Identification of Medicinal Plant Using Image Processing and Machine Learning

Abhishek Gokhale, Sayali Babar, Srushti Gawade(✉), and Shubhankar Jadhav

Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India
abhishekgokhale1599@gmail.com, sayali9779@gmail.com,
srushtigawade1999@gmail.com, shubhankarjadhav33@gmail.com

Abstract. Medicinal plants are the backbone of the system of medicines; they are the richest bioresource of drugs of traditional systems of medicine, modern medicines, nutraceuticals, food supplements, folk medicines, pharmaceutical intermediates, and chemical entities for synthetic drugs. These plants are classified according to their medicinal values. Classification of medicinal plants is acknowledged as a significant activity in the production of medicines along with the knowledge of its use in the medicinal industry. Medicinal plant classification based on parts such as leaves has shown significant results. An automated system for the identification of medicinal plants from leaves using Image processing and Machine Learning techniques has been presented. This paper provides knowledge of the process of identification of medicinal plants from features extracted from the images of leaves and different preprocessing techniques used for feature extraction from a leaf. Many features were extracted from each leaf such as its length, width, perimeter, area, color, rectangularity, and circularity. It is expected that for the automatic identification of medicinal plants, a web-based or mobile computer system will help the community people to develop their knowledge on medicinal plants, help taxonomists to develop more efficient species identification techniques and also participate significantly in the pharmaceutical drug manufacturing.

Keywords: Leaf recognition · Medicinal plants · Feature extraction · Image processing · Machine learning · OpenCV

1 Introduction

Medicinal plants have long been utilized in traditional medicine. Identification of medicinal plants is a very challenging task without external resources or assistance. Identification of the right medicinal plants that are used for the preparation of medicines is important in the medicinal industry. In various countries, there is a trend toward using traditional plant-based medicines alongside pharmaceutical drugs. Therefore, there seems to be immense potential in this field. Various kinds of algorithms are integrated into the application software. Image analysis is one important method that helps segment image into objects and background [1]. One of the key steps in image analysis is feature detection. Transforming the input data into the set of features is called feature extraction. The

B. Iyer et al. (eds.), *Applied Computer Vision and Image Processing*,
Advances in Intelligent Systems and Computing 1155,
https://doi.org/10.1007/978-981-15-4029-5_27

image processing nowadays have become the key technique for the diagnosis of various features of the plant [1].

The non-automatic method is based on morphological characteristics. Thus, classification here is based on the core knowledge of botanists. However, this non-automatic identification is tedious. Hence many researchers support this automated classification system and identification. There are a few systems developed so far where most of the processes are the same.

Following are the steps involved:

Step 1: Preparing the dataset.
Step 2: Preprocessing.
Step 3: Once the preprocessing is done, attributes have to be identified.
Step 4: Training.
Step 5: Classification of the leaves.
Step 6: Result evaluation (Fig. 1).

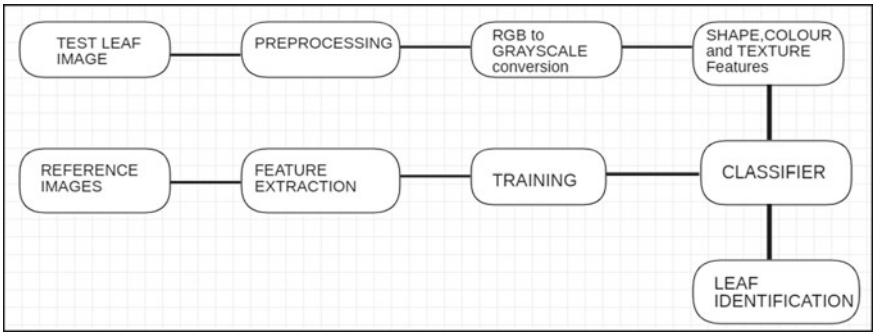

Fig. 1 Design of the proposed solution

This is an effective and efficient automated system (Fig. 1) that can be used by any student, pharmacist, or anyone from the non-botanical background. Motivation to undertake this project was given to us by an incident that happened with the head of the 'National Social Service' cell. He was trying to figure out a way to identify the medicinal plants correctly so that the villagers could make use of them for their pharmaceutical purposes. Seeing his difficulty gave us an idea of building this system.

2 Literature Review

See Table 1.

