Grenze International Journal of Engineering and Technology, January Issue

InMAP: Dataset for Indian Medicinal and Aromatic Plants

V Nitesh[1], C K Roopa[2] and B S Harish[3]
[1-3]Department of Information Science and Engineering JSS Science and Technology University, Mysuru, Karnataka, India
Email: nikenitesh4@gmail.com, ckr@jssstuniv.in, bsharish@jssstuniv.in

***Abstract*— Medicinal and Aromatic plants play an important role in healthcare, agriculture, and the pharmaceutical industry, particularly in countries like India, which have rich biodiversity. Despite significant advancements in image processing and machine learning, the lack of publicly available, diverse, and high-quality datasets has limited the development of automated medicinal and aromatic plant identification systems. This paper introduces the Indian Medicinal and Aromatic Plant (InMAP) dataset, containing a comprehensive collection of high-resolution images representing a wide range of Medicinal and Aromatic species captured under natural conditions. The dataset is designed to address challenges such as varying lighting, complex backgrounds, and differences in leaf orientation. InMAP dataset aims to support and accelerate research in leaf segmentation, plant classification, feature extraction, and disease detection using traditional machine learning and deep learning techniques. The dataset not only contributes to technical advancements but also promotes the conservation and traditional knowledge of India's medicinal plant heritage.**

***Index Terms*— Medicinal Plants, Aromatic Plants, Image Processing, Classification, Machine Learning, Identification.**

I. Introduction

Plants have long been connected with human life, serving as a source of health through medicinal plants and offering well-being, lifestyle, and cultural practices through aromatic plants. Despite developments in the field of modern western drugs during the early 19th century, plant-based medicines remain one of the major sources of drugs in modern as well as traditional systems. Over 60% of all pharmaceuticals are plant-based. The origins of medicinal plant usage are deeply rooted in Indian traditions, documented in three different Vedas: Rigveda (2000 BC), Atharvaveda (2000 - 1000 BC), and Ayurveda (100 - 600 BC), which mentions many medicinal plants native to Indus valley region and their uses. The classical text Charaka Samhita (1000 BC - 100 AD), which was published at Varanasi, contains systematically documented plant-based treatments, diagnostic methods, and healthcare practices that laid the groundwork for traditional Indian medicine, which is being practiced to date.
However, there is a decline in the investigation of local medicinal plants after 1200 AD due to the invasion of Arabs, and the European invasion in 1800 AD onward resulted in the introduction of Western medicinal drugs. The origin and the use of aromatics stretch back through many civilizations but it is difficult to pinpoint when exactly humans started using aromatics, as the famous Chinese traveler Fa-Hien, who visited India during the rule of Chandragupta II has described India as the land of aromatic flowers, fruits, and woods.

Grenze ID: 01.GIJET.12.1.195

In the ancient texts dating back to the Mauryan period, information about gandhikas" can be found who created their blends of perfumes in various forms such as liquid, paste, and powder for trading [5]. In Sanskrit literature, traces can be found as to how a paste was formed using Sandalwood, sweet-smelling flowers, and leaves, which were part of a luxurious self-indulgence. Indian cities like Delhi, Lucknow, Mysore, Agra and Hyderabad have emerged as centers of national and international trade in perfumes and other aromatic-related compounds and were known for their quality across Asia, Europe and Africa [3].

The primary objective of creating an image dataset of Indian medicinal and aromatic plant species is to promote the use of Ayurvedic medicinal practices and the usage of aromatic plants to spread knowledge about commonly available medicinal and aromatic plants native to the Indian region. Increasing awareness in this domain can encourage researchers, educators, and practitioners to explore new challenges and opportunities in the field of medicinal and aromatic plants in combination with computer science. Furthermore, it can stimulate the cultivation and sustainable harvesting of medicinal and aromatic plants.

