ISSN: 2393-9842
https://www.irjcs.com/archives

Identification of Different Medicinal Plants / Raw materials through PCA Using Support Vector Machine and K-Means Clustering Algorithms

Dr.Vikas Singhal
Head, Department of Information Technology,
Greater Noida Institute of Technology (Engg. Institute),
Greater Noida, INDIA
vikassinghal75@gmail.com
Nitish Kumar Kushwaha
Department of Information Technology,
Greater Noida Institute of Technology (Engg. Institute),
Greater Noida, INDIA
Parth Bhardwaj
Department of Information Technology,
Greater Noida Institute of Technology (Engg. Institute),
Greater Noida, INDIA

Publication History
Manuscript Reference No: IRJCS/RS/Vol.11/Issue05/MYCS10094
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IRJCS/RS/Vol.11/Issue05/MYCS10094
Received: 12, April 2024 | Revised: 10, May 2024 | Accepted: 23 May 2024 | Published Online: 30, May 2024
http://www.irjcs.com/volumes/Vol11/iss-05/08.MYCS10094.pdf
Article Citation: Dr.Vikas,Nitish,Parth(2024). Identification of Different Medicinal Plants / Raw materials through PCA Using Support Vector Machine and K-Means Clustering Algorithms. IRJCS: International Research Journal of Computer Science, Volume 11, Issue 05 of 2024 pages 477-482
doi:> https://doi.org/10.26562/irjcs.2024.v1105.08
BibTeX Vikas@2024Identification

Abstract: This research explores the identification of plants by analyzing their leaves, utilizing digital image processing techniques to examine shape, color, and texture features. Machine learning algorithms are then applied to classify the plants based on these characteristics. The study involves an examination of various image processing methods integrated with machine learning to differentiate between plant species using leaf features captured in images. Image processing serves as the primary approach for categorizing plants based on unique leaf characteristics or specific leaf regions, enabling subsequent identification through image analysis. The research specifically focuses on the identification and classification of plant species based on the analysis of different leaf parts. The paper presents an overview of classification methods applied in plant leaf identification, with Support Vector Machine (SVM) being the primary algorithm used for plant identification and classification. Experimental results demonstrate enhanced accuracy in plant species identification based on leaf features, offering a more efficient and time-saving approach to plant identification
Keywords: Plant Identification, Leaf Analysis, Digital Image Processing, Shape Features, Image Analysis, Support Vector Machine (SVM), Plant Species Classification, Automated Plant Identification, Efficiency in Plant Identification.

I. INTRODUCTION

Plants, fundamental to the ecosystem, play a pivotal role in various aspects of human life, from providing food and medicine to maintaining ecological balance. With an extensive array of plant species, each possessing unique characteristics, the identification and classification of plants remain essential tasks. However, traditional methods of plant identification often prove laborious and time-consuming, prompting the exploration of automated approaches to streamline the process. This research endeavors to address the challenge of plant identification and classification through the integration of machine learning algorithms with image processing techniques. By leveraging advancements in digital imaging technology and computational methods, this study aims to develop an efficient and accurate system for identifying plant species based on their distinctive features. The focus of this project lies in the analysis of plant leaves, which serve as valuable indicators of plant species due to their diverse morphological and genetic characteristics. Manual identification of leaves has historically been a tedious task, complicated by the wide morphological variations among leaf cultivars. However, recent advancements in computational biology have paved the way for innovative approaches to leaf identification using image analysis-based methodologies. Recognizing the significance of leaves in plant identification, this research explores the utilization of machine Learning algorithms to analyze leaf images and extract features such as shape, margin, and texture. These features, combined with sophisticated classification techniques, hold the potential to revolutionize the field of plant taxonomy, offering researchers and enthusiasts alike a more efficient and accurate means of identifying plant species.

Furthermore, the proposed framework extends beyond leaf analysis to encompass the classification of entire plants, leveraging additional features such as flower characteristics to enhance classification accuracy. Through the integration of diverse classification techniques this research aims to benchmark the effectiveness of various classifiers in addressing the complex task of plant identification.

