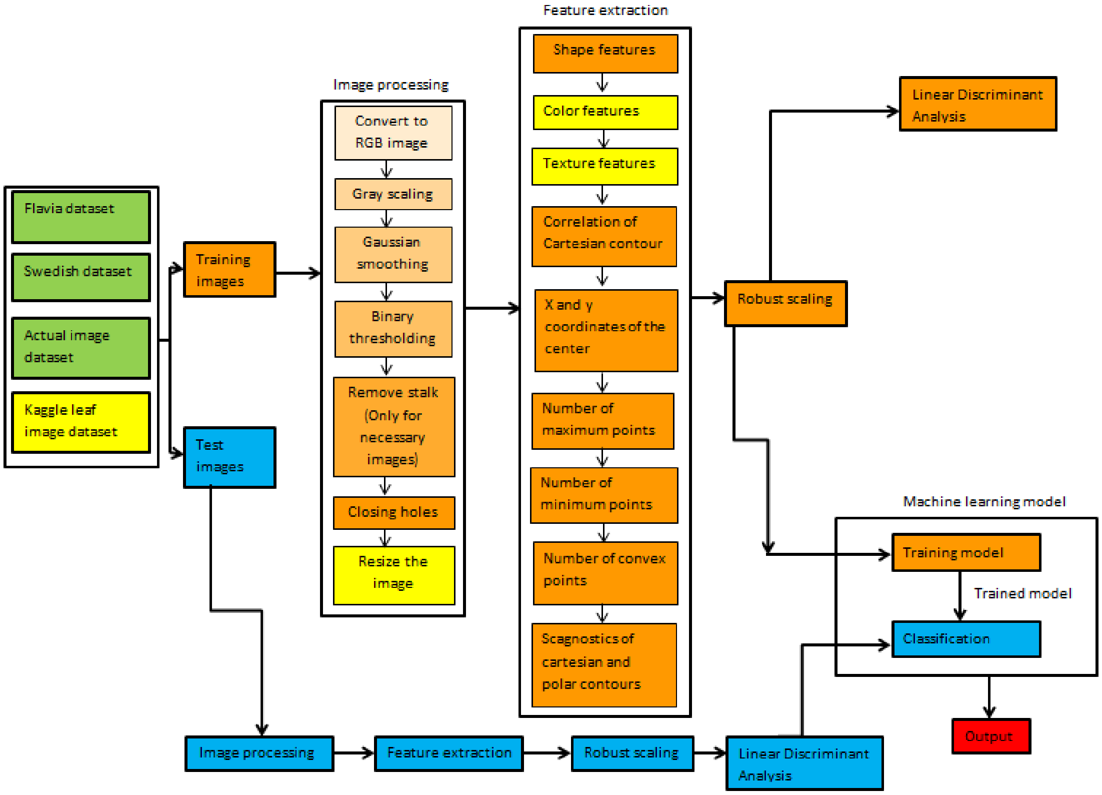


Figure 6: Methodology Diagram

Figure 6 shows the overview of the methodology that we followed. Online phase of the study is colored by orange and offline phase of the study is colored by blue. Firstly we acquire the images of leaves from existing datasets and the leaf image dataset that was collected by ourselves. Then each leaf image data set is divided as training and test images. Training image dataset was contained 80% of the images and test image dataset is contained 20% of the images from each leaf image dataset. We use four datasets to built and evaluate our algorithm. A brief summary of the datasets are given in the table 1.

Next step is image processing. As shown in Figure 6, the main image processing steps are Convert to RGB image, Gray scaling, Gaussian smoothing, Binary thresholding, Remove stalk, Closing holes and Resize image. Since Kaggle leaf image dataset contains only binary images, resizing step is enough as an image processing technique. We can follow remove stalk and closing holes technique only if the dataset contains leaf images with stalk and with holes (eg:- diseased leaves). After applying image processing steps, the images

Dataset	Image format	Total number of leaf images
Actual Leaf Image Dataset	Color	1099
Flavia dataset	Color	1907
Swedish Leaf Image Dataset	Color	975
Kaggle Leaf Image Dataset	Binary	1584

