Medicinal Plant Detection Using Ensemble Learning Techniques

Prof.Premanand Ghadekar, Ubed Shaikh, Sanket Patil, Rajan Ner, Omkar Nimase, Tejas Shinde.

Abstract— Proper identification of plant species delivers significant benefits to various stakeholders: forest service, botanist, taxonomist, physician, Pharmaceutical Research Institute, Combat Organization Endangered species, governments, and the general public. Medicinal plants help in the treatment of several diseases. Ayurvedic medicinal plants are used in the traditional medicine system of India. Incorrect identification of medicinal plants. It can lead to bad results. Plant identification Automated based on visual morphological features Leaf shape, color, texture, etc. Ensemble learning is a machine learning technique to classify various medicine plants. We used ensemble learning to create an accurate classifier capable of identification of medicinal plants. This paper used bagging and boosting ensemble learning techniques, which improves the accuracy of classification of medicinal plants.

Keywords -Medicinal plants, Ensemble learning, Dataset, Machine Learning, Classification.

I. Introduction

One of the key sources of medicine are medicinal plants. Medicinal plants help in the treatment of several diseases and also helps in curing them, they are very effective and they come with no side effects. There are many medicinal plants that are used to cure different diseases like respiratory disorders, gastrointestinal diseases, dermatological diseases, excretory disorders, cardiac disorders, joint issues, fertility issues etc. and the list is increasing day by day. Medicinal plants have been used for the treatment of various human diseases by various communities for centuries. Ayurvedic medicinal plants are plants which are used in ayurveda, the traditional medicine system of India. Ayurvedic texts describe the use of more than a thousand plant-based medicines for a wide variety of conditions. Many of these plants are common in other parts of the world as well, and some have been introduced to other parts of the world by ayurvedic practitioners. Since the best techniques to classify these plants is manual but considering the amount of time it is on the slower side. Hence the classification of the medicinal plants is a difficult task and requires special techniques to classify them as most of these plants are in regions where it is not easy for us to reach. So it is quite beneficial in that case if there are some recognition technologies which will classify these medicinal plants accordingly. In this project there is use of ensemble learning to classify various medicine plants.

There is use of ensemble learning technique to create an accurate classifier capable of identification. Ensemble learning is a machine learning technique that combines the predictions of multiple models to produce better results than any single model. Ensemble learning is a powerful technique because it can help to improve the accuracy of your predictions, and it can be used with a variety of different machine learning algorithms.

II. Literature Review

The authors of this research concentrate on several robotized classification and recognition frameworks for plants. There are thousands of different plant species in the globe, and while many of them offer therapeutic benefits, others are in danger of going extinct, and still others are dangerous to people. Plants must be properly investigated and categorized in order to be used and protected as species. As a result, several researchers have carried out studies on the automatic categorization of plants based on morphological characteristics. In this work [2], classification will be performed using enhanced machine learning-based models. The experimental results of the suggested model are explained in this section. TP + FP = precision TP One parameter for assessing classification models is accuracy. Therefore, a quick and accurate technique of detecting herbs is needed. Improved machine learning classifiers with some preprocessing and feature selection models will be employed in next research in the field of plant identification to address accuracy-related problems and increase performance. Images gathered from 10 plant species with therapeutic significance were used in the simulation. Additionally, the computation of the texture and color features uses the normalized GLCM. These photos have the form characteristics taken from them and shown. As a result, this chart suggests that KNN will have 100% accuracy. But there are additional components in addition to the diagonal ones. Numerous research for the identification of medicinal plants deal with fewer classifications, as has been noted. The program has two main objectives ie. identification of medicinal herbs and retrieval of medicinal herbs. For image classification, we used local binary models to extract leaf texture and a probabilistic neural network. Identification of medicinal plants is based only on the structure of the leaves. In this study, many shape-based elements from

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

the leaves of medicinal herbs were extracted using computer vision algorithms. The classification of leaves from two distinct plant species into the proper groups was then done using machine learning techniques. In addition to assisting taxonomists in the development of more effective species-identification methods, an online or mobile computer system for automatic identification of medicinal plants also makes a substantial contribution to the protection of endangered species.

In the proposed Model [11] use of multispectral and texture datasets, the proposed model develops a machine learning model for medical plant leaves classification. There are many plants that provide air and water to living humans, they all play an essential role in maintaining the earth's biodiversity. In this study, we will collect a refined and standardized dataset, identify edge/line features, fuse extracted features, optimize extracted features, select the most valuable feature, and select efficient ML classifiers. By using the chi-square feature selection approach, one can select the 14 most valuable features that will produce better classification results. Therefore, medical plants are plants that are used to treat human or animal health problems.