II. Data Description

Compared to existing publicly available datasets for medicinal and aromatic plants, the proposed dataset addresses several critical limitations observed in earlier collections. Many existing image datasets are captured under controlled conditions, having simple and uniform backgrounds, fewer lighting variations, and limited environmental complexities. Such idealized settings, while useful for initial model training, do not accurately reflect real-world challenges encountered during practical plant identification. This approach ensures greater diversity and complexity, making the dataset more representative of real-world conditions and enhancing its suitability for developing generalizable machine learning and deep learning models.

III. Experimental Design, Materials and Methods

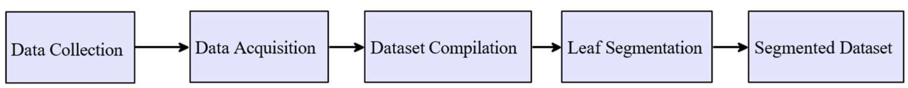

Figure 1. Workflow for medicinal and aromatic plant dataset creation and segmentation.

A. Data Collection

The dataset is created by collecting images from various botanical gardens in and around Bengaluru, Karnataka, India. Field visits were conducted at different times of the day, ensuring differences in lighting conditions and natural backgrounds. Plant species were identified in consultation with local botanists and labeled based on botanical nomenclature. Figure 2 highlights the images captured for the creation of the medicinal and aromatic plants dataset with complex backgrounds (multiple leaf images in the background), leaf shadow present in the image, or occultation of the leaf image and varied lighting conditions.

B. Data Acquisition

The dataset is created without plucking even a single leaf or damaging any plant. Leaves were photographed in their natural growing conditions, using mobile phones. Both indoor (greenhouse) and outdoor environments were utilized to capture the images in a wide range of illumination scenarios, including diffused and direct sunlight. More than 80 species of medicinal and aromatic plants have been identified and captured. Table 1 highlights the literature work on Indian medicinal and aromatic plants.

Figure 2. Samples of Leaf dataset

C. Dataset Compilation

After the data acquisition process, the collected leaf images underwent a systematic organization and cleaning process to prepare a valid dataset. The raw image files were first sorted with respect to plant species, with each species assigned a unique directory labeled using its scientific name to ensure consistency and reduce ambiguity. A quality control step is implemented to filter out images with significant issues such as motion blur, poor lighting, or occlusions. Only high-resolution images with clear visibility of leaves were retained. Additionally, duplicate or near-identical frame images were removed. Each image is further annotated with basic attributes such as leaf count.

D. Segmented Medicinal and Aromatic Leaf Dataset

To extend the usage of the collected dataset, a segmentation pipeline is implemented to isolate individual leaves from the background image. This is achieved using a hybrid segmentation model that integrates You Only Look Once (YOLO) [7] and SAM - Segment Anything Model [4]. The YOLO model is used for object detection and bounding box generation of leaf regions, and the SAM model is used for segmenting the leaf image. This model is used to identify and isolate each leaf within the images. Each input image is processed to generate leaf segmentation masks, which are then stored as images, with each leaf centered and extracted into a separate file. The segmented outputs are systematically organized in subfolders named after the original images.

(a) Input Image (b) Segmentation Output (c) Segmentation Output (d) Segmentation Output

Figure 3. Samples of Segmentation

TABLE I. PUBLICLY AVAILABLE DATASET

Reference	Dataset	Classes	Samples	Attributes
[1]	Medicinal Plant Dataset (Augmented)	30	13500	High resolution, uniform illumination, no background
[6]	DIMPSAR	80	6900	High resolution, uniform illumination, less complex background
[8]	Medicinal Leaf Dataset	40	1500	High resolution, uniform illumination, no background
[9]	MED117 Medicinal Plant Dataset	115	77,700	Video to frame, uniform illumination, white background
[2]	SIMPD Net: South Indian Medicinal Plants	20	2000	Mobile captured Images, High resolution

Table 1 highlights the publicly available datasets related to the medicinal and aromatic plants specific to the Indian region.