3 Methodology

The system will work in four stages:

Table 1 Literature review

Sr. No.	Research paper title	Year of publication	Accuracy
1.	Plant leaf recognition using a convolution neural network [2]	2019	94%
2.	Identification of Indian medicinal plant by using artificial neural network [1]	2018	75%
3.	Automatic recognition of medicinal plants using machine learning techniques [3]	2017	90.1%
4.	Plant identification system using its leaf features [4]	2015	More than 85%

A. Obtaining dataset.
B. Image segmentation/preprocessing.
C. Feature extraction.
D. Classification algorithm (Fig. 2).

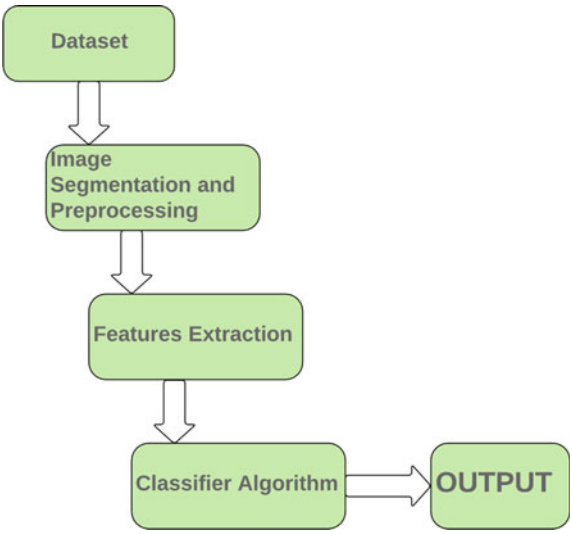

Fig. 2 Methodology overview

A. Obtaining dataset

Leaves are a feasible means to identify plants [5]. The image dataset used in this paper is Flavia leaves dataset which is obtained from http://flavia.sourceforge.net/. This image dataset consists of approximately 1900 image instances of leaves of 32 different species of plants. Sample images from one class are shown (Fig. 3).

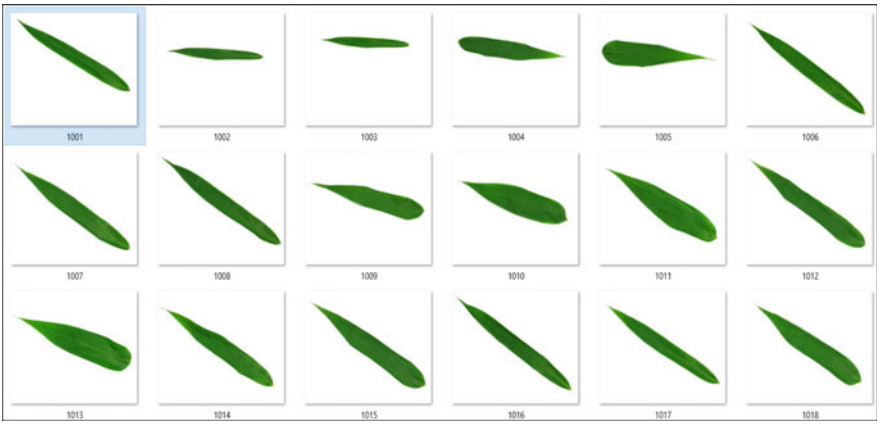

Fig. 3 Image dataset

For training and testing the model, a dataset was created using the extracted features of the leaf. The dataset was divided into two sets namely Train set (70%) and Test set (30%).

B. Image segmentation/preprocessing

Pixel values play a very important role in image analysis. They can be used to segment distinct objects. If there's a significant difference in the contrast values of the object and the image's background, then the pixel values will also differ. In this case, a threshold value can be set. Accordingly, an object or the background can be classified on the basis of the pixel values being less than or greater than a threshold value. This method is also known as Threshold Segmentation. It converts original image (Fig. 4) to grayscale (Fig. 5). If the image has to be divided into two regions, i.e., object and background, a single threshold value is defined. This is known as the global threshold (Fig. 6). If there are multiple objects along with the background, multiple thresholds need to be calculated. These thresholds are collectively known as the local threshold. This technique is preferred when there is a high contrast between object and background.

Fig. 4 Original leaf image

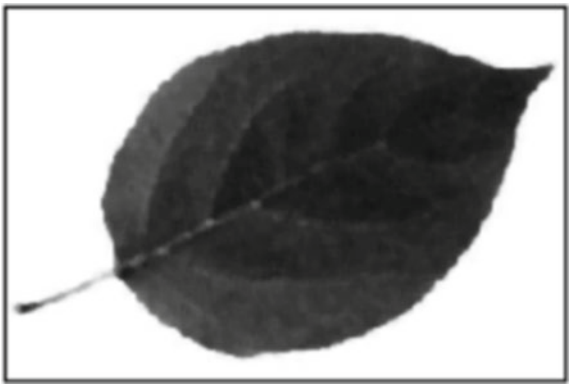

Fig. 5 Grayscale image

Two adjacent regions with different grayscale values are always differentiated based on the edge present between them. The discontinuous local features of an image can be considered as the edges. This discontinuity may prove to be helpful in defining a boundary of the object. This helps in discovering multiple objects present in an image along with their shapes. Filter and Convolutions are used in Edge detection. Edge detection is fit for images having better contrast between objects. When there are too many edges in the image and if there is less contrast between objects, it should not be used.

