iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** V **Month of publication:** May 2018

DOI: http://doi.org/10.22214/ijraset.2018.5443

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

Review of Leaf Classification Techniques using Machine Learning

Shubham Kumar Singh[1], Bittu Dhiman[2], Dev Pratap Singh[3]
[1, 2, 3]*Department of Computer Application, National Institute of Technology, Kurukshetra*

Abstract*: Plants form an indispensable part of our lives and as such the need for their identification and conservation is very important. A leaf is an organ of vascular plant and is the key parallel member of the stem. Each leaf has an arrangement of highlights that separate it from alternate leaves, for example, edge and shape. This paper proposes a comparative study of different classification approaches of supervised learning, based on different orientation and description of leaves and chosen algorithms. Starting with orientation of leaves, we exhibited leaves by a fine-scale edge feature histogram, by a Centroid Contour Distance Curve shape signature. At that point we assess the characterization utilizing cross validation. The acquired outcomes are extremely intriguing and demonstrate the significance of each component. The plant leaf image is classified with biometric features. Customarily, the prepared ordered play out this procedure by following different undertakings. Leaf biometric include are examined utilizing computer based strategy like morphological element examination and artificial neural system based classifier. KNN show take contribution as the leaf venation morphological element and group them into four distinct species.Keywords: Leaf Identification, Artificial neural network, Canny edge detection, k-NN Classification, Leaf venation pattern, Morphological Features

I. INTRODUCTION

Plants form an important part of our lives. We are a lot more dependent on plants than we think we are. Our mere existence is dependent on them. So it is even more important that we know about the plants around us and their properties. They provide a lot of stuff like oxygen, fruits, grains, cereals and many more. The list goes on and on. But it is extremely unfortunate that we know so little about things that are so crucial to our survival. We don't know about most plants around us.

We don't know about their properties, uses and benefits. We don't know whether they are vulnerable or not. If we have all the information regarding plants, we not only would benefit immensely from this information but we would also take better care of both ourselves and the plants around us. Many plants are on the verge of extinction and we don't even know which one are they. Many plants are poisonous, many seem like poisonous.

Many have important medicinal properties. Obviously we can't keep a physical copy of all that data at our disposal every time. There are too many different fields of data involved here. So to achieve this target, all we need is a centralized source of knowledge and the easiest way to achieve this is to have a computer or smartphone to identify and classify a plant. Plus, we also need a database for storing and retrieving plant features, properties and its uses. To rectify this situation all we need is an application that is capable of not only identifying a plant but also providing important and necessary information to the user in a snap. Seems easy but once we take all the variables into account it seems improbable let alone feasible.

A. Our Problem and Machine Learning

Identification of plant using morphological features has been an interesting topic for researchers around the globe. The main reason being the challenges faced. The easiest way to identify a plant is by using the unique morphological features nature gifted them. But even if we can extract all the unique features correctly and accurately, there is no way any traditional problem solving algorithm will be able to solve it. The only way it seems possible is to make a computer learn how to recognize leaves.

And as you may have guessed, this can be possible by machine learning only. In our problem we have to classify the inputs into different plants, which again fall into different categories. So it is a supervised multi-variate classification problem. Multi-variate because the inputs will fall into many categories.

The impact of such an application can be measured by the fact that it will have multi-disciplinary applications ranging from agricultural use to military applications. And the need to create it is what motivates us to keep going. Since it is a machine learning problem we want to know which algorithm is best for us. And the only way to know this is by reviewing the algorithms on different datasets. A brief introduction of machine learning is provided below.