TABLE II. SUMMARY OF INDIAN MEDICINAL AND AROMATIC PLANT DATASET

SL No	Common Name	No. of Samples	SL No	Common Name	No. of Samples
1	All Spice (Pimento Tree)	31	2	Alovera	28
3	Amla Krishna	54	4	Amruthaballi	12
5	Ankole Mara	33	6	Annatto	36
7	Areca Nut	34	8	Ashoka	44
9	Ashwagandha	31	10	Ayyapana Plant	34
11	Babchi	32	12	Bergamot Mint	31
13	Betel Vine	39	14	Bhoomyamalakee	2
15	Black Pepper	38	16	Black Purification plant	47
17	Blood Flower	15	18	Bone Setter	42
19	Brahmi	38	20	Bryophyllum	44
21	Calabash Tree	48	22	Camphor Basil	31
23	Chives	28	24	Cinnamon	67
25	Clematis	20	26	Clocimum	35
27	Coleus	49	28	Cotton Lavender	24
29	Country Mallow	44	30	Crape Jasmine	28
31	Curry Leaf	32	32	Garden Rue	43
33	Geranium	3	34	Henna	24
35	Holy Basil	33	36	Honagone Soppu	34
37	Indian acalypha	53	38	Indian Indigo	3
39	Indian Liquorice	45	40	Indian Privet	41

41	Indian Spinach	36	42	Insulin Plant	29
43	Japanese Mint	33	44	Justicia	53
45	Kaadu Haralu	35	46	Kalmegh	37
47	Kokam Fruit	34	48	Krishna Tulsi	14
49	Lemon Balm	35	50	Lemon grass	12
51	Long Peper	41	52	Madhunashini	31
53	Makandhi beru	25	54	Malbar Nut	41
55	Marjoram	17	56	Mint	29
57	Multi Vitamin Plant	38	58	Noni	22
59	Nutmeg	43	60	Ocimum Basil	16
61	Origano	31	62	Patchouli	24
63	Pepper Mint	36	64	Red Priwinkle	34
65	Red Sanders	9	66	Red wood tree	31
67	Rose Mary	35	68	Safed Chitrak	32
69	Sage	39	70	Santolina (Lavender Cotton)	41
71	Sarpagandha	39	72	Scented Geranium	33
73	Shoe Flower	40	74	Silk Cotton	32
75	Simarouba	37	76	Spanish Mint	33
77	Spear Mint	35	78	Stevia	30
79	Swallow Wort	10	80	Sweet Bay	19
81	Thorn Apple	30	82	Thumbe	33
83	Thyme	18	84	Tincture Plant	39
85	Tooth Ache Plant	37	86	Tulsi-Holy-Basil	34
87	Turmeic	31	88	Uttharani	24
89	Varigated coleus	38	90	Vetiver	11
91	Vishnu Kraanthi	48	92	Wild Cumin	24
93	Wild Spinach	14	94	Wood Apple	35
95	Ylang-Ylang	32			

IV. EXPERIMENTAL RESULTS

A newly created dataset is evaluated empirically using various classifiers.

TABLE III. CLASSIFICATION RESULTS USING VARIOUS CLASSIFIERS

Classifiers	Train vs Test	Accuracy
Naive Bayes	50 vs 50	64.85
	60 vs 40	65.39
Decision Trees	50 vs 50	65.78
	60 vs 40	67.46
Logistic Regression	50 vs 50	66.32
	60 vs 40	67.24
K-Nearest Neighbors (KNN)	50 vs 50	62.40
	60 vs 40	64.59
Support Vector Machines (SVM)	50 vs 50	69.75
	60 vs 40	70.21
Neural Networks	50 vs 50	68.60
	60 vs 40	69.73

V. CHALLENGES IN DATASET CREATION

The creation of a real-world medicinal and aromatic plant dataset presented several unique challenges. One of the primary challenges involved capturing high-quality images in natural settings, which introduced uncontrollable variables such as inconsistent lighting, shadow formation, leaf motion due to wind, and background (such as overlaying soil textures). Unlike controlled lab conditions, field environments limited the ability to standardize image composition. Another major challenge was dealing with leaf shadowing, especially under direct sunlight or in dense greenery, which affected contrast and visibility, making the segmentation task more complex.

Another crucial step involved the accurate identification and labeling of plant species, which required the assistance of experts in the field of medicine and aromatic plants. The ambiguity resulted in visual similarity between a few species, especially in the partially grown leaves, which posed a major risk of mislabeling and required verification from experts. To maintain the quality and integrity of the dataset, many leaf images had to be removed due to motion blur, over-exposure, or lack of clarity.

Overcoming these challenges, the dataset has been created to depict various real-world scenarios and challenges such as varying backgrounds, multiple leaf objects in the background, non-uniform lighting conditions, indoor and outdoor leaf images being captured. These efforts in creating a diverse medicinal and aromatic leaf dataset help to enhance the robustness of models trained on the dataset, allowing for improved generalization and better performance in real-time applications.

VI. CONCLUSION

The InMAP dataset is a collection of diverse Indian medicinal and aromatic leaf images captured under various real-world conditions, unlike the existing dataset that has been created in controlled environment conditions, such as having uniform lighting and a simple background. In the InMAP dataset, we have addressed these challenges by capturing images in varied lighting conditions, complex or natural background scenarios, and multiple orientations. Through this dataset, we have tried to provide a more realistic representation of the real-world conditions that a model would experience when deployed in real-time tasks, with these complexities and challenges making the dataset practical for the application of machine learning and deep learning tasks.

Additionally, the dataset also contains a folder containing individual segmented leaf images of all the plant species that are part of the actual dataset. This process of segmenting individual leaves opens up new research opportunities supporting tasks such as precise leaf shape analysis, disease detection, species classification, and even the exploration of other morphological characteristics that are unique to each species. The InMAP dataset is designed not only to support the advancement of research in artificial intelligence and machine learning tasks, such as plant classification, recognition, but also to serve as a valuable resource for promoting and preserving the knowledge of medicinal and aromatic plants in India. By documenting the plant species that have been a part of Indian culture, health, and daily life for centuries. By providing better identification and documentation of plant species, the dataset supports sustainable agricultural practices and conservation efforts. This could lead to a deeper understanding of the ecological and pharmacological importance of plants, opening new areas for research and health solutions. In-MAP can serve as a resource for researchers, educators, and conservationists aiming to advance automated plant identification systems and contribute to the preservation of India's rich botanical heritage.

REFERENCES

[1] Medicinal plant dataset (augmented).

[2] Muthukumar Arunachalam, T. Gopu, Sabari Nathan, K. Uma, V. Srivatsav, and M. Thangaraj. Simpd net: A novel dataset for south indian medicinal plants in natural conditions. In 2024 IEEE Silchar Subsection Conference (SILCON 2024), pages 1–6, 2024.

[3] A.A. Farooqi and B.S. Sreeramu. Cultivation of Medicinal and Aromatic Crops. Universities Press, 2001.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollr, and Ross Girshick. Segment anything, 2023.

[5] Rakesh Kumar, Ajay K Singh, and Amit Kanawjia. E-content on medicinal and aromatic crops hfl 311 b.sc. (hons.) horticulture.

[6] Pushpa B R and N. Shobha Rani. Dimpsar: Dataset for indian medicinal plant species analysis and recognition. Data in Brief, 49:109388, 8 2023.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[8] Roopashree S and Anitha J. Medicinal leaf dataset. 1:17632, 2020.

[9] Parismita Sarma and Parvez Aziz Boruah. Med117medicinalplantleafdatasetnametable.4,2023.