Digital image processing techniques are used for the classification of medicinal plants in the Plant Leaf Identification system. Firstly, all the images are preprocessed, for removing background area [6]. Then their features based on color, texture, and shape [7] are extracted from the processed image. The subsequent steps were followed for preprocessing the image.

(1) In this technique, we convert RGB to a grayscale image.
(2) After conversion, we smoothen the image using a Gaussian filter.
(3) Then Otsu's thresholding method is used for adaptive image thresholding (Fig. 7).
(4) Morphological Transformation is used for the closing of the holes.
(5) The last step for preprocessing is that the boundary extraction is done using contours.

C. Feature extraction

The major problem in image analysis arises due to the number of variables involved. These variables require a large amount of memory and computation. If the dataset is used as it is, it becomes less instructive and more redundant for doing analysis. When an algorithm has to process large datasets, then by applying this method, the dataset will be reduced to minimum dimensions. Extracting useful features from images in the dataset is the feature extraction process. Various types of leaf features were extracted (Fig. 8) from the preprocessed image which are listed as follows:

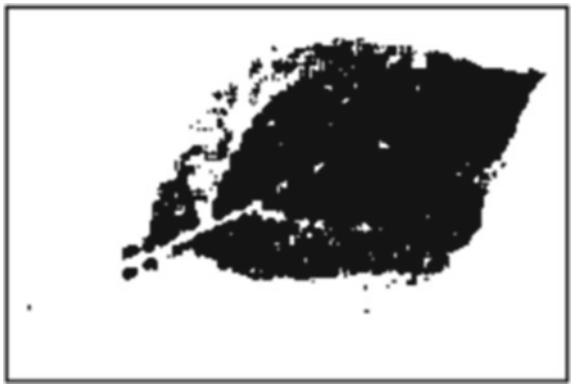

Fig. 6 Global threshold

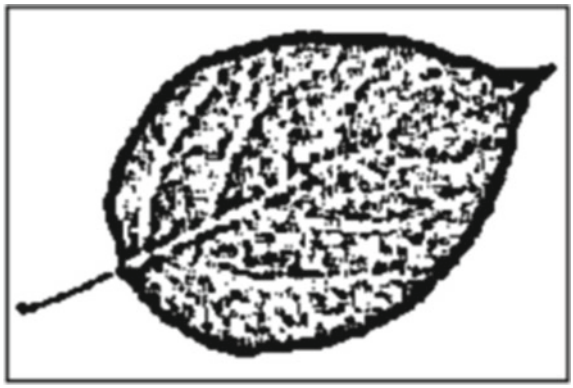

Fig. 7 Adaptive mean threshold

1. Features related to shape:

    - Length.
    - Width.
    - Total area.
    - Perimeter.
    - Proportional relationship between width and length (aspect ratio).
    - Rectangularity.
    - Circularity.

2. Features related to color:

   - The sum of channels divided by the number of channels of R, G, and B (mean).
   - Amount of variation of a set of values of R, G, and B channels (standard deviations).