B. Machine Learning

The The only thing that sets Humans apart from animals is our ability to learn, reason and think, observe and put that into our actions. But the ability to put them into computer programs gives them the same abilities. Well almost half a decade from the inception of the term 'machine learning', we are still nowhere close to it. But interestingly, experimentations with the Human brain led us to many interesting discoveries. In one such experiment, the brain of a ferret was rewired in a way that the part which usually processes the audio signals received by the brain was connected to the eyes and made to process the visual signals [1]. Impressively, the ferret was able to process vision perfectly. This made scientist to think that the brain works on just one learning algorithm and computer programs should be designed to learn in the same way. That is by the experiences received during the learning period. As per Wikipedia, "Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to learn with data, without being explicitly programmed" [2]. As per Tom M. Mitchell Machine Learning "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [3]. This definition pretty much describes what actual machine learning is. When we think of machine learning another term comes into our minds, that is Artificial Intelligence. Well both are different in their own respect. While Artificial Intelligence deals with inductive logic programming and knowledge based reasoning, machine learning is more concerned with tackling solvable problems of practical nature by models borrowed from statistics, probability theory and linear algebra. Basically, machine learning can be classified into three broad categories:

1) *Supervised Learning*: In this category the computer is given example inputs with their desired outputs and the task is to generate a function that maps the inputs to outputs.
2) *Unsupervised Learning*: In this category the computer is not given any labels. It has to find structure in the input by itself.
3) *Reinforced Learning*: Training data is given as feedback or reward or punishment to the program's action in the dynamic environment. There are practically dozens of algorithms for each of the above mentioned category. Each of them having their own merits and demerits. The problems which can be solved by machine learning can also be categorized into three categories:
4) *Classification*: In classification problems the output is discrete. That is the model must divide the inputs into two or more categories. The basis of classification is the characteristics (also known as features) of the input examples. This typically comes under supervised learning.
5) *Regression:* In regression problems, the output is continuous. That is the model must predict the value of a certain feature by analysing the other features. The output can take any value in a range. This also typically comes under supervised learning.
6) *Clustering:* In clustering, the input is divided into multiple groups. But unlike classification, the group of the training input sets are not known beforehand, thus making it a typical unsupervised learning problem.

II. IMPORTANCE

Even to a normal person, plant form an integral part of their lives. But still he/she is not able to utilize the full potential of the plants around us. The main reason behind this is the ignorance and lack of knowledge about the plants. Even with the amount of research already made in Botany, the general public is still unaware of the most important and useful details of the plants around them. This doesn't only ends here, due the reasons mentioned above, the diseases in plants and their remedies are also not very well known to the public. This can be easily remedied by an application which classifies the plant and then fetches the details of the plant. But at the core of any such application lies the classifier that deals with the classification. And the faster the classifier is, the faster the application will be. The main focus of this paper is finding the best algorithm for the task. From a plethora of algorithms to choose from, it's very difficult to arrive at the right choice. But the algorithm is just one part of the image recognition process. The complete process comes in three stages which is explained below:

A. Information Gathering(Collection of leaf images)

The most important stages of all is the information gathering stage. This is the stage in which the information regarding plants is collected and used for the classification. The data collected in this phase consist of the image samples of a particular leaf and its corresponding label. The importance of this stage can be estimated from the fact that accuracy of classifier is directly dependent on the accuracy of the data gathered. The images collected should be sharp, shadow free and noiseless. An ideal image consist of scanned leaves on a white background so that data extraction is as easy as possible. Plus, the location information regarding the images will be helpful in making decisions easier. The accuracy of the classifier also depends of the number of the images of each plant. So the magnitude of this stage is immense. The number of plants in the world is almost unimaginable and so is the duration of

this phase. Normally this stage goes around in cycles where each cycle consist of updating the dataset with new plants and more samples of each plant.

B. Data Extraction(Image Pre-processing)

This stage consists of the extraction of features from the images gathered. The feature extraction consist of different data fields required for their classification. The type and number of fields varies from algorithm to algorithm and technique to technique. But some steps are common to all. The common steps are as follows:

1) *Image Enhancement*
2) *Noise Reduction*
3) *Image Segmentation*
4) *Feature Extraction*

The importance of this phase is also the same as the previous one. Accuracy in feature extraction plays a vital role in the classification this stage is all about feature extraction.

