The identification of edible plants using supervised machine models for traditional medicine

*Note: Sub-titles are not captured in Xplore and should not be used

1st Nduvheni Mulavhelesi
science faculty CSAM
University of the Witwatersrand
Johannesburg, South Africa
email address : 1664849@students.wits.ac.za

2nd Dr. Ritesh Ajoodha
science faculty CSAM
University of the Witwatersrand
Johannesburg, South Africa

3rd Dr.Shalini Dukhan
science faculty CSAM
University of the Witwatersrand
Johannesburg,South Africa

***Abstract*—The paper will discuss the significance of plant recognition using machine and deep learning to identify and classify edible plants using a kaggle data set containing approximately 62 wild plants. The plant identification models will assist traditional healers in South Africa in determining which plants they can use on they rituals.**

This research trained and tested six autonomic models for plant classification: Support vector machine, Convuntional neutral networks, Naive Bayes, Recurrent neutral Network, and Bayesian network'

In South Africa there are 200 000 traditional healers and more than half of the population consults with them, implying that they play an essential role in people's well-being. Using models to identify plants would boost their collection of muthi, which will also benefit the existing medical sector because African plants such as aloe, buchu, and devil's claw contribute or used in medical medicine.

***Index Terms*—Muthi, Autonomic models and supervised models**

I. Introduction

Before visiting a medical doctor more than 60% of people in South African rural communities seek advice and treatment from traditional healers and those who seek professional health care also consult traditional healers. The songoma uses a herbal mixture (Muthi) to heal their patients however they are unable to cure a number of diseases [9]. Healers could use the models to help them accurately identify a wide range of plant species for their collections, allowing them to find a cure for such diseases.

Traditional healer's knowledge of plants must be passed down from generation to generation in order for traditional rituals to continue and implementing a machine to preserve that information eliminates the possibility of human error. Since the plant species are getting extinct as a result of climate change .

Identify applicable funding agency here. If none, delete this.

If some of the plants used by traditional healers become extinct in the coming years the models will identify and assist traditional healers in locating similar plants that they can use in replacement of the plants that are in danger of extinction and also the way that they cultivate the roots and leaves is not sustainable. The models will also help traditional healers in reliably identifying a wide range of plant species for their collections.

In this research supervised machine learning models are used to assist traditional healers in selecting which plants they may use to make Muthi to heal their patience and to use in traditional ceremonies.

Subhankar Ghosh, H Kumar, and Jyothi S Nayak [5] investigated image classification using a combined classifier, which allowed them to use leaf morphological features on the following classification techniques: support vector machine, K-neutral neighbor, Probabilistic Neural Network, Decision Tree, Naive–Bayes classifier, and Learning Vector Quantization. They were able to demonstrate the advantages and disadvantages of models by combining them for feature selection and selecting the one with the highest output accuracy.

Jagadeesh D Pujari, Rajesh Yakkundimath, and Abdulmunaf S Byadgi [8] used probabilistic neutral network to identify and classify the crops or plants that are fungal affected. The method is developed to detect the early symptoms on plants that are affected. The algorithm is designed to process colour images after each feature is extracted using discrete wavelet transformation and principal component analysis to eliminate mistakes from the previous extraction. These enhanced features are utilized as inputs to a classifier and used to test and train the model.

Using automatic models the research demonstrate a considerable change in edible plant identification that can be used by traditional healers and also show between the

6 (six) models which one perform best to accurately identify and classify the plants .

The research contribute to current literature by being able to use about six different autonomic models in terms of identifying plants for traditional healers who are still using the traditional methods for then to be able to identify large amount of data at once while avoiding human errors.

The rest of the paper is organized as follows: chapter 2 is related work that some papers have done linked to my research. Chapter 3 shows the features that i have extracted and the methods on how i used them to extract those features and a small description of the models i have used in plant identification. The fourth chapter covers the results and discussion on how the models have performed.