3. Features related to texture:

   - The difference between the textures (contrast).
   - The similarities between the textures (correlation).
   - Inverse difference.
   - Entropy.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	area	perimeter	physiologi	physiologi	aspect_rat	rectangula	circularity	mean_r	mean_g	mean_b	stddev_r	stddev_g	stddev_b	contrast	correlatio	inverse_di	entropy
2	197484	3479.036	1416	759	1.865613	5.442183	61.28948	6.395667	13.64341	4.388007	24.02533	40.20093	21.44841	12.63969	0.997666	0.911738	1.688689
3	101248	2490.382	1190	130	9.153846	1.527931	61.25555	7.049316	9.232018	10.87607	33.81621	37.38222	46.47923	8.137424	0.997191	0.944818	1.193795
4	86570.5	2290.683	1095	119	9.201681	1.505189	60.61222	3.434303	6.371511	2.644757	19.9757	29.05737	19.27505	8.553729	0.99661	0.959023	0.848758
5	190214	2856.479	1318	254	5.188976	1.759976	42.89629	7.670415	13.3036	6.049157	28.82289	40.22185	26.9486	8.440064	0.998419	0.914331	1.673914
6	227727	2917.249	1324	286	4.629371	1.662798	37.3708	8.992028	16.67117	6.294281	30.96716	45.0402	28.59533	8.641447	0.998568	0.898644	1.968081
7	233724	3689.81	1434	953	1.504722	5.847076	58.25117	7.31924	15.73062	4.432931	24.81753	42.22132	21.53427	14.56146	0.997708	0.895654	1.957151
8	258395	3543.678	1396	874	1.597254	4.721856	48.59866	9.674849	18.43227	6.657215	28.70796	46.43166	24.42757	11.79534	0.998101	0.884639	2.126015
9	244401	3732.957	1479	912	1.621711	5.518995	57.01682	9.033226	16.53425	6.947844	29.5701	43.37798	28.11403	13.62784	0.997945	0.889018	2.070195
10	223690	3142.317	1404	388	3.618557	2.435299	44.14214	8.594646	16.13478	5.935715	28.13373	43.99787	24.5651	12.59348	0.997737	0.896627	1.891695
11	288344.5	3083.27	1329	450	2.953333	2.074082	32.96944	9.552538	19.34629	6.716408	27.72883	45.99022	24.64435	10.30296	0.998595	0.874376	2.286608
12	28444	726.8183	224	183	1.224044	1.441148	18.5721	73.99541	94.84775	37.14786	75.08527	88.4658	66.17187	499.0942	0.940857	0.455425	8.665403

Fig. 8 Feature set of different leaf samples

D. Classifier algorithms

Four machine learning classifier algorithms were applied to the data, which are as follows:

1. KNN (k-Nearest Neighbor) Algorithm.
2. Logistic Regression.
3. Naïve Bayes Algorithm.
4. SVM (Support Vector Machine).

These classifier algorithms were applied to the preprocessed data. The results are shown in Table 2. The Logistic Regression classifier achieves the best performance with an accuracy of 83.04% (Table 2).

However, due to resource constraints, for finding the highest accuracy the important parameters of every classifier were varied. The k-Nearest Neighbor (KNN) classifier gave the best accuracy of 79.49% (Table 3).

Apart from the accuracy, the performance was also assessed on a class proportion of leaves, for each class, that was accurately chosen from the entire set [3]. Precision here is the proportion of precisely identified leaves out of the total leaves that are predicted to be a specific plant while F-measure here is considered as the average of these two values [3].

Table 4 shown gives useful knowledge which can be used to test the strong aspects of the system and address its weaknesses. Plants that have low precision and recall

Table 2 Performance of machine learning classifiers

Sr. no.	Classifier	Accuracy
1.	SVM	82.69
2.	Logistic Regression	83.04
3.	Naïve Bayes	72.90
4.	KNN	81.99

Table 3 Performance of machine learning classifiers after cross-validation

Sr. no.	Classifier	Accuracy
1.	SVM	78.74
2.	Logistic Regression	78.85
3.	Naïve Bayes	71.23
4.	KNN	79.49

must be reassessed. For example, new features must be designed and extracted that give uniqueness in such leaves and are determinative of their species [3].

In Fig. 9, the confusion matrix is shown which is obtained when using the k-Nearest Neighbor (KNN) classifier with specified attributes in each iteration. The identification was successful which is indicated by the highest values in the diagonal line. Classes ranging from 0 to 31 represent the different 32 species of plants.

4 Conclusion

The main aim of this paper is to identify the medicinal plant from a given sample of a leaf. For this, we proposed an automated system for the identification of species of plants from leaves on the basis of their Color, Shape, and Texture features by using image processing techniques. Accordingly, the features were extracted from the Flavia image dataset, which consists of a total of 1907 images, and machine learning algorithms like SVM, Logistic Regression, Naïve Bayes, and KNN were applied. Accuracies of 82.69%, 83.04%, 72.90%, and 82.99% were observed, respectively. After cross-validation of the extracted features, the accuracies changed to 78.74%, 78.85%, 71.23%, and 79.49%, respectively. As a result, an inference was deduced from the observed accuracies that KNN would be best suited to the proposed solution. This system takes less processing time with increased accuracy for identification .