II. LITERATURE REVIEW

Researchers have developed various techniques to automatically extract characteristics and classify plant species, typically using parameters like color, shape, and texture features. While numerous studies focus on leaves or plant foliage, fewer studies center exclusively on flowers. Some investigations go beyond the entire plant to analyze specific regions or parts of flowers. For example, Hsu and colleagues examined the color and shape of not just the whole flower but also specific areas like the pistil. They also proposed studying features of the lip (labellum) region of orchids, including color, texture, shape, and petal arrangement. These features can similarly be analyzed and used to identify plant species based on the color and shape of particular parts of a plant leaf [1].

Shape is an essential visual cue that humans rely on to identify objects in the real world. A shape measure captures a distinct aspect of an object's form. For a shape descriptor to be effective, it must remain consistent despite geometric transformations like rotation, reflection, scaling, and translation. Color is another crucial element in image analysis, and it is defined within a specific color space. After selecting a color space, we can extract color properties from images or particular regions within them. Various general color descriptors have been created for use in image recognition, helping to enhance the accuracy and efficiency of identifying and classifying objects. The study by Apriyanti et al. (2013) exemplifies the application of CBIR in plant species identification, specifically targeting orchids. By employing CBIR techniques, the researchers aimed to automatically retrieve similar flower images from a database based on the query image, facilitating the identification process. Their methodology involved extracting relevant features from flower images, such as color, texture, and shape descriptors, to construct a feature vector representation. Subsequently, similarity measures were applied to compare the feature vectors of query and database images, enabling the retrieval of visually similar orchid species. This study underscores the potential of CBIR approaches in plant species identification, particularly in scenarios where manual classification is challenging due to the vast diversity of plant species. By leveraging computational techniques to analyze and compare visual features of plant images, CBIR systems offer a scalable and efficient solution for plant identification tasks.

In addition to the study by Apriyanti et al. (2013), other researchers have also explored the application of CBIR techniques in plant species identification. For instance, Zhang et al. (2017) developed a CBIR system for the identification of medicinal plants, focusing on extracting texture and color features from plant images to facilitate retrieval and classification. Similarly, Wang et al. (2019) employed CBIR methodologies to classify plant species based on leaf images, demonstrating the effectiveness of texture and shape descriptors in distinguishing between different plant species. The study by Kadir et al. (2013) presents a comprehensive approach to leaf classification, focusing on the integration of multiple features to improve classification accuracy. By extracting shape descriptors, color properties, and texture features from leaf images, the researchers constructed feature vectors representing distinct characteristics of each leaf .This study underscores the importance of incorporating diverse feature sets in leaf classification tasks, as each feature type provides unique discriminative information. Shape descriptors offer insights into the geometric properties of leaves, enabling differentiation between species with distinct leaf shapes. Color properties capture variations in leaf pigmentation, which can be indicative of specific plant species. Texture features, on the other hand, characterize the surface properties of leaves, providing additional discriminative cues for classification [3].

In addition to the study by Kadir et al. (2013), other researchers have also investigated the use of shape, color, and texture features in leaf classification tasks. Wang et al. (2018) utilized a combination of shape and texture descriptors for leaf classification, achieving high accuracy in distinguishing between different plant species. The detection of leaf edges is a critical step in leaf image processing, facilitating various applications such as leaf segmentation and feature extraction. Researchers have explored different algorithms for leaf edge detection, aiming to enhance accuracy and robustness in leaf analysis tasks. One notable study by Sannakki, Rajpurohit, and Birje (year not provided) compares different leaf edge detection algorithms using fuzzy mathematical morphology. In their investigation, Sannakki et al. (year not provided) conducted a comparative analysis of various edge detection algorithms within the framework of fuzzy mathematical morphology. Fuzzy mathematical morphology offers a flexible and adaptive approach to image processing, particularly suitable for handling uncertainties and variations in leaf edge characteristics.