II. Related work

This study seeks to determine which model best classifies different plant species with the least amount of error, nevertheless, [12] has stated that errors can occur when features are extracted. These are the writers who demonstrate the effects that have an impact on the accuracy. According to [12], there are concerns such as shifting leaf morphology with plant age and differences in overall shape due to leaf content. There are other mistakes that people make while photographing leaves since the leaf is captured with the background and the lighting of the camera which influences the quality of features extracted.

There are studies that illustrate the impact of using shape, color, or texture for plant recognition, possibly two or all of them. [14] shows a significant difference on accuracy when there were using K-NN classifier with only curvature feature which they obtain 71 % but when they use Curvature and Veins they were able to obtain 87% accuracy.

S Anubha Pearline, V Sathiesh Kumar, and S Harin [1] exposed the difference that makes by using certain type of data set. They were able to used Folio, Swedish leaf, Flavia and Leaf12. And they classify the plants using these models naıve Bayes, k-nearest neighbor and random forest. All these models gave out different results when there were train and tested on different data sets.

[Adams Begue, Venitha Kowlessur, Fawzi Mahomoodally, Upasana Singh and Sameerchand Pudaruth] [10] They conducted a study utilizing Random Forest with leaves from twenty-four distinct plants, and photographs were captured using a smartphone in a lab setting. After shooting the photo, they remove any shadows by applying HSV format which slit the image into different color saturation. The next step is to convert the image to a binary image that has two colours black and white then they extract the features like length, width, area of leaf, perimeter of leaf, hull area and perimeter, vertices, horizontal and vertical distance.Then 10-fold cross-validation technique was used to train the random forest for identification.

[Mohamed Elhadi Rahmani, Abdelmalek Amine, Mohamed Reda Hamou] used k-Nearest Neighbor (k-NN) on 100th plant leaves after pre-processing extract three distinct features which are margin, shape and inside texture of the leave, then we train the model by performing 10-fold cross-validation by separating randomly the data set into 10 columns in iteration exclusive folds and the accuracy of these models improves significantly by using all three features instead of just the margin, shape, or texture for classification.

On [Aakif et al] They pre-processed 817 distinct leaf photos from 14 distinct fruit trees and used morphological features and Fourier descriptions to derive the shape of the leaves. The data was then sorted and utilized to classify the leaves and tree types.

Artificial neural network was used by [vijay salti el al] and [Aakif et al] for both authors obtained good accuracy but used different method for the model.[Vijay salti el al] used publicly available leaf image flavai data set which leaves images are scanned or a picture is taken.The background of the picture is one colour and the leave does not have petiole.The Flavia data set, which features 1907 RGB leaf pictures of 33 plants was used for testing. After obtaining the leaves photo they undergo pre-prossesing to reduce the environment noise by grayscale conversion, binarization, smoothing, filtering and edge detection.

After pre-processing that is when they can extract features like colour which are calculated by the mean of all the columns and rows of RGB that we obtain in the processing stage and the we extract shape of the leave by using geometric features, morphological and tooth features.The extraction data is utilized to train the classifier and classify the plant based on prior knowledge, this includes training data set, general statistical categorization which is the process of determining a collection of categories or classes to which a new observation belongs. Categorization will refer to the process of assigning a specific plant species to an image based on its feature set.

On [Aakif et al] They pre-processed 817 distinct leaf photos from 14 distinct fruit trees and used morphological features and Fourier descriptions to derive the shape of the leaves. The data was then sorted and utilized to classify the leaves and tree types.

Support vector machine was tested by [14] using 28,046 environmental images of 109 plant taxa in Vietnam they tested four deep convolutional feature extraction models (MobileNetV2, VGG16, ResnetV2 and Inception) to see which one has the highest accuracy using the support vector machine. The best extraction method for plant recognition is mobileNetV2.