In the Image Enhancement phase consist of tweaking the colour scale and the saturation of the leaf. This is done to differentiate the leaf from the surroundings. It also helps in smoothening the image. This is done for better noise reduction and getting enhanced images. The objective of image enhancement is to transform RGB image to a suitable colour space which is device independent in nature**.** It also creates sharp borders which aids in the Noise Reduction phase.

The Noise Reduction phase consist of the reduction of noise from the images. The noise consist of shadows, blurring of a part of the image or any other kind of distortion. These obstruct the feature extraction process so it is extremely vital that the noise in each and every sample is properly reduced. This phase consist of applying different filters which consequently reduce the noise in the each images. A proper balanced setting of the filters is required for the proper noise reduction. Filters such as the Mean Filter, Median Filter, Gaussian Filter, Anisotropic Diffusion etc. are consistently and frequently employed in this stage.

extraction. The background of the image is also removed as it also obstructs the feature extraction process. The main objective of this stage is to speed up the feature extraction process so that it can efficiently extract all the features required for the algorithm.

The Feature Extraction process involves extraction of the various features such as diameter, physiological length etc. These features form the dataset of the leaf. The classifier is trained on this dataset. So the precision required in this stage is immense. These features define how the computer sees the image.

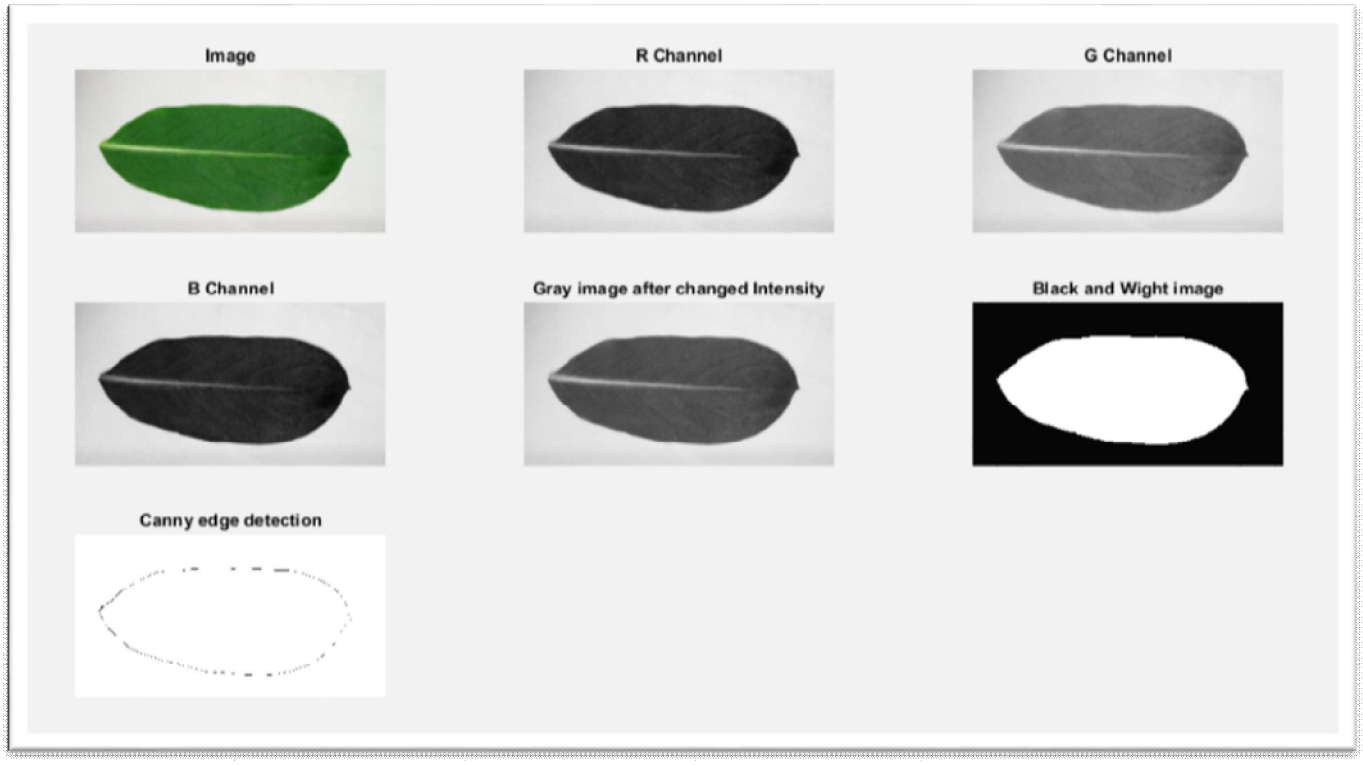


Fig. 1 Image Pre-processing

C. Classification(Training and Prediction)

This phase consist of the classification of the leaves using a machine learning classifier. The speed, efficiency and practicality of the classifier depends on the algorithm used and the kind of data we have. From a sea of algorithms, it is very difficult to choose one and hence extra care should be taken in the choice of the algorithm. The tools that can help in the choice of the algorithm is very limited and most researchers are mostly dependent on their intuition for this. But a comparative study of different algorithms definitely helps in the long run. Then there is the problem of evaluation metrics too. The evaluation metrics grades the result of the classifier and it is very important that the evaluation metrics should be chosen with no prejudice or bias. The evaluation metrics differ from problem to problem. For example, for a skewed binary classification problem, the general evaluation metric is area under the receiver operating characteristic curve (ROC AUC). For a multi-variate classification problem, the general evaluation metric is categorical cross-entropy or multiclass log loss. Plus optimization techniques such as feature scaling, mean normalization and learning rate optimization play a vital role in the training phase.