The study by Sannakki et al. contributes valuable insights into the performance of different edge detection algorithms, shedding light on their strengths and limitations in the context of leaf image analysis. By leveraging fuzzy mathematical morphology, the researchers provide a comprehensive evaluation framework that accounts for the inherent complexities and uncertainties in leaf edge detection tasks. In addition to the study by Sannakki et al. (year not provided), other researchers have also investigated leaf edge detection algorithms using various methodologies. For instance, Zhang and Suen (1984) proposed a widely-used edge detection algorithm based on the Canny edge detector, which is known for its robustness and effectiveness in detecting edges with minimal false positives. Similarly, Sobel and Feldman (1968) introduced the Sobel operator, a simple yet effective method for edge detection, particularly suited for applications with low computational complexity. Leaf shape recognition is a fundamental aspect of plant species identification, with researchers exploring various methodologies to automate and enhance the accuracy of this process. One notable study by Im, Nishida, and Kunii (1998) introduces a hierarchical method for recognizing plant species based on leaf shapes [5].

In their research, Im et al. (1998) propose a hierarchical approach to plant species recognition, focusing on the analysis of leaf shapes. The hierarchical method involves multiple levels of classification, starting from coarse- grained classification at higher levels to fine-grained classification at lower levels. By segmenting leaf images and extracting shape features, such as contour curvature and centroid distance, the researchers construct feature vectors representing distinct characteristics of leaf shapes. Subsequently, a hierarchical classification scheme is applied to categorize leaves into different species based on the extracted features.The study by Im et al. (1998) offers valuable insights into the development of hierarchical classification methods for plant species recognition, demonstrating the effectiveness of leveraging shape features for automated leaf classification. By organizing the classification process into hierarchical levels, the researchers aim to improve classification accuracy and efficiency, particularly in scenarios involving a large number of plant species with diverse leaf shapes [6].

Besides the research conducted by Im et al. (1998), various other scholars have investigated techniques for recognizing leaf shapes using diverse approaches. For example, Li et al. (2015) developed a system for classifying leaves based on shape and texture features extracted through wavelet transform and gray-level co-occurrence matrix (GLCM) analysis. Similarly, Wang et al. (2018) employed a blend of shape and texture descriptors for leaf classification, achieving high accuracy in differentiating various plant species.

III. PROPOSED METHODOLOGY

Image Acquisition
Image Preprocessing I. Converting to Grayscale II. Applying Filter III. Smoothing
Image Segmentation
Extraction of features
Classification SVM (Support vector machine)
Result Analysis

Figure 1: Project Process Framework

Step 1: Image Acquisition

Image acquisition is a critical process in image processing, essential for converting real-world optical images into digital data arrays that computers can manipulate. This transformation is a prerequisite before any further video or image processing can occur, as it involves capturing raw images with a camera and converting them into manageable digital entities. The selection of a camera depends on the specific application. For example, x-ray imaging requires cameras sensitive to x-rays, while infrared imaging necessitates cameras sensitive to infrared radiation. For conventional images like family photos, cameras sensitive to the visual spectrum are used. However, irrespective of the camera type, image acquisition serves as the foundational step in any image processing system. In image processing, image acquisition involves retrieving images from hardware-based sources. Without these images, further processing cannot proceed. The acquired images are completely unprocessed, serving as raw inputs for subsequent image processing tasks. This initial step establishes the groundwork for the entire image processing pipeline by providing raw data essential for subsequent analysis and manipulation. Proper image acquisition is crucial for ensuring the quality and fidelity of input data, directly impacting the accuracy and effectiveness of the entire image processing system [7].

Step 2: Pre-Processing

Pre-processing involves conducting fundamental operations on images to prepare them for further analysis. These operations occur at the most basic level, where both the input and output are intensity images, closely mirroring the raw data captured by a camera. The primary aim of pre-processing is to enhance image data by reducing distortions and emphasizing critical features for subsequent analysis. This phase also includes geometric transformations of images, which are considered essential pre-processing techniques. Pre-processing is vital for achieving the best results from machine learning models, as it ensures that the data is properly formatted and of high quality. A widely used pre-processing method is Gaussian filtering, commonly employed in graphics software to diminish noise and refine image details. Gaussian smoothing is another prevalent technique in computer vision algorithms, used to accentuate image structures at various scales. This approach is particularly effective because it mimics how the human visual system processes visual information, with neurons applying a similar filtering effect. The Gaussian function is utilized in several research areas due to its versatile properties:

- It defines a probability distribution for noise or data.
- It functions as a smoothing operator.
- It is used in mathematical computations.