Table 1 : Papers Summaries				
Paper	models	Data	Features	Accuracy
Nguyen Van Hieu1 and Ngo Le Huy Hien	support vector machine	The Vietnamese plant image	shape, colour and Mobile NetV2	96%
Vijay S, Anshul S, and Shanu S	Artificial neural network	flavia database	shape, structure, colour, pattern and size	93.3%
Mohamed E.R, Abdelmalek A, Mohamed R.H	k-Nearest Neighbor (k-NN)	leaf12	margin, shape and texture	94.687%
Hossain and Amin (2010)	PNN	Flavia dataset	shape	91%
Adams B, Venitha K,Fawzi M ,Upasana S and Sameerchand P	Random forest	Lifeclef (2015)	hull area, perimeter,and vertices	90.1%

III. Methodology

In this study i have used six different automatic supervised models by comparing them to find out which one has the higher accuracy in identification of wild plants that are edible. Using kaggle data set which have 62 plants that was collected by gverzea off the internet. The issue with the data is that is not tagged correctly and not balanced. The data will be chosen randomly for training and testing the models. The photographs depict the leaves, stems, and flowers of the plants that bloom.

A. Leaf Pre-processing

Some models used a Pre processed data which is taking raw data so it can be more understandable computationally. This process will take a coloured picture to a gray scale picture to make it easy for it to be analyzed so we can extract geometric features and the texture of the leaf. I will use the RGB histogram for extraction of colours on the plants

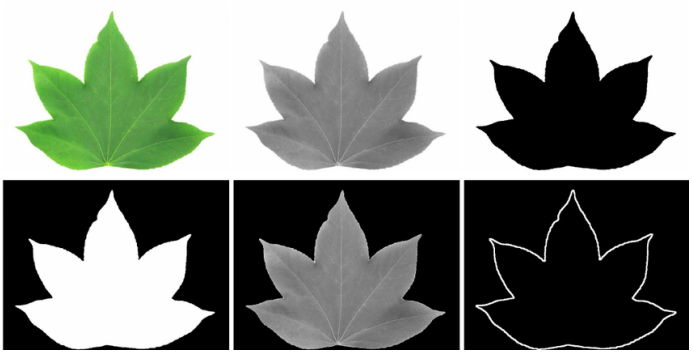

Fig. 1. pre processing on a leave

B. Feature Extraction

For leaves have extracted geometric features and texture because the colour of leaves is generally green which would not affect the results. For the plants with flowers i was able to extract the shape and color because flowers come in a variety of colours.

1) Geometric Features: 1) Equivalent diameter is a diameter of a centre of leaf with an equal overall cross section is calculate

$$ED = 2 * \sqrt{\frac{areaofaleaf}{\pi}}$$

2)solidity it shows the state firmness of a leaf

3)Eccentricity =

$$\sqrt{1 - (\frac{minoraxislength^2}{majoraxislength^2})}$$

4)Extent is an area covered by the leaf =

$$\frac{numberofpixelintheregion}{numberofpixelinaboundarybox}$$

5) Compactness is the ability to neatly fit all necessary components or features into a relatively small space.
6) Aspect ratio is the ratio between the height and width of the leaf
7)Entirety =

$$\frac{convexarea - area}{area}$$

8) Perimeter ratio of length and width =

$$\frac{P}{L + W}$$

2) Colour Extraction: Color extraction is useful for searching based on color tags and obtaining accurate results with a short computing time and high performance. Use the following methods for colour extraction:

1) Global colour histogram, it shows the frequency distribution of colour bins in an image. It counts comparable pixels and saves them which examines each statistical color frequency in a picture.

2) Histogram intersection, it is a method that it gives good results even though the background is not clean when comparing the two histogram distribution.

3)Colour correlogram, in contrast to a color histogram which captures merely the colour distribution in an image. It expresses the spatial correlation of colour changes with regard to distance change.