III.LITERATURE SURVEY

A few frameworks utilize portrayals utilized by botanists [8] – [11]. Be that as it may, it is difficult to concentrate and exchange those features to a computer consequently. This paper tries to counteract human impedance in include feature extraction. It is likewise a since quite a while ago talked about theme on the most proficient method to concentrate or measure leaf features [12] – [18]. That makes an application to recognize the pattern in this field is a new challenge [4] [19]. As per [4], information securing from living plant naturally by the computer has not been implemented.

A few different methodologies utilized their pre-characterized features. Miao et al. proposed a proof hypothesis based rose grouping [6] in light of numerous feature of roses.

Gu et al. attempted leaf recognition utilizing skeleton division by wavelet change also, Gaussian introduction [20]. Wang et al. utilized a moving middle focus (MMC) hypersphere classifier [21]. Comparable technique was proposed by Du et al. [4]. Their paper proposed an adjusted dynamic programming algorithm for matching the leaf shape [22]. Ye et al. thought about the closeness between features to characterize plants [5].

Abdul Kadir et.al. [2012]This paper reports the aftereffects of analyses in enhancing execution of leaf recognizable proof framework utilizing Principal Component Analysis (PCA). The framework included mix of features got from shape, vein, shading, and surface of leaf. PCA was consolidated to the distinguishing proof framework to change over the features into orthogonal features and afterward the outcomes were inputted to the classifier that utilized Probabilistic Neural Network (PNN).

Foliage and Flavia are the two datasets on which the PNN is tried, that contain different shading leaves (foliage plants) and green clears out separately. The outcomes demonstrated that PCA can expand the precision of the leaf distinguishing proof framework on both datasets. [7]

Samuel E. Buttrey et.al. [2002].

We build a hybrid (composite) classifier by consolidating two classifiers in basic utilize—classification trees and k-closest neighbour (kNN). In our plan we isolate the feature space up by a classification tree, and then characterize test set things utilizing the k-NN run just among those preparation things in an indistinguishable leaf from the test thing. This lessens to some degree the computational load associated with k-NN, and it produces a classification run the show that performs superior to either trees or the typical k-NN in a number of understood informational indexes. [20]

Among all methodologies, ANN has the speediest speed and best precision for classification work. [23] Demonstrates that ANN classifiers (MLPN, BPNN, RBFNN and RBPNN) run speedier than k-NN (k=1, 4) and MMC hypersphere classifier while ANN classifiers progress different classifiers on precision.

