Statistical Machine Learning for Medicinal Plant Leaves Classification

Jayani P.G. Lakshika[1], Thiyanga S. Talagala[2]

[1] Department of Statistics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka

Introduction

Medicinal plant species identification is costly, time-consuming, and requires lots of effort, and expert knowledge. Recently, many researchers use deep learning methods like CNN (Convolution Neural Network) to classify plants directly using plant images. Even though deep learning models have achieved great success, their interpretability, and transparency of the deep learning models are limited. Also, deep learning models require a large memory space and become computationally prohibitive. To conquer this, we introduce an image feature-based statistical machine learning algorithm.

- **Main Objective**: Develop an automatic algorithm to classify medicinal plants by using statistical machine learning approach.

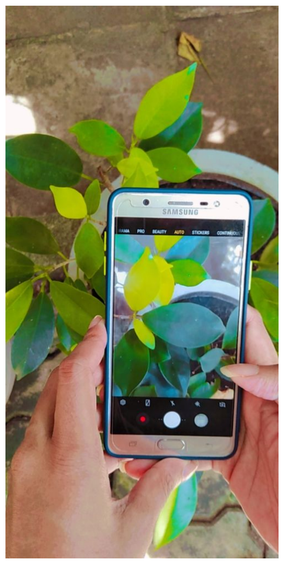

Figure 1: Taking a photograph of a leaf.

- **Significance of the study**: To avoid misidentifying medicinal plants.
- **Why leaf images? Leaf images** are considered as they contain large number of diverse set of features.

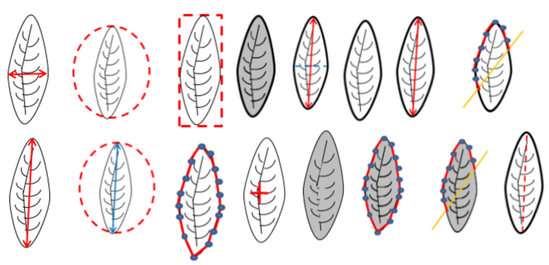

Figure 2: Shape-based features.

- Our classification algorithm operates on the features extracted from the image leaves.

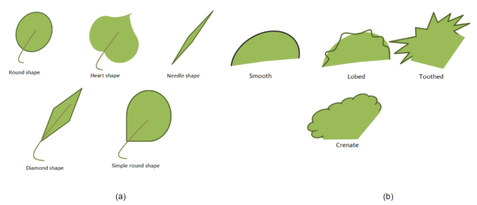

Figure 3: (a) Leaf shape (b) Leaf margin

Methodology

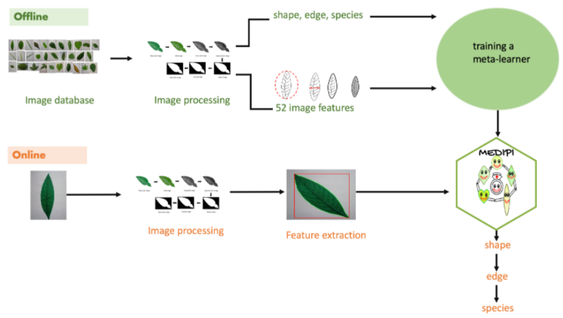

Figure 4: Overview of algorithm.

- **Offline phase** contains 4 main steps: i) Image processing, ii) Feature extraction, iii) Label images, and iv) Trained a algorithm.
- We trained our algorithm using random forest, gradient boosting, and extreme gradient boosting.

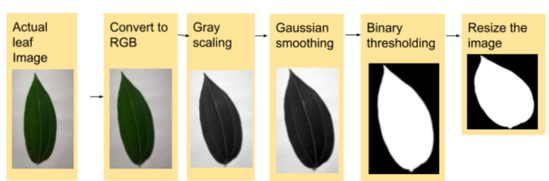

Figure 5: Image processing.

- **Online phase**, the pre-trained classification model is used for real-time leaf image classification.

Data

- Five leaf datasets of different countries (i) Flavia 1907 images collected from China, (ii) Swedish 975 images collected from Sweden, and (iii) Kaggle 1584 images collected from United Kingdom (UK), and (iv) MedLEA 1099 images collected from Sri Lanaka.

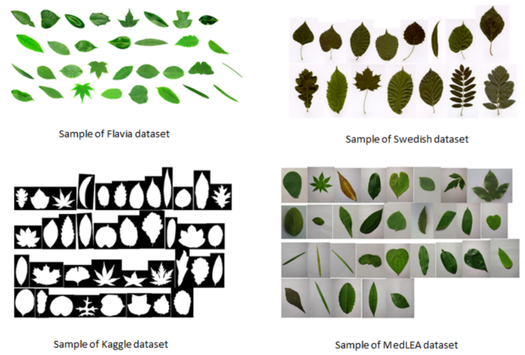

Figure 6: Datasets.

Visualization

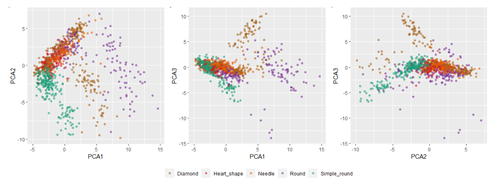

Figure 7: Distribution of Swedish leaf images on the projection space created based on principal component analysis. All projected points are coloured according to their shape labels.

We explore the ability of features to discriminate the classes of interest using PCA and LDA. Under both experimental settings clear separation of classes are visible in their projection spaces (see 3D view).

MedLEA

The MedLEA: **Med**icinal **LEA**f package is an open-source R software (https://CRAN.R-project.org/package=MedLEA) for research **reproducibility**. This repository provides morphological and structural features of 471 medicinal plant leaves and 1099 leaf images of 31 species and 29-45 images per species in Sri Lanka.

MEDIPI

Our medicinal plant classification algorithm is defined as MEDIPI: **MEDI**cinal **P**lant **I**dentification (https://github.com/JayaniLakshika/MEDIPI). The MEDIPI is divided into the offline phase and an online phase.

Features

We introduced 52 computationally efficient interpretable features that are useful in leaf classification (Lakshika and Talagala 2021).

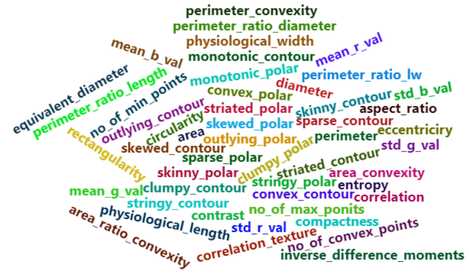

Figure 8: Features.

Conclusions

- 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.
- The model trained with random forest algorithm provides the highest accuracy.
- Our algorithm works as a **hierarchical classification system** which classifies according to shape, edge type (see Figure 3), and plant species. We observed that shape features are more important when classify the leaf images in the first level of the hierarchy. Scagnostic features are more important in identifying leaf species in the bottom level of the hierarchy.
- The **MEDIPI** algorithm yields accurate results to the state-of-the existing techniques in the field.
- We used high dimensional visualization approach as Linear Discriminant Analysis (LDA) to visualize what is happening inside the trained algorithm and provides transparency to our black−box model.

References

Lakshika, Jayani P. G., and Thiyanga S. Talagala. 2021. “Computer-Aided Interpretable Features for Leaf Image Classification.” http://arxiv.org/abs/2106.08077.