4)RGB histogram, it shows the distribution of primarily colour (green, red, blue) on an image

3) Texture Extraction: Texture is an important feature that can be designed to recognize items or find regions of interest in an image. Texture provides important information about the structural arrangement of surfaces. The spatial distribution is encoded in textural properties. I have used Gray Level Co-occurrence Matrix which is a matrix that calculates the frequency which pairs with pixels at a certain value and displacement appears in the image.

C. Models

I have used six supervised classification models to group or categorize data. To do so i have trained them on the data that i tested with. The data was tagged and balanced after that the models were able to identify the plants based on prior knowledge.The following models are the ones that i have used use for image classification:

1)Convolutional Neural Network is a type of neural network that is classified as deep neural networks which is used for visual classification. These models have three types of layers. The first layer is a convolution layer the second one is a pooling layer and third is fully connected layer. Before classification I will pre process the images one by one so that they all have the same resolution by turning color images to grayscale images and remove the background of the image so that the extraction of features can done easier.

I am going to divide my data set 60% for training , 20% for validation and 20% for testing the model. After pre processing the images i will insert it on the convolutional neural network model and this model will take every pixel and assign a number to it that will form a matrices.

The filter then generates a convolution movement together with the input picture, moving by one unit right along the image.The values are then multiplied by the original image values. All of the multiplied figures are combined together to get a single number. The process is repeated for the entire image, yielding a matrix that is smaller than the original input image. The other layers remove the negative values for greater accuracy then it is repeated for the other images then use pooling layer for identifying the max values.

The fully connected layer now adds an artificial neural network for use with CNN. This artificial network incorporates many features and improves in the prediction of image classes with improved accuracy.

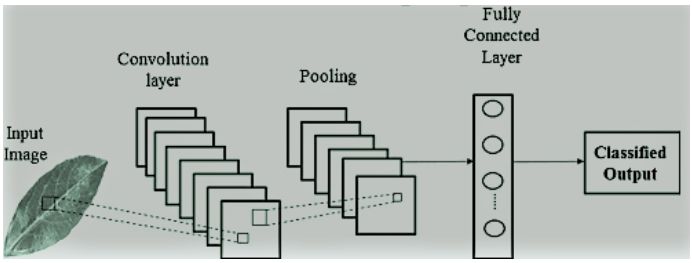

Fig. 2. Convolutional Neural Network

2) Support vector machine is the model that analyses data for identification and regression analysis it has a hyperplane that is used to dividing the edible plants for classification. In this model there is no need for pre-processing the images. I will use 1/4 data for training the the 3/4 for testing this model does not require large amount data for train .

Making $(x_i, y_i)i = 1)$ be a data set of N training samples, where x i is the ith portion in the input space x, and y_i+1, 1 is the class of $x_i$label. The decision function of Support Vector Machine that classifies a new test sample x can be shown as

$$\frac{1}{n}[\sum_{i=1}^{N} Max(0, 1 - y_i(w * x_i - b)] + \lambda||w||^2$$

We put data for training the model that will distinguish the edible wild plants into classes that are divided by the hyperplane and the distance between the closest edible plant on different classes should be equal that is called distance margin.

3) Naïve Bayes is a model that is based of the bayes theorem which state that the possibility of an event based on past knowledge of factors that may be associated with the event on image classification by making identification based on the features extracted independently. I will use 4:1 data for training and testing. With a vector $X = (x_1, \ldots, x_n)$ representing some n features that are independent, then we use these variables to calculate the probability

$$p(C_k|X) = \frac{p(x_k) * p(X|c_x)}{p(x)}$$

4) Logistic regression is statistical model that predicts the binary situation of whether the plant is edible or not. On this

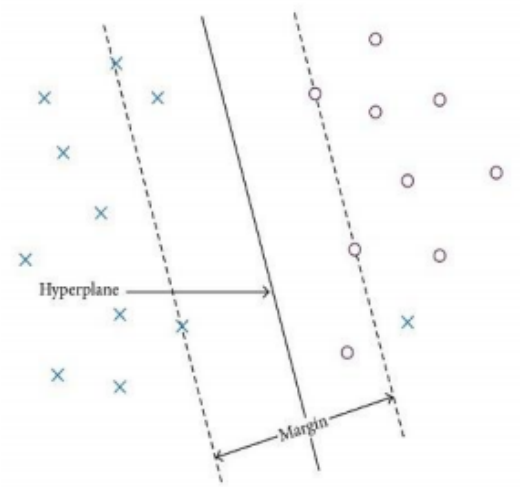

Fig. 3. hyperplane separation

researcher i will use 66% in training and 33% for testing the model. Which is the most known and simple method.That have (Y) the dependent variable and (X) the independent variable that have a linear relationship. Which is used to show how the independent variables have an influence on the dependent variables.

$$Y_i = \lambda_0 + \lambda_1 * X_1 + \ldots + \lambda_i * X_i$$

5) Bayesian Network is a form of probabilistic graphical model that employs Bayesian inference to compute probability, a set of variables represented as nodes on a directed acyclic graph (DAG). It depicts these variables conditional independence.

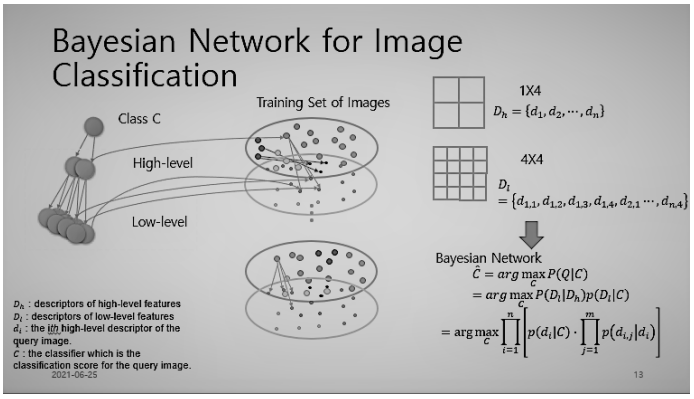

Fig. 4. Bayesian Network

6) Recurrent neural network is a model that takes the input data for training feature into hidden layers that make a sequence data. which is being used for prediction when is tested. For this model i will use 80% data for training the model and 20% data for testing it

$$h_t = f(h_{t-1}, X_t)$$

where :
h_t - a function of the previous time step and input in the current
time step x-input
x - input
f is nonlinear so the equation become

$$h_t = tanh(W_{hh} * h_{t-1} + W_{xh} * x_{[}t])$$

where:
W_{hh} weight of recurrent
W_{xh} weight of input

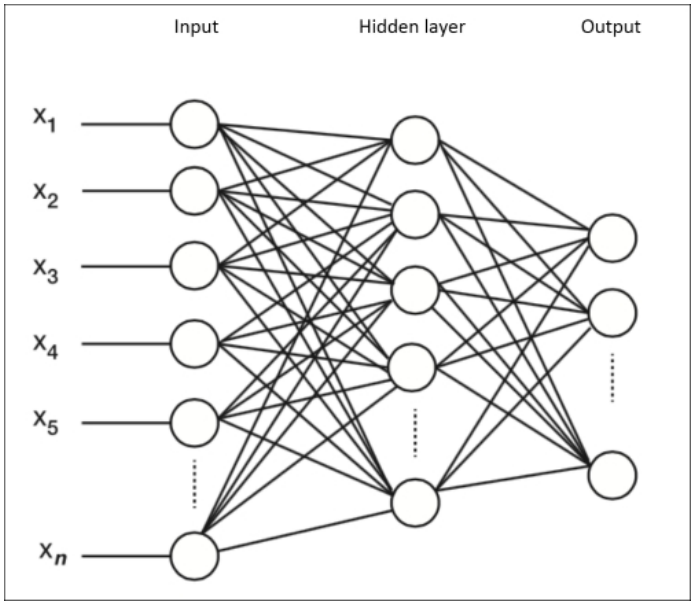

Fig. 5. RNN

D. Evaluation

1.To evaluate the classification of plants i have used accuracy, recall, f-measure and precision which were used to evaluate how the supervised machine learning model were performing by doing it increases the sensitivity of the plant identification models.

a) True positive(TP) is when the model predict the result is false but is true

b) True negative(TN) is a result in which the model predicts the negative class

c) False positive(FP) is when the model predict the result is false but is true

d) False negative(FN) is when the model predict the results to be true when is false

1.Accuracy, it shows the percentage of when the model correctly identify plant using training data knowledge.

$$\frac{TP+TN}{TP+TN+FP+FN}$$

2. Precision, it shows the percentage of when the model identifies if the plant is edible and is true.

$$\frac{TP}{TP+FP}$$

3. Recall, it measures when the model identifications are correct.

$$\frac{TP}{TP+FN}$$

4. F-measure, a measure of models reliability on plants identification. It is derived from the precision and recall evaluation methods.

$$\frac{2*Recall*Precision}{Recall+Precision}$$

5..Relative error is a measure of relative uncertainty or it compares a measurement to a precise value. It provides us with the exact number as well as the units of the quantity that differ from the genuine one.

$$Relative_error = \frac{x-x_0}{x}$$

where :
x = present approximately
x_0 = past approximately

6. ROC is a graph that depicts a classification model's performance over all categorization thresholds.

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

Where :
TPR = True positive Rate
FPR = False Positive Rate
FP = False Positive
TN = True Negative
FN = False Negative

IV. Results and Discussion

Since they are widely used by many people in Africa, traditional plant used for medicine has gotten a lot of attention. Proper plant identification on the other hand helps a broad range of audiences, including consumers, forestry services, botanists, taxonomists, physicians, pharmaceutical laboratories, endangered species organizations, government, and the general public.

Six distinct identification models was applied, as shown in the table below and shows accuracy, relative error, F-measure, and recall, where the percentages of the accuracies was close in all four methods of evaluation.

When I used shape and texture on CNN which result in 82 % accuracy, with the data set divided 80 - 20 for training and testing respectively, with 7 convolutional layers but when i used shape, texture and color the accuracy improved to 90%Ẇhen texture and shape features were used for the Bayesian network received a low 48% but when shape, texture, and color were used as features can observe a sightly increase to 51%

Table 2 : Accuracy for shape and texture

	Accuracy	Recall	Fmeasure	precision
CNN	82%	81.2 %	81.12%	81.5%
support vector machine	78.5%	78.6 %	78.9 %	78.34%
Naıve Baye	63%	63.13%	63.22%	63.12 %
Logistic regression	50.2%	50.3 %	50.4 %	50.56%
Bayesian Network	48.2%	49.2 %	49.3 %	49%
Recurrent neural network	60.2%	60.6 %	60.05 %	60.3%

Table 3 : Accuracy for shape, texture and colour				
	Accuracy	Recall	Fmeasure	precision
CNN	90%	90.1 %	90.15%	91.5%
support vector machine	82.5%	83 %	83.12 %	82.34%
Naive bayes	70. 3%	70.2%	70.1 2%	70.101%
Logistic regression	69.11%	69.21%	69.14 %	69.14%
Bayesian Network	51%	51.11 %	51.14 %	51.14%
Recurrent neural network	75%	75.2 %	75 %	75.1%

Table 2 shows the accuracy for models when are only tested on shape and texture of the plants which CNN model have the highest accuracy compare to all the models and Bayesian network has the lowest the accuracy which the accuracy have been improved by addition of colour feature but still Bayesian network have the lowest accuracy caused by not having enough features for the model to perform well and CNN perform well in image classification because of the layers can be adjusted to improve the accuracy.

V. Conclusion and future work

The goal was to assist traditional healers in identifying their plants, preserving the knowledge they know for future generations and finding alternative or similar plants in the event that the plants they use now become extinct. This goal was not met due to a lack of data on which plants they use but i was able to use the models to classify and identify edible plants which is very helpful for the next project now that they have information on which model works best.

According to the models I was able to use, CNN is very efficient at classifying and recognizing plants, which may be used in the future project to identify the plants that traditional healers actually use in their practice and to locate alternative or similar plants. To improve the accuracy of Bayesian model can be done by adding more features.

Acknowledgment

I'd want to thank my supervisors, Dr. Ritesh Ajoodha and Dr. Shalini Dukhan for guiding and advising me during this research as well as my family, friends and myself.

[1] [10] [15] [13] [7] [8] [4] [11] [14] [2] [6] [5] [12] [3] [9]

References

[1] S Anubha Pearline, V Sathiesh Kumar, and S Harini. A study on plant recognition using conventional image processing and deep learning approaches. *Journal of Intelligent & Fuzzy Systems*, 36(3):1997–2004, 2019.

[2] Oluleye Hezekiah Babatunde, Leisa Armstrong, Jinsong Leng, and Dean Diepeveen. A survey of computer-based vision systems for automatic identification of plant species. *Journal of Agricultural Informatics*, 6(1):61–71, 2015.

[3] Adams Begue, Venitha Kowlessur, Fawsi Mahomoodally, Upasana Singh, and Sameerchand Pudaruth. Automatic recognition of medicinal plants using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 8(4):166–175, 2017.

[4] Ali Caglayan, Oguzhan Guclu, and Ahmet Burak Can. A plant recognition approach using shape and color features in leaf images. In *International Conference on Image Analysis and Processing*, pages 161–170. Springer, 2013.

[5] Subhankar Ghosh, H Kumar, and Jyothi S Nayak. Study on classification of plants images using combined classifier. *International Journal*, 3(4), 2015.

[6] Kang Liu, Azian Azamimi Abdullah, Ming Huang, Takaaki Nishioka, Md Altaf-Ul-Amin, and Shigehiko Kanaya. Novel approach to classify plants based on metabolite-content similarity. *BioMed research international*, 2017, 2017.

[7] P Manojkumar, CM Surya, and P Gopi Varun. Identification of ayurvedic medicinal plants by image processing of leaf samples. In *Third ınternational conference on research in computational ıntelligence and communication networks*, 2017.

[8] Jagadeesh D Pujari, Rajesh Yakkundimath, and Abdulmunaf S Byadgi. Image processing based detection of fungal diseases in plants. *Procedia Computer Science*, 46:1802–1808, 2015.

[9] Zhang Qi. Who traditional medicine strategy. 2014-2023. *Geneva: World Health Organization*, page 188, 2013.

[10] Angie K Reyes, Juan C Caicedo, and Jorge E Camargo. Fine-tuning deep convolutional networks for plant recognition. *CLEF (Working Notes)*, 1391:467–475, 2015.

[11] Vijay Satti, Anshul Satya, and Shanu Sharma. An automatic leaf recognition system for plant identification using machine vision technology. *International journal of engineering science and technology*, 5(4):874, 2013.

[12] Divya Srivastava, Rajesh Wadhvani, and Manasi Gyanchandani. A review: color feature extraction methods for content based image retrieval. *International Journal of Computational Engineering & Management*, 18(3):9–13, 2015.

[13] Nguyen Van Hieu and Ngo Le Huy Hien. Recognition of plant species using deep convolutional feature extraction.

[14] Nguyen Van Hieu and Ngo Le Huy Hien. Automatic plant image identification of vietnamese species using deep learning models. *arXiv preprint arXiv:2005.02832*, 2020.

[15] B Yanikoglu, Erchan Aptoula, and Caglar Tirkaz. Automatic plant identification from photographs. *Machine vision and applications*, 25(6):1369–1383, 2014.