A. Artificial Neural Network

Artificial Neural Network are computing system consist of the large numbers of highly, elementary interlinked processing nodes that abstractly similar structure of the biological nervous system. Each nodes is capable of sending or receiving from one to another. The signal at the nodes is a real value, and the output of each node is calculated by the non-linear function of the sum of its input. Learning in ANNs is done with special training algorithms developed on the base of learning rules presumed. Training algorithms are an integral part of ANN model development.

A good training algorithm will shorten the training time, while achieving a better accuracy. The algorithm is not guaranteed to find the global minimum of the error function since gradient descent may get stuck in local minima, where it may remain indefinitely.

B. Logistic Regression:

Logistic Regression is a statistical method for analysing a dataset in which there is presence of two or more independent variable that determine an output. The output is in the form of dichotomous variable. Logistic regression models the probability of the default class. The main objective of the logistic regression is to find the best fitting model describing the relationship between the outcome variable and a set of independent variables. It predict the coefficients of the formula:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_K$$

This formula is used predict a logit (logistic regression) transformation of the probability of presence of the outcome variables, where p is the probability of the presence of the characteristic of outcome variables.

C. k-NN

k-NN (K nearest neighbour) is an algorithm which can further used in both classification as well as in regression predictive problem of machine learning. KNN has no model other than storing the entire dataset, so there is no learning required. Efficient implementations can store the data using complex data structures like "k-d trees" to make look-up and matching of new patterns during prediction efficient. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this might be the mode (or most common) class value.

To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance

TABLE I COMPARISON OF ALGORITHMS

Algorithm	Advantage	Disadvantages	Accuracy
Logistic Regression	1. It is robust. 2. It may handle nonlinear effects. 3. It does not require that the independents be interval.	1. If the set of independent variables include the wrong independent variables, the model will have little to no predictive value. 2.Independent Observations Required	89% [23]
k-nearest neighbours algorithm	1. It is a very simple classifier that works well on basic recognition problems. 2. Effective if the training-set is large	1. It is a lazy learner. 2. It does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data.	90%[23]
Artificial neural network	1.Relatively easy to use 2.Can approximate any function, regardless of its linearity 3.Great for complex or abstract problems like image recognition	1.Often abused in cases where simpler solutions like linear regression would be best 2.Requires a shit load of training and cases 3.Increasing accuracy by a few percent can bump up the scale by several magnitudes	93%[23]

IV. CONCLUSIONS

The main motive behind this is to give an approximate idea about which algorithm to use. The idea is to compare the algorithms so that we can efficiently and effectively use them. Plant identification is an important topic and hence we need to be extremely careful in the choice of the algorithm. The reason behind the identification of plants is to improve the usability of plants and maximize their