Table 1: Summary of datasets used in the algorithm

are ready to extract features. There are four classes of features: (i) Shape features, (ii) Color features, (iii) Texture features, and (iv) Scagnostics features of Cartesian and polar coordinates. In our research we also introduce some new features: Correlation of Cartesian contour, x and y coordinates of the contour, Number of minimum and maximum points, Number of convex points. Now the dataset contained all the features with the leaf image id. But Kaggle leaf image dataset doesn't have Color and Texture features. Robust scaling is applied to scale the data. To visualize the feature dataset with labels, Linear Discriminant Analysis is used. Our algorithm operates according to a hierarchical classification system. First the leaves are classified according to the shape such as; (i) diamond, (ii) simple round, (iii) round, (iv) needle, and (v) heart shape. The second level classifies according to the edge types. The bottom level classifies the plant species. Before training the model we labeled all leaves according to shape type, edge type, leaf arrangement, apex type, base type etc. These labels are identified by exploring "Ayurvedic Medicinal Plants of Sri Lanka", medicinal leaf repository maintained by, Barberyn Ayurveda resort and University of Ruhuna. The information we gathered by exploring the "Ayurvedic Medicinal Plants of Sri Lanka", medicinal leaf repository are made available through our R package MedLEA. Next step is to train the model for the training dataset by using machine learning techniques. This is a multi-class supervised learning classification problem. The trained model is used to predict labels in the the test dataset.

- MEDIPI

Our medicinal plant classification algorithm is defined as MEDIPI: **MEDI**cinal **P**lant **I**dentification. The MEDIPI is divided into offline phase and online phase. The classification algorithm is trained in the offline phase. In the online phase the pre-trained classification model is used to real-time leaf image classification for general users.

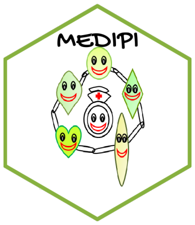


Figure 7: Hexsticker of MEDIPI

2.2 Data

We use four datasets. There is one primary dataset and three secondary datasets. The primary dataset is named as MedLEA. The secondary datasets are;

- MedLEA
- Flavia
- Swedish
- Kaggle

2.2.1 Secondary Data

- Flavia Leaf Image Dataset

The Flavia dataset contains 1907 leaf images. There are 32 different species and each have 50-77 images. Scanners and digital cameras are used to acquire the leaf images on plain background. The isolated leaf images contain blades only, without petiole. These leaf images are collected from the most common plants in Yangtze, Delta, China (Waldchen and Mader 2018). Those leaves were sampled on the campus of the Nanjing University and the Sun Yat-Sen arboretum, Nanking, China (Waldchen and Mader 2018). (https://sourceforge.net/projects/flavia/files/Leaf%2520Image%2520Dataset/)

- Swedish Leaf Image Dataset

The Swedish dataset contains 1125 images. The images of isolated leaf scans on a plain background of 15 Swedish tree species, with 75 leaves per species. This dataset has been captured as part of a joined leaf classification project between the Linkoping University and the Swedish Museum of Natural History (Waldchen and Mader 2018). (https://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/)

- Kaggle Leaf Image Dataset

This dataset consists of 1,584 images of leaf images of 99 species and 16 images per species. These leaf images were already converted to binary images. This dataset originates from leaf images collected by James Cope, Thibaut Beghin, Paolo Remagnino, & Sarah Barman of the Royal Botanic Gardens, Kew, UK. (https://www.kaggle.com/c/leaf-classification)

Sample of Flavia dataset

Sample of Swedish dataset

Sample of Kaggle dataset

Sample of MedLEA dataset

2.2.2 Primary Data

Image collection process contains 5 main steps as shown in figure 8. This approach is very simple, easy, and can be followed without any expertise knowledge.