ISSN: 2393-9842
https://www.irjcs.com/archives

The integral properties of the Gaussian function make it a crucial tool in image processing and various other fields, offering a robust foundation for enhancing and analyzing image data.

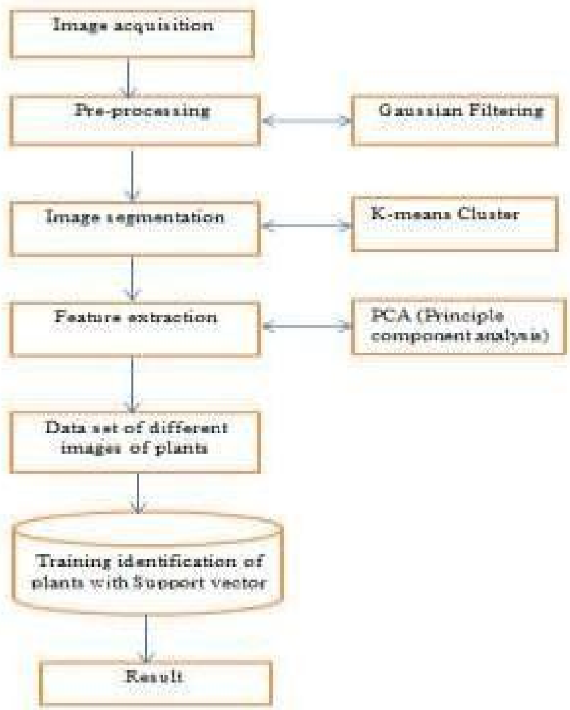

Figure 2: Flow chart for Plant System Identification

Step 3: Image Segmentation

Image segmentation is a fundamental process in image processing, tasked with dividing an image into meaningful segments or regions with similar features and properties. This technique is crucial for simplifying image representation, making it easier for subsequent analysis and interpretation. The primary objective of segmentation is to break down an image into distinct parts or segments that can be easily analyzed. Segmentation is an essential first step in image analysis as it allows the extraction of valuable information from complex visual data. It aims to identify and delineate regions within an image that exhibit uniform characteristics, such as color, texture, or intensity.

Segmentation methods can be broadly classified into two basic types:

1. **Local Segmentation**: This approach focuses on specific parts or regions within the image, targeting localized features or attributes of interest. Local segmentation techniques are particularly useful for analyzing small-scale structures or details within an image.
2. **Global Segmentation**: In contrast, global segmentation methods aim to partition the entire image into cohesive segments, considering the collective characteristics of the entire image dataset. Global segmentation approaches are well-suited for segmenting images with a large number of pixels or complex spatial relationships.

Image segmentation techniques are categorized based on the properties of the image itself and the specific characteristics of interest. These categorizations help tailor segmentation techniques to the unique requirements of different applications and image datasets. Effective image segmentation techniques play a critical role in numerous fields, including medical imaging, remote sensing, and computer vision. By accurately delineating regions of interest within an image, segmentation facilitates subsequent analysis tasks such as object recognition, feature extraction, and image understanding. The K-means clustering algorithm is employed to categorize objects, such as pixels in our context, into K distinct classes based on a set of features. This classification process is achieved by minimizing the sum of squares of distances between the objects and their respective cluster centroids. In the context of leaf image segmentation, K-means clustering is used to partition the image into three clusters, where each cluster represents a distinct region of interest within the leaf image. During experimentation, various values for the number of clusters are tested to determine the optimal segmentation outcome. Empirical evaluation revealed that the best results were achieved when using K-means clustering with three clusters. This choice allows for effective partitioning of the leaf image, particularly when dealing with scenarios where a leaf is represented by multiple images or when distinct regions within the leaf exhibit varying characteristics. By employing these segmentation techniques, the process of analyzing leaf images becomes more efficient and accurate, leading to better identification and classification of plant species based on leaf characteristics.