In This Algorithm [13] Ayurvedic medicines are prepared by identifying and classifying medicinal plants based on their shapes, colors and texture. Using these features, researchers classify plants based on their spatial and morphological characteristics. The AyurLeaf model outputs the accuracy-loss graph during training, as well as the confusion matrix during classification. In addition to its own dataset, AyurLeaf has been trained and tested against the DLeaf dataset as well. In addition to using the AyurLeaf dataset, we also train and test AyurLeaf.

III. Methodology/Experiments

A. DESCRIPTION OF THE DATASET

In this research, to experiment on the Ayurvedic medicinal plant classification, we have used the MepcoTropicLeaf-V1 dataset and Medicinal Leaf Dataset which we obtained from Kaggle . The dataset consists of 40 classes and 3000 images. The following tables gives a short description and range/levels of the x-attributes in the dataset:

A. Spinach Dataset

Class	No. of Images
Amaranthus Green	123
Amaranthus Red	89
Balloon vine	123
Betel Leaves	127
Black Night Shade	108
Celery	82
Chinese Spinach	60
Coriander Leaves	120
Curry Leaf	109
Dwarf Copperleaf (Green)	88
Dwarf Copperleaf (Red)	79
False Amarnath	101
Fenugreek Leaves	80
Giant Pigweed	103
Gongura	53
Indian Pennywort	64
Lagos Spinach	84
Lambs Quarters	69
Lettuce Tree	64
Malabar Spinach(Green)	106
Mint Leaves	125
Mustard	80
Palak	86
Siru Keerai	68
Water Spinach	55

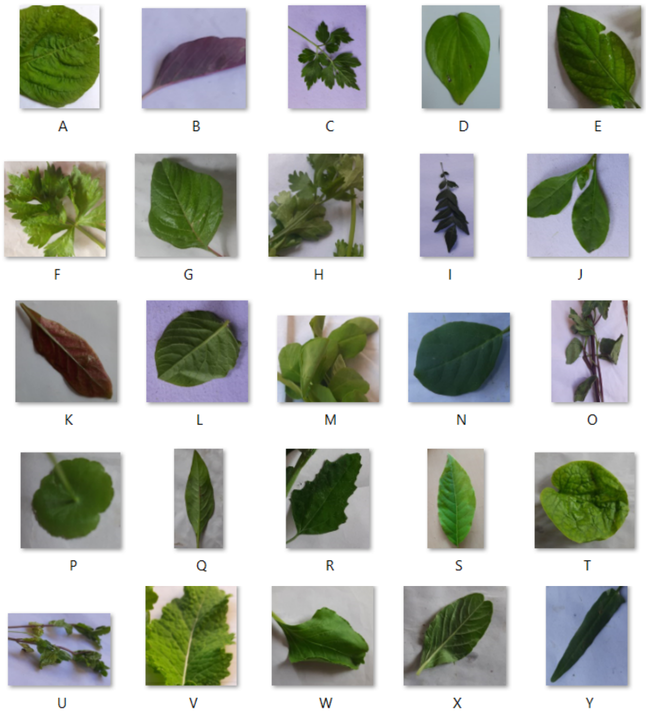

Fig 1. Sample Images of 25 Classes From Dataset

B. DATA PREPROCESSING

Pre-processing, when both the input and output are intensity images, is a phrase used to describe operations on images at their most fundamental level. An intensity picture is often represented by a matrix of image function values, and these images are of the same kind as the original data captured by the sensor (brightnesses). The purpose of pre-processing is an improvement of the image data that suppresses unintentional distortions or enhances some image features essential for further processing, even though geometric transformations of images (such as rotation, scaling, and translation) are categorized here as pre-processing methods because similar techniques are used.

Train and Test splitting:

For training and testing the model, the given dataset is split into training and testing data where training data is used to train the prediction model and then the model is tested on the testing data to check the accuracy of the prediction model.

In this experiment, we have split the dataset into training and testing data in the ratio 8:2 (i.e. 80% training data and 20% testing data) using in-built methods which randomly splits the data in the mentioned ratio but in a balanced manner to maintain the overall class contribution equal in both training and testing sets.

Fig 2. Sample Images After Preprocessing (resize,rescale,rotate)

C. ALGORITHMS USED

1.Convolutional Neural Network(CNN):

CNNs are made up of neurons that learn to optimize themselves and are similar to conventional ANNs. The foundation of innumerable ANNs, each neuron will continue to take in input and carry out an action (such as a scalar product followed by a nonlinear function). The complete network will express one perceptive scoring function from the input raw picture vectors to the final output of the class score (the weight). The final layer will include loss functions related to the classes, and all of the standard best practices created for conventional ANNs still hold true.

2. Decision Tree (DT):

Decision trees are tree-like structured prediction models that can be built for regression as well as classification. A decision tree is developed incrementally by breaking down the dataset into smaller chunks or subsets by applying classification techniques on the attributes. As a result, the DT is constructed having decision and leaf nodes. The decision node at the very top is called the root node and it indicates the best predictor variable.Both Categorical and numerical data work well with DTs. DTs models have a very high tendency of overfitting and this results in negative accuracy. Overfitting occurs when the learning algorithm gives very high accuracy on the training data but fails to give accurate results on testing data. This problem can be solved by pruning the tree and then setting some constraints on the model. Also, Decision trees cannot be used well with continuous numerical variables because a slight change in the input data can drastically change the structure of the tree and this causes instability.

3. Random Forest (RF):

It is a supervised learning technique that works on a Tree-based classification. RF uses ensemble learning methods for regression. RF works very well with regression as well as classification problems. Just like its name suggests, the algorithm is to select random samples from the training dataset and n-number of DTs are constructed based on a subset of selected variables from the dataset.

Algorithm:

1) It randomly selects 'm' features from the dataset.
2) For every node it finds the best fit.
3) Split the node using the best split.
4) It keeps repeating the above 3 steps and builds a forest of DT's.

The feature of this algorithm is that it provides very accurate results when dealing with sample amounts of data. Missing or null data values are very efficiently handled in RF. The main disadvantage of RF is that it is very complex, slow and ineffective because for real time prediction, it requires more DT's and hence, has high processing costs.

4. Support Vector Machine (SVM):

SVMs are ML tools used mostly for regression and classification analysis that examine data and identify patterns or decision boundaries within the dataset. SVM creates hyper-planes in a multidimensional space that divide various class borders; the dataset's feature vector refers to the number of dimensions. As seen in Fig. 2 [5], SVM can handle several continuous and categorical variables. There are two types of circles: full circles and outlined circles. The SVM's objective is to classify the two kinds according to their characteristics. Three lines make up the model. One is the marginal line or margin, which is,

$$w.x-b=0.$$

The closest data points for both classes are shown by the lines,

$$w.x-b =1 \text{ and } w.x-b =-1$$

The support vectors are the circles located on the hyper-plane. An outlier is the filled circle in the other class. To prevent overfitting and achieve a virtually flawless categorization, it is disregarded. In order to reduce the likelihood of generalization error, the SVM aims to maximize the perpendicular distance between the two edges of the hyper-plane. The generalization capacity rises as the number of support vectors decreases since the hyper-plane depends on them.

5. K-Nearest Neighbor Algorithm(kNN):

k-Nearest Neighbor (kNN) algorithm is a machine learning algorithm which is very productive and effortless. kNN makes consecutive clusters by grouping data and the newly entered data is classified based on it because of similarity with previously trained data. The input is assigned to the class with which it shares the most nearest neighbors. k-nearest-neighbor classification was created to carry out characteristic analysis when it was uncertain or challenging to obtain unambiguous parametric approximations of probability densities. It is known for its effectiveness and simplicity. In a 1951 report from the US Air Force School of Aviation, Fix and Hodges introduced the k- nearest neighbor rule, a non-parametric algorithm for pattern classification.

It is a classification algorithm. Mainly there are two steps in classification:

1. Learning Step: A classifier is constructed using the training dataset.

2. Assessment of the classifier.

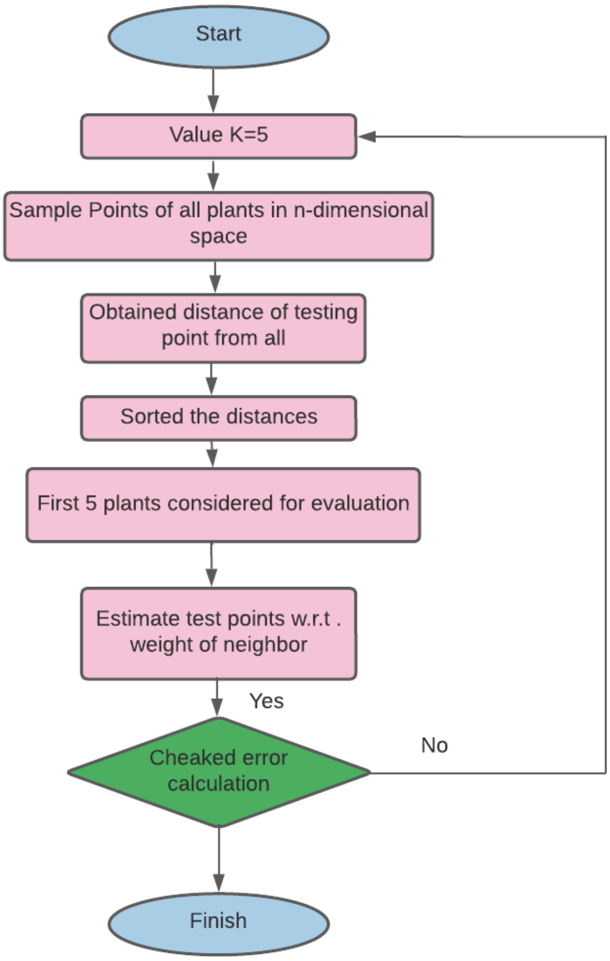
Start
Value K=5
Sample Points of all plants in n-dimensional space
Obtained distance of testing point from all
Sorted the distances
First 5 plants considered for evaluation
Estimate test points w.r.t . weight of neighbor
Yes
No
Cheaked error calculation
Finish

Fig 3. KNN Model

D. ENSEMBLE LEARNING MODEL

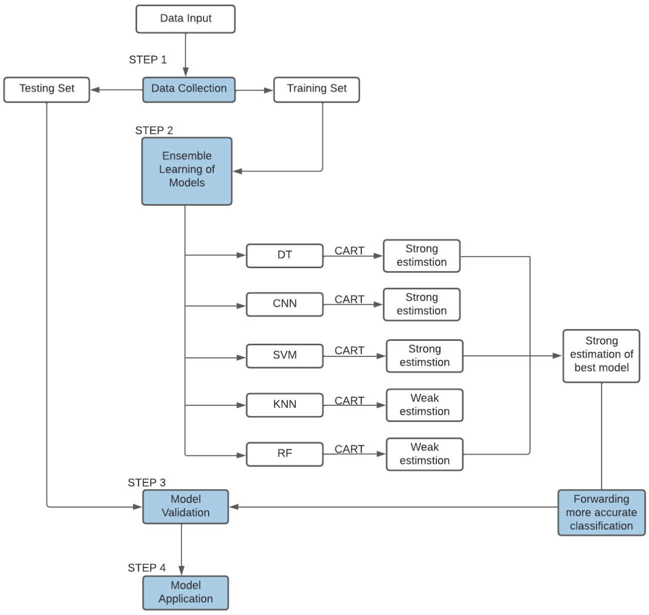

Fig 4. Project flow diagram

IV. Experiment and results

1. Convolutional Neural Network(CNN):

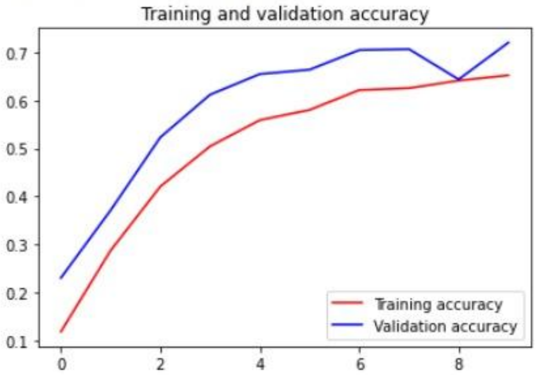

Fig 5. Training and validation accuracy

The whole dataset mentioned in the above table splitted into train, test and valid in the ratio 70:20:10 and the above graph (fig 3) shows epoch vs accuracy(training and validation).The accuracy we achieved is 89%.

2. Decision Tree (DT):

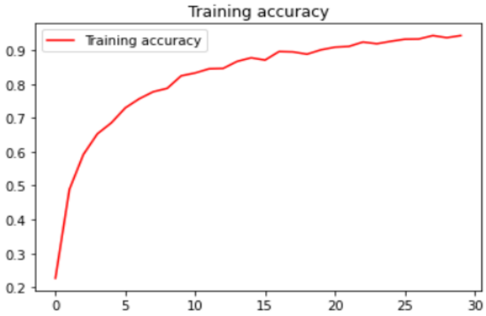

Fig 6. Training and validation accuracy

Dataset mentioned in the above table splitted up into train, test in the ratio 80:20 and the above graph (fig 4) shows epoch vs accuracy(training and validation).The accuracy we achieved is 98%.

3. Random Forest (RF):

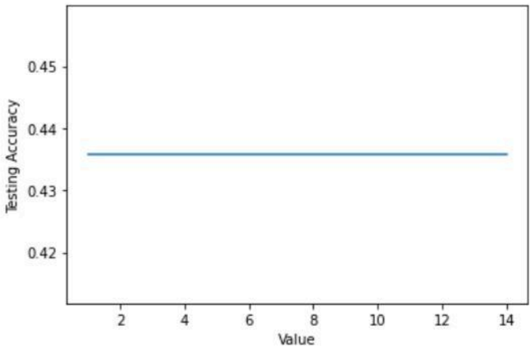

Fig 7. Training and validation accuracy

Dataset mentioned in 11the above table splitted up into train, test in the ratio 80:20 and the above graph (fig 4) shows epoch vs accuracy(training and validation).The accuracy we achieved is 98%.

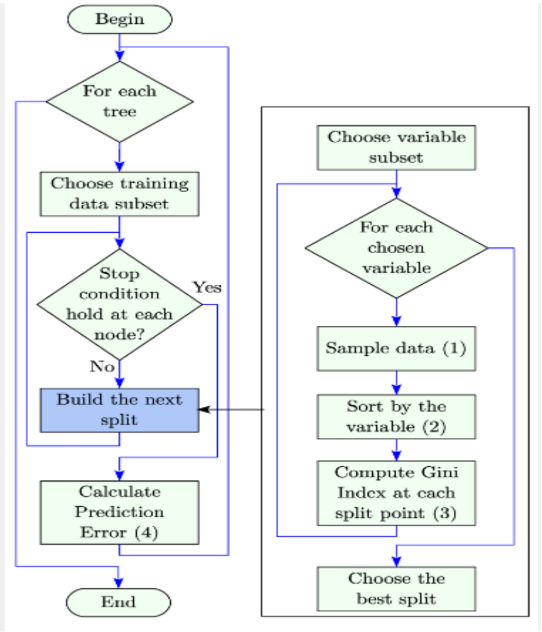

Fig 8. RF Algorithm

4. Support Vector Machine (SVM):

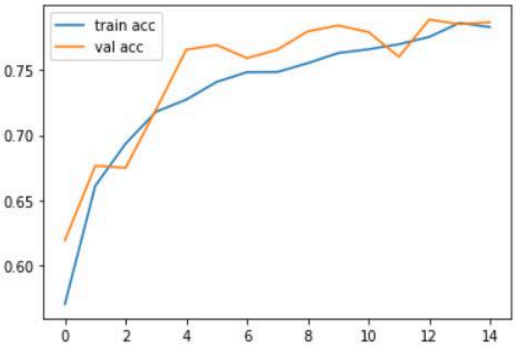

Fig 9. Training and validation accuracy

Dataset mentioned in the above table splitted up into train, test in the ratio 80:20 and the above graph (fig 4) shows training vs testing accuracy(training and validation).The accuracy we achieved is 80%.

5. K-Nearest Neighbor Algorithm(kNN):

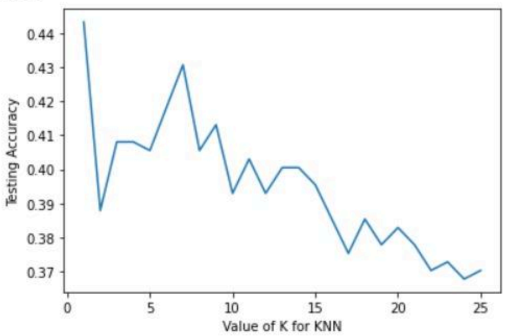

Fig 10. Testing accuracy for different values of k

In the analysis of the kNN algorithm the data split into the training and testing subsets with k-fold cross validation methods and the ratio is 80:20. In the above graph (fig 4) the accuracy is gradually decreasing along the incrementing value of "K". The accuracy we achieved is 45%.

MODELS	EXISTING ACCURACY (SAMPLES)	CURRENT ACCURACY (3000 SAMPLES)
DT	**85** (850)	**98%**
CNN	**95%** (1200)	**89 %**
SVM	**96%** (1200)	**80%**
KNN	**42%** (350)	**45%**
RF	**40%**(400)	**43%**

Fig 11. Evaluation Metrics for Algorithms Used

The overall accuracy of ensemble learning of algorithms is 89%.

V. CONCLUSION

This research paper has reviewed, analyzed and examined most of the current research on noteworthy attributes and algorithms of the medicinal plant classification system, used to predict the class of a medicinal plant. This research paper has used ensemble learning techniques to increase the accuracy of Medicinal plant classification. The model proposed by us will definitely help the scientist or botanist to classify between different types of Medicinal Plant and these plants can be used for various medicinal purposes. The paper also shows the everlasting potential of the SVM, Random Forest and CNN, KNN, DT and Naive Bayes models while classifying the medicinal plants. These models are based upon some of the input attributes and also work very well while classifying any type of medicinal plant. The impact of this study was intended to assist the other researchers in creating a realistic model that could easily and accurately classify medicinal plants.

VI. REFERENCES

[1] Banita Pukhrambam , Dr. R.Rathna , "A Smart Study on Medicinal Plants Identification and Classification using Image Processing Techniques ", Vels Institute of Science, Technology & Advanced Studies Chennai, India, [2021].

[2] Maibam Maikel Singh , Thounaojam Rupachandra Singh, "A Survey on Different Methods for Medicinal Plants Identification and Classification System", Department of Computer Science, Manipur University, India [2021].

[3] Amrutha M Raghukumar , Gayathri Narayanan , "Comparison Of Machine Learning Algorithms For Detection Of Medicinal Plants ", Amrita Vishwa Vidyapeetham, Amritapuri, India, [2020].

[4] K. Priya , S. Ranjana , R. Manimegala , " Medicinal Plant Identification Using Android Application Based On Leaf Image ", European Journal of Molecular & Clinical Medicine ISSN 2515-8260. [2020].

[5] Adams Begue, Venitha Kowlessur , Fawzi Mahomoodally , Upasana Singh , "Automatic Recognition of Medicinal Plants using Machine Learning Techniques ", Faculty of Science, University of Mauritius .

[6] Upasana Gitanjali Singh , Fawzi M. Mahomoodally , "Automatic Recognition of Medicinal Plants using Machine Learning Techniques", University of Mauritius,

[7] Vishakha A. Metrea , Sudhir D. "Research Review on Plant Leaf Disease Detection utilizing Swarm Intelligence" 2021

[8] Dhanuka Nadeeshan , Geeshani Amarawansha , Dasuni Nawinna , "Identification of Medicinal

Plants by Visual Characteristics of Leaves and Flowers” ,[2019]

[9] Nayana G. Gavhale, Dr.A.P.Thakare, “Identification of Medicinal Plant Using Machine Learning Approach”,[2020]

[10] Owais A. Malik, Nazrul Ismail , Burhan R. Hussein and Umar Yahya, “Automated Real-Time Identification of Medicinal Plants Species in Natural Environment Using Deep Learning Models—A Case Study from Borneo Region”,[2022]

[11]Samreen Naeem, Aqib Ali, Christophe Chesneau, Muhammad H. Tahir, Farrukh Jamal, Rehan Ahmad Khan Sherwani and Mahmood Ul Hassan, “The Classification of Medicinal Plant Leaves Based on Multispectral and Texture Feature Using Machine Learning Approach”,[2021]

[12] Md. Khairul Islam, Sultana Umme Habiba, Sk. Md. Masudul Ahsan, “Bangladeshi Plant Leaf Classification and Recognition Using YOLO Neural Network”,[2019]

[13] Dileep M.R. , Pournami P N. “ A Deep Learning Approach For Classification Of Medicinal Plants”, IEEE [2019].

[14] Nghia Duong-Trung, Luyl-Da Quach, Minh-Hoang Nguyen, Chi-Ngon Nguyen, “A Combination of Transfer Learning and Deep Learning for Medicinal Plant Classification”,ICIIT '19: Proceedings of the 2019 4th International Conference on Intelligent Information Technology,[2019]

[15] A.D.A.D.S. Jayalath, T.G.A.G.D Amarawanshaline, D. P. Nawinna, P.V.D. Nadeeshan, H.P Jayasuriya, “Identification of Medicinal Plants by Visual Characteristics of Leaves and Flowers”,[2019]

[16] Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma, “A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification”, International Conference on Intelligent Sustainable Systems (ICISS),[2019]