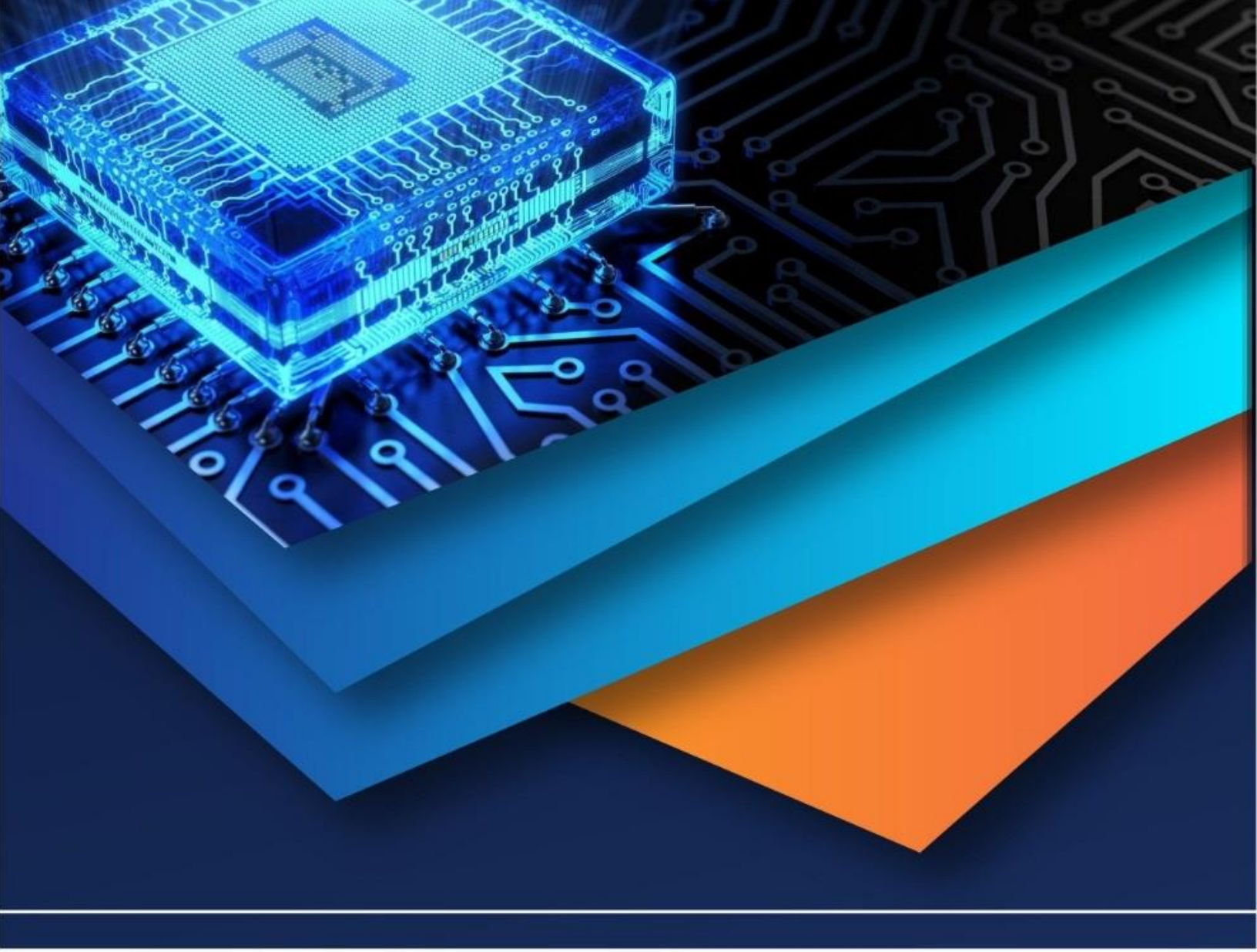
usefulness. It will also help in the protection and preservation of plants. We owe a lot to our mother earth and it is our moral responsibility to save the planet and its beautiful ecosystem.

V. ACKNOWLEDGMENT

We are extremely grateful to our mentors Dr Kapil Gupta and Mrs. Deepika Saxena for their valuable guidance and cooperation. This project cannot be possible without their efforts and thank them form within our hearts. We also thank Miss Neha Goyal for the valuable time she has found for us. We also thank the cooperative staff of the Department of Computer Applications for their ever ready support and help. The success of this project is dependent on them also. Finally we thank the Lord of the Lords Shiva for making this happen.