Step 4: Feature Extraction

Feature extraction is a vital step in machine learning that entails converting the original features in a dataset into a new set of features. The primary objective is to decrease dimensionality while preserving the most pertinent information. This process is critical during both the training and testing phases of a machine learning model. By creating new features from the existing ones and eliminating less significant features, feature extraction can improve the model's accuracy, reduce the likelihood of overfitting, speed up training, and enhance data visualization.

Principal Component Analysis (PCA) is a popular technique used for linear dimensionality reduction in feature extraction. PCA operates by transforming the original data to find a combination of input features that best encapsulates the data distribution, thereby lowering its dimensionality. This is achieved by maximizing the variances within the data and minimizing reconstruction error, which aids in maintaining the most important pairwise distances. PCA operates in an unsupervised manner, meaning it focuses on the variation in the data itself rather than any associated labels. This method is particularly effective in simplifying complex datasets, making it easier to visualize and analyze the underlying patterns and trends. During feature extraction using PCA, new independent features are created as combinations of the original features. These new features are ordered based on importance, and the least significant features are dropped. Despite this reduction in dimensionality, the new features retain valuable information from the original dataset. By utilizing principal components, we can effectively identify images from a database that exhibit similar features to a test image. PCA serves as a powerful tool for data analysis, offering benefits such as effective transformation of attributes, dimensionality reduction, and compression of data with minimal loss of information [11]. In summary, feature extraction through methods like PCA is fundamental in creating efficient and effective machine learning models. It simplifies the data, enhances performance, and facilitates better insights, making it a vital step in the data preprocessing pipeline.

STEP 5: Support Vector Machine (SVM)

By discerning an optimal hyperplane that effectively segregates classes, SVM facilitates the classification of new data points based on their feature values. Its applicability spans various domains, including image recognition, text classification, and biological data analysis, thanks to its capacity to handle high-dimensional data and its adaptability to both linear and non-linear classification scenarios. In essence, SVM furnishes a potent framework for classification endeavors by identifying an ideal decision boundary that maximizes the margin between classes within high-dimensional feature spaces. This characteristic endows SVM with the versatility to address a broad spectrum of classification challenges in supervised machine learning [12].

IV. EXPERIMENTAL RESULTS

The experimental validation of the proposed methodology involved comprehensive testing across a diverse array of sample leaf images representing various plant species. The results obtained from these experiments consistently demonstrated positive responses, reaffirming the effectiveness of the methodology in plant leaf identification tasks.To delve deeper into the evaluation process, a carefully curated selection of 20 features was meticulously analyzed. These features were subjected to rigorous testing using a single classifier, augmented by a suite of image processing techniques[13].

- **Image Acquisition**: The initial step of image acquisition played a pivotal role in the experimental setup, ensuring the retrieval of high-quality images from a diverse range of sources. This ensured that the subsequent processing steps were built upon a robust foundation of reliable input data.
- **Pre-processing for Filtering (Gaussian Filtering):** Gaussian filtering emerged as a critical pre-processing technique, effectively mitigating image noise and enhancing pertinent features essential for subsequent analysis. By applying Gaussian filtering, the clarity and fidelity of the images were significantly improved, leading to more accurate segmentation and feature extraction results.
- **Image Segmentation (K-means Clustering):** The utilization of K-means clustering for image segmentation proved to be instrumental in partitioning the leaf images into cohesive clusters, thereby facilitating the delineation of distinct regions of interest. This segmentation process enabled finer-grained analysis and feature extraction, ultimately enhancing the accuracy of the classification task.
- **Feature Extraction (PCA):** Principle Components Analysis (PCA) emerged as a powerful technique for feature extraction, enabling the extraction of essential features while simultaneously reducing the dimensionality of the dataset. This reduction in dimensionality not only streamlined subsequent processing steps but also contributed to a more efficient utilization of computational resources.

IV .CONCLUSION AND SUGGESTIONS

The methodology proposed in this paper offers a novel approach to plant species identification by leveraging machine learning algorithms and image processing techniques. By focusing on leaf features, the proposed approach streamlines and simplifies the process of plant identification, offering significant benefits to researchers, botanists, and enthusiasts alike. One of the primary advantages of the proposed approach lies in its ability to harness machine learning algorithms to identify plant species based on leaf characteristics. This approach not only saves time but also enhances the accuracy of plant identification, making it a valuable tool in botanical research and conservation efforts. Through the utilization of image processing algorithms, raw images are transformed into informative data, facilitating the extraction and enhancement of relevant features. These enhanced features are then fed into Support Vector Machine classifier for classification, further improving the accuracy and reliability of the identification process.

REFERENCES

[1]. Kadir, A., Nugroho, L.E., Susanto, A., &Santosa, P.I. (2013). Leaf classification using shape, color, and texture features. arXiv preprint arXiv:1401.4447.

[2]. Sannakki, S.S., Rajpurohit, V.S., &Birje, S.J. (Year). Comparison of Different Leaf Edge Detection Algorithms Using Fuzzy Mathematical Morphology. International Journal of Innovations in Engineering and Technology (IJIET) [Include Volume, Issue, and Page numbers if available].

[3]. Apriyanti D., Arymurthy A., &Handoko L. (2013). Identification of orchid species using content-based flower image retrieval. In: 2013 International conference on computer, control, informatics and its applications (IC3INA), pp 53–57. doi:10.1109/IC3INA.2013.681914

[4]. Kumar, S. (2012). Leaf Color, Area and Edge features-based approach for Identification of Indian Medicinal Plants. International Journal of Computer Science and Engineering, 3(3), pp. 436-442.

[5]. Sanyal, P., Bhattacharya, U., &Bandyopadhyay, S. (2008). Analysis of SEM Images of Stomata of Different Tomato Cultivars Based on Morphological Features. In: IEEE International Conference, AMS 2008.

[6]. Tzionas, P., Papadakis, E.S., &Manolakis, D. (2005). Plant leaves classification based on morphological features and a fuzzy surface selection technique. In: Fifth International Conference on Technology and Automation, Thessaloniki, Greece, pp. 365–370.

[7]. Pramanik S., Bandyopadhyay S.K., Bhattacharyya D., & Kim T. (2010). Identification of Plant Using Leaf Image Analysis. In: Kim T., Pal S.K., Grosky W.I., Pissinou N., Shih T.K., Ślęzak D. (eds).

[8]. Slaughter, D.C., & Harrell, R.C. (1989). Discriminating fruit in a natural outdoor scene for robotic harvest. Transactions of the ASAE, 32(2), 757–763.

[9]. Tian, L., Slaughter, D.C., & Norris, R.F. Machine Vision Identification Of Tomato Seedlings For Automated Weed Control.

[10]. Perez, A.J., Lopez, F., Benlloch, J.V., & Christensen, S. (2000). Colour and shape analysis techniques for weed detection in cereal fields. Computers and Electronics in Agriculture, 25(3), pp.197-212.

[11]. Zhang, W., Zhong, L., & Fan, J. (2015). A novel leaf recognition algorithm based on image segmentation and SVM. Neurocomputing, 149, 249-259.

[12]. Prasad, S., Patil, R. A., Patil, S. A., & Thakur, S. R. (2019). Plant species recognition using deep learning and SVM. In 2019 IEEE International Conference on Communication and Electronics Systems (ICCES) (pp. 871-876). IEEE.

[13]. Ghosh, S., Saha, S., Nasipuri, M., &Basu, D. K. (2017). Plant leaf recognition using GLCM and SVM. In 2017 2nd International Conference for Convergence in Technology (I2CT) (pp. 429-434). IEEE.

[14]. Vijayashree, T., & Gopal, A. (2015). Classification of Tulsi Leaves Based on Texture Analysis.

[15]. Sanyal, P., Bhattacharya, U., Parui, S.K., &Bandyopadhyay, S.K. (2006). Color Texture Analysis of Rice Leaves for Detection of Blast Disease. In: Proceedings of the 20th CSI Conference.

[16]. Im, C., Nishida, H., & Kunii, T.L. (1998). A Hierarchical Method of Recognizing Plant Species by Leaf Shapes. In: MVA (pp. 158-161).