Firstly we have to select a plant that we are going to use for this classification. Then have to find a leaf and pick it. In this step, have to be more careful about selecting the leaf. Our algorithm considers only

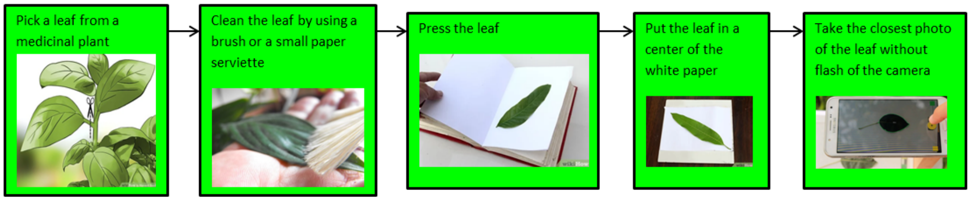

Pick a leaf from a medicinal plant
Clean the leaf by using a brush or a small paper serviette
Press the leaf
Put the leaf in a center of the white paper
Take the closest photo of the leaf without flash of the camera

Figure 8: Image collection process of medicinal plants in Sri Lanka

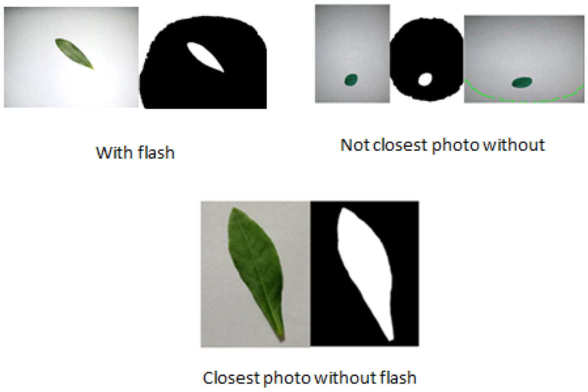
the leaf images without any diseases. When picking the leaf, use a scissor to pick the leaf without petiole. Because the algorithm considers only the leaf without petiole. Make sure that the leaf has to pick in the morning time. Because the leaf looks fresh in the morning time.

After picking the leaf, have to clean it by using a small brush or a piece of paper serviette. Because there are small water bubbles, soil seeds and mud patches.

In some cases, the leaf looks like rounding from the apex or base or margin of the leaf can't put on a flat surface. Therefore will be problematic when putting it to the algorithm. Because the algorithm is difficult to capture the shape of the leaf correctly. To avoid these problems, press the leaf approximately 1 or 2 days (In some cases less than 1 day is enough), before taking the photos.

Then keep the pressed leaf in a white paper. In this step, we have to consider about where we have to keep it. Make sure to keep the leaf in the centre of the white paper. The reason is that the converting to binary image work well when the leaf is in the centre of the white paper.

Finally when taking the photo, have to take the closest photo without the flash of the camera (see figure 9). Closest photo because algorithm is difficult to extract the contour of a very small leaf (see figure 9), decrease the amount of computational load that is exerted upon the graphic processing unit, and reduce the unnecessary foreground region (Goyal, Kapil, and Kumar 2018). When converting to the binary images, to capture the shape of the leaf correctly have to remove the shadow of the leaf much as can. Therefore by using the camera without flash, can remove the shadow (see figure 9). Make sure the photo is taken in the daylight to ignore the effect of light illumination.



With flash

Not closest photo without

Closest photo without flash

Figure 9

- MedLEA

Through this research, we establish a repository of medicinal plant images in Sri Lanka. This repository is made available to the public through an open-source R software MedLEA, available at https://CRAN.R-project.org/package=MedLEA for research reproducibility.

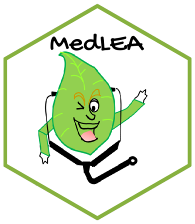


Figure 10: Hexsticker of MedLEA

There are 1099 images of leaf images of 31 species and 29-45 images per species of medicinal plants in Sri Lanka. These leaves have simple arrangement. A single leaf that is never divided into smaller leaflet units is know as a leaf with simple arrangement. That leaf is always attached to a twing by its stem or the petiole. The margins, or edges, of the leaf can be smooth, lobed, or toothed. The photos were taken from the device, Huawei nova 3i. The closest photos are captured on a white background.

2.3 Image Processing

Image processing plays a vital role in leaf image identification. Image processing is applied to reduce noise, background subtraction and content enhancement in the identification process (Goyal, Kapil, and Kumar 2018). The workflow we use to process images in this paper is shown in Figure 11. This includes seven main steps. They are: i) converting BGR (Blue-Green-Red) image to RGB (Red-Green-Blue), ii) gray scaling, iii) Gaussian filtering, iv) binary thresholding, v) remove stalk, vi) close holes, and vii) image resizing. Some of these steps applicable only for specific images. For example, apply remove stalk is applicable only to leaf images which has stalk.

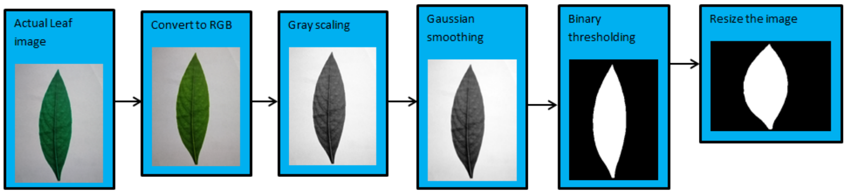


Figure 11: Image processing

Feeding RGB images with gray scaling, optimize the contrast and intensity of images by reducing dimensions and complexity. Smoothing techniques are applied to remove noise and make the image less clear or distinct. Furthermore, as the result of binary thresholding is used to separate foreground from its background. Removal of stalk and closing holes in foreground is important when capturing the shape of the leaf.

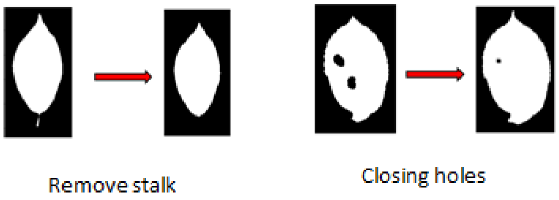


Figure 12

Figure 12 shows that binary image after removing the stalk and closing holes according to the order.

More details about the image processing steps are discussed in the Computer-aided Interpretable Features for Leaf Image Classification paper.

2.4 Feature extraction

Most crucial part is to extract distinctive leaf features from the images. Therefore most of the time research more focused on neural network models like CNN (Wu et al. 2007; Azlah et al. 2019; Herdiyeni and Wahyuni 2012) which are complicated and hard to understand what happening inside the algorithm. We introduced pre-calculate features which can be easy to interpret and generalize. They are also computational efficient. Mainly we focused on four types of features of leaf images as shape features, texture features, color features and scagnostics features. We identified altogether 52 features. More details about the features of the leaf are discussed in the Computationally Efficient Features paper. The following table shows the summary of all features.

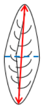
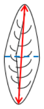
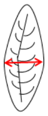
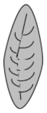
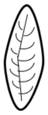
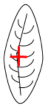
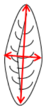
Image type	Feature category	Feature	Feature name	Figure	Formula	Range	Software	Software package
Binary	Shape	F_1	Diameter		$F_1 = max(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2})$; $\forall i, j, i \neq j$	[0,∞]	Python	combinations, numpy
		F_2	Physiological length		$F_2 =$ Length of the rectangle	[0,∞]		OpenCV
		F_3	Physiological width		$F_3 =$ Width of the rectangle	[0,∞]		OpenCV
		F_4	Area		$F_4 =$ Number of zero pixels covered by the contour	[0,∞]		OpenCV
		F_5	Perimeter		$F_5 = \sum_{i=0}^{n} d_i$; where n is the number of distances around the contour	[0,∞]		OpenCV
		F_6	Eccentricity		$F_6 = \sqrt{1 - \frac{b^2}{a^2}}$; where a is semi major axis and b is semi minor axis	[0,1]		OpenCV
		F_7, F_8	x and y coordinate of center					scipy.ndimage
		F_9	Aspect ratio		$F_9 = \frac{F_2}{F_3}$	[0,∞]		

Table 2 continued from previous page

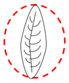
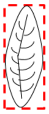
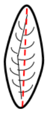
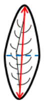
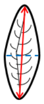
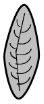
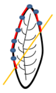
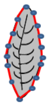
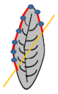
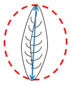
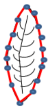
Image type	Feature category	Feature	Feature name	Figure	Formula	Range	Software	Software package
		F_{10}	Roundness/ Circularity		$F_{10} = \frac{4\pi F_4}{{F_5}^2}$	[0,∞]		numpy
		F_{11}	Compactness		$F_{11} = \frac{{F_5}^2}{F_4}$	[0,∞]		
		F_{12}	Rectangularity		$F_{12} = \frac{{F_5}^2}{F_4}$	[0,∞]		
		F_{13}	Narrow factor		$F_{13} = \frac{F_1}{F_2}$	[0,∞]		
		F_{14}	Perimeter ratio of diameter		$F_{14} = \frac{F_5}{F_1}$	[0,∞]		
		F_{15}	Perimeter ratio of physiological length		$F_{15} = \frac{F_5}{F_2}$	[0,∞]		
		F_{16}	Perimeter ratio of physiological length and width		$F_{16} = \frac{F_5}{F_2 * F_3}$	[0,∞]		
		F_{17}	Perimeter convexity		$F_{17} = \frac{\text{Perimeter of convex hull}}{F_5}$	[0,∞]		OpenCV
		F_{18}	Area convexity		$F_{18} = \frac{(\text{Area of convex hull} - F_4)}{F_4}$	[0,∞]		OpenCV
		F_{19}	Area ratio of convexity		$F_{19} = \frac{F_4}{\text{Area of convex hull}}$	[0,∞]		OpenCV
		F_{20}	Equivalent diameter		$F_{20} = \sqrt{\frac{4 * F_4}{\pi}}$	[0,∞]		numpy
		F_{21}	Number of convex points		$F_{21} =$ Number of vetices of the convexHull	[0,∞]		OpenCV

Table 2 continued from previous page

Image type	Feature category	Feature	Feature name	Figure	Formula	Range	Software	Software package
Gray scale	Texture	F_{22}	Contrast		$\frac{\sum_{a=1}^{columns}\sum_{b=1}^{rows}(a-b)^2 h(a,b)}{\text{Number of gray levels}-1}$	[0,∞]	Python	mahotas
		F_{23}	Entropy		$-\sum_{a=1}^{columns}\sum_{b=1}^{rows} h(a,b)log_2(h(a,b))$	[-∞,0]		
		F_{24}	Correlation		$\frac{\sum_{a=1}^{columns}\sum_{b=1}^{rows}(ab)h(a,b)-\mu_x\mu_y}{\sigma_x\sigma_y}$	[−1,1]		
		F_{25}	Inverse difference moments		$\sum_{a=1}^{columns}\sum_{b=1}^{rows}\frac{h(a,b)}{(a-b)^2}$	[0,∞]		
Color	Color	F_{26}	Mean red intensity value		$F_{26} = \frac{\text{Total intensity value of red channel of the image pixels}}{\text{Total intensity value of the image}}$	[0,∞]	Python	numpy
		F_{27}	Mean blue intensity value		$F_{27} = \frac{\text{Total intensity value of blue channel of the image pixels}}{\text{Total intensity value of the image}}$	[0,∞]		
		F_{28}	Mean green intensity value		$F_{28} = \frac{\text{Total intensity value of green channel of the image pixels}}{\text{Total intensity value of the image}}$	[0,∞]		
		F_{29}	Standard deviation of red intensity value		$F_{29} = \frac{\sqrt{\sum_{j=0}^{h}\left(\text{Red channel intensity}_j - \text{Red mean value}\right)^2}}{\text{Total intensity value of the image}}$; where h is number of pixels	[0,∞]		
		F_{30}	Standard deviation of blue intensity value		$F_{30} = \frac{\sqrt{\sum_{j=0}^{h}\left(\text{Blue channel intensity}_j - \text{Blue mean value}\right)^2}}{\text{Total intensity value of the image}}$; where h is number of pixels	[0,∞]		
		F_{31}	Standard deviation of green intensity value		$F_{31} = \frac{\sqrt{\sum_{j=0}^{h}\left(\text{Green channel intensity}_j - \text{Green mean value}\right)^2}}{\text{Total intensity value of the image}}$; where h is number of pixels	[0,∞]		
Binary	Scagnostics	F_{sc1}	Outlying		$F_{sc1} = \frac{\text{Total length of edges adjacent to outlying points}}{\text{Total edge length of minimum spanning tree}}$	[0,1]	R	binostics
		F_{sc2}	Skewed		$F_{sc2} = 1 - \text{weight} * (1 - qu_{skew})$;	[0,1]		
		F_{sc3}	Sparse		$F_{sc3} = \text{weight} * 90$ th percentile of the distribution of edge lengths in the minimum spanning tree where weight $= 0.7 + \frac{0.3}{1+\text{Number of vertex}^2}$	[0,1]		
		F_{sc4}	Clumpy		$F_{sc4} = max_j[1 - \frac{max_k[length(e_k)]}{length(e_j)}]$	[0,1]		
		F_{sc5}	Striated		$F_{sc5} = \frac{1}{\lvert Ve\rvert}\sum_{\nu\in Ve^{(2)}} I(\cos\theta_{e(\nu,a)e(\nu,b)} < -0.75)$ where $Ve^{(2)} \subseteq Ve$ and $I()$ be an indicator function	[0,1]		
		F_{sc6}	Convex		$F_{sc6} = \text{weight} * \frac{\text{Area of alpha hull}}{\text{Area of convex hull}}$ where weight $= 0.7 + \frac{0.3}{1+\text{Number of vertex}^2}$	[0,1]		
		F_{sc7}	Skinny		$F_{sc7} = 1 - \frac{\sqrt{4*\pi*\text{Area of alpha hull}}}{\text{Perimeter of alpha hull}}$	[0,1]		
		F_{sc8}	Stringy		$F_{sc8} = \frac{\lvert Ve^{(2)}\rvert}{\lvert Ve\rvert - \lvert Ve^{(1)}\rvert}$ Ve is the number of vertices	[0,1]		
		F_{sc9}	Monotonic		$F_{sc9} = r^2_{Spearman}$	[0,1]		

Table 2 continued from previous page

Image type	Feature category	Feature	Feature name	Figure	Formula	Range	Software	Software package
		F_{32}	Number of minimum points		$F_{32} =$ Number of global minimum points	$[0,\infty]$		
		F_{33}	Number of maximum points		$F_{33} =$ Number of global maximum points	$[0,\infty]$		
		F_{34}	Correlation of cartesian contour		$F_{34} = \frac{\sum_{i=0}^{m}(x_i-\overline{\mathrm{x}})(y_i-\overline{\mathrm{y}})}{\sqrt{\sum_{i=0}^{m}(x_i-\overline{\mathrm{x}})^2(y_i-\overline{\mathrm{y}})^2}}$	$[-1,1]$		

Table 2: Definitions of features

2.5 Classification framework

Our algorithm works as a hierarchical classification system. The hierarchy contains 3 levels. The first level classifies images according to the shape. The second level classifies according to the edge types. The bottom level classifies the plant species. If we use species names rather than hierarchical classification, it is difficult to manage and occur class imbalance problem due to large number of class labels. In order to develop hierarchical system we explore actual plant image repository: “Ayurvedic Medicinal Plants of Sri Lanka” (http://www.instituteofayurveda.org/plants/) which describes about most commonly used medicinal plants for practice of Ayurveda in Sri Lanka. This website was created as the result of the project implemented by Barberyn Ayurveda resort and University of Ruhuna. The website was updated on 11th May 2017. Our pilot study was based on this database.

We investigate 471 medicinal leaves in this repository. By investigating leaf images of each medicinal plant, we recorded the physical appearances (morphological characteristics) of the leaf. Rather than adding variables that regarded to physical appearances, we recorded Sinhala name (Local Name), Family name, and Scientific name as well. There were 22 variables with the primary key (unique) as “Id”. There were 18 variables that described about physical appearances of the medicinal leaf images.

3 Exploratory Data Analysis of Ruhuna Dataset

In this section, we present Exploratory Data Analysis to get an idea about the common morphological features of leaves.

Figure 13: Composition of Sample of Ruhuna Dataset by Arrangement of Leaves

According to the Figure 13, most of the leaves are arranged in Simple arrangement. Therefore further analysis, we selected the leaves that have simple arrangement.

According to Table **??**, most common shapes of the leaves are (i) Diamond, (ii) Simple round, (iii) Heart, (iv) Needle, and (v) Round. Therefore we used these common 5 leaf shapes as the first level of the hierarchy.

According to Table **??**, most of the leaves have smooth edges. Based on the results of the study we identify 4 common edge types as (i) Smooth, (ii) Toothed, (iii) Lobed, and (iv) Crenate.

As seen in Table **??**, most of diamond, heart, round, needle, and simple round shaped leaves have Smooth edges. All of the needle shaped leaves have smooth edges. There are no crenate edged heart shaped leaves. Furthermore, diamond, round, and simple round shaped leaves have smooth, toothed, lobed, and crenate edges. We used edge types to go to the bottom level of the hierarchy.

By conducting the study, we identified that what are the common shapes and edge types of the medicinal leaves in Sri Lanka. We proceed with 5 main shapes as diamond, heart, round, needle, and simple round. By

Figure 14: Classification hierarchy

using main 4 edge types as smooth, toothed, lobed, and crenate, we go to the bottom level of the hierarchy. Based on this dataset hierarchy (see figure 14) is created. Based on these information we develop a hierarchical structure to classify the leaf images by identifying main features. The general hierarchical structure is shown in Figure 15.

Figure 15: General hierarchical structure

4 Results

5 Discussion and Conclusions

Automatic medicinal plant species identification using leaf images is a popular research field with several critical applications. Through this research, we introduce an automatic algorithm to classify medicinal plants using medicinal plant leaves. Leaf images are considered as they contain large number of diverse set of features such as shape, veins, edge features, apices, etc that are useful in identifying medicinal plants.

In order to identify medicinal plant species using leaf images, we first do a preliminary study to get an idea about the morphological characteristics like shape, edge type, apex, base, arrangement etc. We identify five main shapes as: (i) Diamond, (ii) Simple round, (iii) Heart shaped, (iv) Needle, and (v) Round and four main edge types as: (i) Smooth, (ii) Toothed, (iii) Lobed, and (iv) Crenate by observing images in medicinal plant repository maintained by Barberyn Ayurveda resort and University of Ruhuna available at http://www.instituteofayurveda.org/plants/. Our observed results are converted into a open source R software package called MedLEA: **Med**icinal **LEA**f (https://CRAN.R-project.org/package=MedLEA). Furthermore, most of the researches are based on the existing databases like Flavia, Swedish etc. These existing databases contain few plant species and it is not sufficient to train a reliable model properly. In addition to that, a database of leaf images of medicinal plants in Sri Lanka is not yet available. Hence through this research, we establishe a repository of medicinal plant images which is available at MedLEA. We collect the leaf images by following simplest and reliable approach which can be followed without expertise knowledge. The images were taken on a white background, positioning center of the white paper and the images are obtained from a normal smartphone without flash light to remove the shadow.

Furthermore, we introduce our medicinal plant classification algorithm as MEDIPI : **MEDI**icinal **P**lant **I**dentification. The MEDIPI is divided into offline phase and online phase. The classification algorithm is trained in the offline phase. In the online phase, the pre-trained classification model is used to real-time leaf image classification for general users. Our classification algorithm operates on the features extracted from the image leaves. Through this research, we introduce 52 computer aided, interpretable features for leaf image recognition. There are four main categories of features that are used to classify leaf images. Many researches are based on shape, texture, and color features. In this research, we introduce new feature category called scagnostics for leaf image classification. Other than that correlation of cartesian coordinate, number of convex points, number of minimum and maximum points are introduced as new shape features. We explore the ability of features to discriminate the classes of interest under supervised learning and unsupervised learning settings using principal component analysis and linear discriminant analysis. Under both experimental settings clear separation of classes are visible in their projection spaces.

In addition to that, the offline phase of the algorithm contains four main steps: (i) Image processing, (ii) Feature extraction, (iii) Label images, and (iv) Trained a algorithm. The purpose of image processing is to improve the leaf image by removing undesired distortion. The main image processing steps are (i)

Convert original image to RGB image, (ii) Gray scaling, (iii) Gaussian smoothing, (iv) Binary thresholding, (v) Remove stalk, (vi) Closing holes, and (vii) Resize image.

Furthermore, we train our algorithm using random forest, gradient boosting, and extreme gradient boosting. The model trained with random forest algorithm provides the highest accuracy. Our algorithm works as a hierarchical classification system. The hierarchy contains 3 levels. The first level classifies images according to the shape. The second level classifies according to the edge types. The bottom level classifies the plant species. We observe that shape features like (i) x value of Center (cx), (ii) y value of Center (cy), (iii) Entropy, (iv) Perimeter ratio of length and width, (v) Diameter, (vi) Area convexity, (vii) Perimeter convexity, (viii) Narrow Factor, (ix) Area ratio convexity, (x) Physiological length, (xi) Physiological width, (xii) Rectangularity, and (xiii) Eccentricity are more important when classify the leaf images in the first level of the hierarchy. Scagnostic features like (i) Monotonic contour, (ii) Convex polar, (iii) Convex contour, (iv) Striated polar, (v) Striated contour, (vii) Skinny contour, and (vii) Skinny contour are more important in identifying leaf species in the bottom level of the hierarchy.

In addition to that, we use high dimensional visualization approaches as Linear Discriminant Analysis (LDA) to visualize what is happening inside the trained algorithm and provides transparency to our black-box model. We compare the accuracy of our proposed algorithm against several benchmarks and other commonly used algorithms for medicinal plants classification. The MEDIPI algorithm yields accurate results to the state-of-the existing techniques in the field. We have to use training/test from same dataset to get accurate results. Most of the literatures are based on shape feature. By train the algorithms (i) Only with shape features, and (ii) With all feature categories (Shape, color, texture, scagnostic), we observe that shape feature is not sufficient to classify leaf images.

Reference

Azlah, Muhammad, Lee Suan Chua, Fakhrul Rahmad, Farah Abdullah, and Sharifah Alwi. 2019. “Review on Techniques for Plant Leaf Classification and Recognition.” *Computers* 8 (October): 77. https://doi.org/10.3390/computers8040077.

Devalaraja, Samir, Shalini Jain, and Hariom Yadav. 2011. “Exotic Fruits as Therapeutic Complements for Diabetes, Obesity and Metabolic Syndrome.” *Food Research International (OTTAWA, ONT.)* 44 (August): 1856–65. https://doi.org/10.1016/j.foodres.2011.04.008.

Goyal, N., Kapil, and N. Kumar. 2018. “Plant Species Identification Using Leaf Image Retrieval: A Study.” In *2018 International Conference on Computing, Power and Communication Technologies (Gucon)*, 405–11.

Gunawardana, Shehara, and W. J. A. Banukie Jayasuriya. 2019. “Medicinally Important Herbal Flowers in Sri Lanka.” *Evidence-Based Complementary and Alternative Medicine* 2019 (May): 1–18. https://doi.org/10.1155/2019/2321961.

Herdiyeni, Yeni, and Ni Wahyuni. 2012. “Mobile Application for Indonesian Medicinal Plants Identification Using Fuzzy Local Binary Pattern and Fuzzy Color Histogram.” In *2012 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2012 - PROCEEDINGS.*

Nilaweera, D. D. 2010. “Journal of Machine Learning Research.”

Saslis-Lagoudakis, C. Haris, Julie Hawkins, Simon Greenhill, Colin Pendry, Mark Watson, Will Tuladhar-Douglas, Sushim Baral, and Vincent Savolainen. 2014. “The Evolution of Traditional Knowledge: Environment Shapes Medicinal Plant Use in Nepal.” *Proceedings. Biological Sciences / the Royal Society* 281 (February): 20132768. https://doi.org/10.1098/rspb.2013.2768.

Waisundara, Viduranga Y, and Mindani I Watawana. 2014. “The Classification of Sri Lankan Medicinal Herbs: An Extensive Comparison of the Antioxidant Activities.” *Journal of Traditional and Complementary Medicine* 4 (3): 196–202. https://doi.org/10.4103/2225-4110.126175.

Waldchen, Jana, and Patrick Mader. 2018. "Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review." *Archives of Computational Methods in Engineering* 25 (April): 507–43. https://doi.org/10.1007/s11831-016-9206-z.

Wu, S. G., F. S. Bao, E. Y. Xu, Y. Wang, Y. Chang, and Q. Xiang. 2007. "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network." In *2007 Ieee International Symposium on Signal Processing and Information Technology*, 11–16.