Table 4 Performance assessment of species using KNN classifiers

Class	Precision	Recall	F1-score	Support
0	0.92	0.61	0.73	18
1	0.76	0.94	0.84	17
2	0.92	1.00	0.96	22
3	0.93	1.00	0.97	28
4	0.89	0.96	0.92	25
5	0.88	1.00	0.94	15
6	0.71	0.85	0.77	20
7	0.85	0.73	0.79	15
8	0.60	0.50	0.55	12
9	0.78	0.93	0.85	15
10	0.79	0.83	0.81	18
11	0.72	0.76	0.74	17
12	0.77	0.71	0.74	14
13	0.67	0.67	0.67	21
14	0.93	0.88	0.90	16
15	0.25	0.06	0.10	17
16	1.00	1.00	1.00	26
17	0.88	1.00	0.94	22
18	0.95	0.90	0.92	20
19	0.85	0.94	0.89	18
20	0.74	0.93	0.82	15
21	0.89	0.57	0.70	14
22	1.00	0.93	0.97	15
23	0.71	1.00	0.83	17
24	0.90	0.56	0.69	16
25	0.71	0.67	0.69	15
26	0.88	0.88	0.88	16
27	0.82	0.90	0.86	20
28	0.83	0.88	0.86	17
29	0.87	0.91	0.89	22
30	0.73	0.79	0.76	14
31	0.92	0.73	0.81	15
Average	0.814	0.813	0.805	17.85

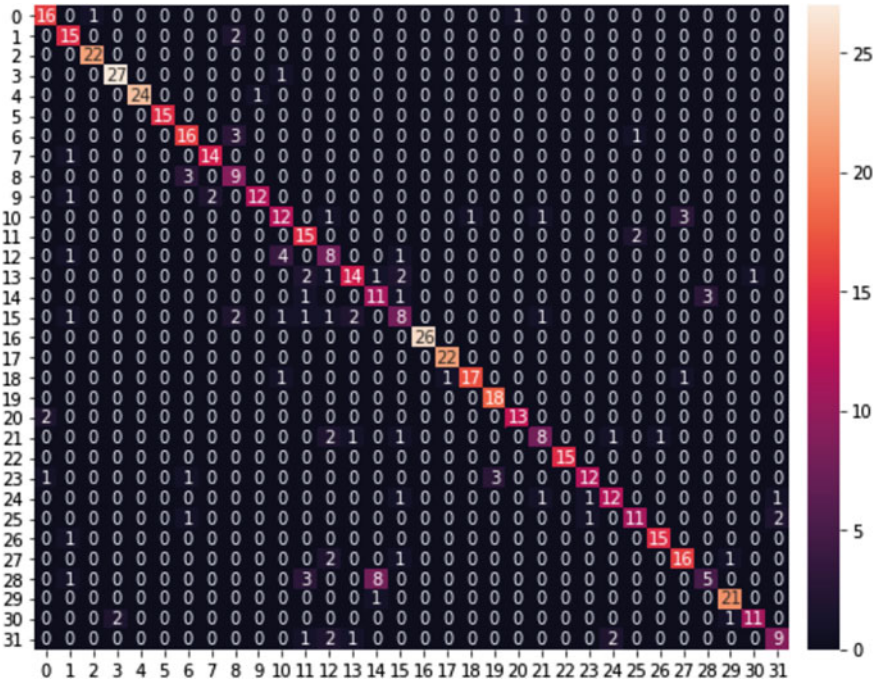

Fig. 9 Confusion matrix for KNN classifier

Acknowledgements. The authors would like to extend gratitude toward the faculty guide Dr. Anuradha Thakare and H.O.D Department of Computer Engineering Dr. K. Rajeswari for their constant support and guidance.

References

1. Aitwadkar, P.P, Deshpande, S.C, Savant, A.V.: Identification of Indian medicinal plant by using artificial neural network. Int. Res. J. Eng. Technol (IRJET) **5**(4), 1669–1671 (2018)
2. Jeon, W.-S., Rhee, S.-Y.: Plant leaf recognition using a convolution neural network. Int. J. Fuzzy Logic Intell. Syst. **17**(1), 26–34 (2017)
3. Begue, A., Kowlessur, V., Mahomoodally, F., Singh, U., Pudaruth, S.: Automatic recognition of medicinal plants using machine learning techniques. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **8**(4) (2017)
4. Nijalingappa, P., Madhumathi, V.J.: Plant identification system using its leaf features. In: International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (2015)
5. Khmag, A., Al-Haddad, S.A.R., Kamarudin, N.: Recognition system for leaf images based on its leaf contour and centroid. In: IEEE 15th student conference on research and development (SCOReD) (2017)
6. Sabu, A., Sreekumar, K., Nair, R.R.: Recognition of ayurvedic medicinal plants from leaves: a computer vision approach. In: Fourth International Conference on Image Information Processing (ICIIP) (2017)
7. Venkataraman, D., Mangayarkarasi, N.: Computer vision based feature extraction of leaves for identification of medicinal values of plants. IEEE International Conference on Computational Intelligence and Computing Research (2016)