REFERENCES

[1] S. L. P. &. M. S. Laurie von Melchner, "Visual behaviour mediated by retinal," NATURE, vol. 404, no. 20 APRIL, pp. 871-876, 2000.
[2] Wikipedia, "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning. [Accessed April 2018].
[3] T. M. Mitchell, in Machine Learning, McGraw-Hill, 1997, p. 02.
[4] J.-X. Du, X.-F. Wang, and G.-J. Zhang, "Leaf shape based plant species recognition," Applied Mathematics and Computation, vol. 185, 2007.
Y. Ye, C. Chen, C.-T. Li, H. Fu, and Z. Chi, "A computerized plant species recognition system," in Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, October 2004.
[5] Miao, M.-H. Gandelin, and B. Yuan, "An oopr-based rose variety recognition system," Engineering Applications of Artificial Intelligence, vol. 19, 2006.
[6] Abdul Kadir et.al. ―Performance Improvement of Leaf Identification System Using Principal Component Analysis‖ International Journal of Advanced Science and Technology Vol. 44, July, 2012
[7] M. J. Dallwitz, "A general system for coding taxonomic descriptions," Taxon, vol. 29, 1980.
[8] H. Fu, Z. Chi, D. Feng, and J. Song, "Machine learning techniques for ontology-based leaf classification," in IEEE 2004 8th International Conference on Control, Automation, Robotics and Vision, Kunming, hina, 2004.
[9] D. Warren, "Automated leaf shape description for variety testing in chrysanthemums," in Proceedings of IEE 6th International Conference Image Processing and Its Applications, 1997.
[10] T. Brendel, J. Schwanke, P. Jensch, and R. Megnet, "Knowledge based object recognition for different morphological classes of plants," Proceedings of SPIE, vol. 2345, 1995.
[11] Y. Li, Q. Zhu, Y. Cao, and C. Wang, "A leaf vein extraction method based on snakes technique," in Proceedings of IEEE International Conference on Neural Networks and Brain, 2005.
[12] H. Fu and Z. Chi, "Combined thresholding and neural network approach for vein pattern extraction from leaf images," IEE Proceedings-Vision, Image and Signal Processing, vol. 153, no. 6, December 2006.
[13] Y. Nam, E. Hwang, and K. Byeon, "Elis: An efficient leaf image retrieva system," in Proceedings of International Conference on Advances in Pattern Recognition 2005, ser. LNCS 3687. Springer, 2005.
[14] H. Fu and Z. Chi, "A two-stage approach for leaf vein extraction," in Proceedings of IEEE International Conference on Neural Networks and Signal Processing, Nanjing, China, 2003.
[15] Z. Wang, Z. Chi, and D. Feng, "Shape based leaf image retrieval," IEE Proceedings-Vision, Image and Signal Processing, vol. 150, no. 1, February 2003.
[16] H. QI and J.-G. YANG, "Sawtooth feature extraction of leaf edge based on support vector machine," in Proceedings of the Second International Conference on Machine Learning and Cybernetics, November 2003.
[17] S. M. Hong, B. Simpson, and G. V. G. Baranoski, "Interactive venation based leaf shape modelling," Computer Animation and Virtual Worlds, vol. 16, 2005.
[18] F. Gouveia, V. Filipe, M. Reis, C. Couto, and J. Bulas-Cruz, "Biometry: the characterization of chestnut-tree leaves using computer vision," in Proceedings of IEEE International Symposium on Industrial Electronics, Guimar˜aes, Portugal, 1997.
[19] Samuel E. Buttrey et.al. ―Using k-nearest-neighbor classification in the leaves of a tree‖ Computational Statistics & Data Analysis 40 (2002) 27 – 37 www.elsevier.com/locate/csda.
[20] X.-F. Wang, J.-X. Du, and G.-J. Zhang, "Recognition of leaf images based on shape features using a hypersphere classifier," in Proceedings of International Conference on Intelligent Computing 2005, ser. LNCS 3644. Springer, 2005.
[21] J.-X. Du, D.-S. Huang, X.-F. Wang, and X. Gu, "Computer-aided plant species identification (capsi) based on leaf shape matching technique," Transactions of the Institute of Measurement and Control, vol. 28, 2006.
[22] J. Du, D. Huang, X. Wang, and X. Gu, "Shape recognition based on radial basis probabilistic neural network and application to plant species identification," in Proceedings of 2005 International Symposium of Neural Networks, ser. LNCS 3497. Springer, 2005.
[23] Abdulhamit Subasi, Ergun Erçelebi, " Classification of EEG signals using neural network and logistic regression,
Computer Methods and Programs in Biomedicine," Volume 78, Issue 2, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY