A statistical approach to identify diseased leaf from healthy

E. Priya
Department of ECE
Sri Sairam Engineering College
Chennai, India
priya.ece@sairam.edu.in

N. Dhanavarsha
Department of ECE
Sri Sairam Engineering College
Chennai, India
sec20ec076@sairamtap.edu.in

S.V.Gayathri
Department of ECE
Sri Sairam Engineering College
Chennai, India
sec20ec146@sairamtap.edu.in

N.Pavithra
Department of ECE
Sri Sairam Engineering College
Chennai, India
sec20ec194@sairamtap.edu.in

2022 International Conference on Communication, Computing and Internet of Things (IC3IoT) | 978-1-6654-7995-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC3IOT53935.2022.9767995

Abstract— **The economic advancement of a country depends on the green revolution. Productivity enhancement plays a significant part in improvising green revolution for country like India. Plant nutrition should be monitored to improve the yield in agriculture. This could be done either manually or by automatic procedure. Manual procedure drags the process. So the best way is to identify the mal nutrient of a plant by categorizing the diseased plant from the healthy plant. In this work, the leaf of Pongamia Pinnata species is taken from Mendeley Data. These leaves possess RGB color. They are converted into gray space along with contrast enhancement for further processing. Texture and tone features such as gray-level co-occurrence matric, statistical, histogram and probability measures are extracted from the contrast enhanced images. Statistical-based t test is conducted to find the significant features in categorizing the leaves into diseased and healthy. Among the 29 features, results demonstrate that energy is the most significant $(p = 0.0002)$ feature followed by maximum probability $(p = 0.0047)$ and information measure of correlation $(p = 0.0073)$. A good correlation $(r = 0.566)$ is observed for the feature energy with the out class, namely healthy and diseased. This work thus involves automation in the process of identifying the diseased leaf from healthy.**

Keywords— Plant pathology, Leaf images, Spatial gray tone, Image texture, Paired samples t test, statistically significant

I. INTRODUCTION

The economy of any developing country like India depends on Agriculture as majority of the population are engaged in agricultural produce. The majority of the rural population is dependent on agriculture for their income and per acre productivity determines the profit or loss to the farmers. Productivity depends on the fertileness of the soil and disease free crops. Diseases if detected early can be controlled otherwise will hamper the productivity of crops, resulting in low agricultural produce. In the upcoming years, there is a competent growth in crop yield, source-based competences. There is a vast growth expected in global crop through optimal use of land, water and nutrients in an exponential manner [1]. Report says, that there is a loss of 50% with the intervention of pests and disease that exist in a more general way [2].

Identification of crop disease at a beginning stage will increase the yield. Many approaches have been formulated to increase the production of food crops and others. Plant clinics take the role of identification and remedial measures to prevent the loss of the crops due to mal nutrients and diseases. The well-known manual procedure is the application of pesticides or pest management. This step is a crucial point in pest management [2]. Another drawback is that the farmers at remote places need to travel a long way to reach the expert to get an opinion. The farmers have to spend money and their time for the expert opinion [3].

The computer-based automation has evolved as a niche in identifying the disease infected plant from the healthy. The computer-based algorithms take lesser in time in generating a conclusive report. This also increase the accuracy of identification and throughput. But manual procedure needs expert in the proposed field, which could not be expected all the time. Pattern recognition and computer vision has made a revolution in automation. This has paved a way to compete among the algorithms thus developed [3, 4].

But this again results in a threat, is there is a lack of necessary setup to carry out the process. Food quality becomes a major menace when crops are infected with diseases. The infection, stages of crop disease and severity identification need to be done at a rapid rate. This still remains strenuous in most region of the globe due to the deficit of required technology and development [2].

Literature reveals several computer vision methods are in use for the identification of healthy and infected leaves [5, 6]. Various methods include pre-processing, segmentation, transformation-based methods and machine learning procedures. Support vector machine, neural network, k-means clustering, k-nearest neighbor, decision tree, adaboost, random forest, probabilistic method such as naïve Bayes, regression methods such as linear, logistic are the machine learning methods that are adopted. Recently Deep Neural Networks (DNN) are attempted in categorizing the leaves into healthy and unhealthy. Many variants of DNN are used by the authors. Sequential learning also has started to be implemented for online processing of the datasets. Conventional neural network is generally termed as data hungry, as it requires huge amount of data. Apart from data, there are lot many tunable hyper-parameters. It also requires higher end computational hardware for its processing. These drawbacks are overcome in transfer-based learning [3]. There is a tremendous use of transform-based methods that finds

978-1-6654-7995-0/22/$31.00 ©2022 IEEE

application in the identification of the plant leaves. Transform-based methods include wavelet, slantlet, ridgelet, contourlet, curvelet, ripplet, surflet, beamlet, ranklet, seislet and others.

In this work, Pongamia Pinnata commonly known as Pongam Tree leaves including both healthy and unhealthy are taken from the Mendeley Data. These leaves are used in the preparation of ointment that reduces the pigmentation due to lesion [7]. Computer vision procedure is carried out, that involves pre-processing via contrast enhancement and extraction of texture and tone-based features. Statistically significant feature is chosen by paired t test whose correlation with the outcome (healthy or unhealthy) is computed.

II. METHODOLOGY

A. Dataset description

The dataset of leaf images is referred from Mendeley Data. The dataset has 4503 images in total. Out of these images, 2278 are healthy and 2225 images are diseased. These leaf images were acquired at Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir in a closed noise free environment. The leaves were collected in the year 2019 from the month of March to May. The images were acquired with a Nikon D5300 camera of 18 to 55 mm lens. The images are of RGB color space with 24-bit depth and 2-unit resolution. Leaves of 12 different species of both healthy and diseased are found in the dataset. This study involves species from Pongamia Pinnata [8].

B. Methods

The considered species has 322 healthy and 276 diseased leaves. The schematic diagram of the work carried out is presented in Figure 1. Since the contrast of the acquired images are to be improved, the images from RGB color space are converted to gray scale. Also to extract texture features, the contrast of the images is boosted by clipping the image intensity below and above the mean times standard deviation value of the individual image under consideration.

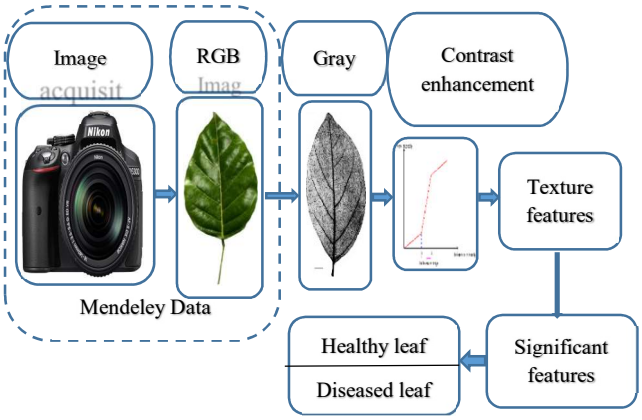


Figure 1.Schematic flow of the work

Extraction of relevant features help in distinguishing the leaves as healthy and diseased. Many authors have reported its importance in terms of geometric, texture and spectral features [9-13]. Meaningful features convey fruitful information. Spectral features convey the portion of electromagnetic spectrum thereby represent the tonal information. Texture and tonal features bear most of the important data of an image [14].

Totally 29 texture features from 4 groups namely: gray-level co-occurrence matric-based (G1), statistical-based (G2), histogram-based (G3) and probability-based (G4) are extracted from the contrast enhanced images. Of these features, most significant $(p \le 0.0001)$ features are alone considered for further segregating the healthy from diseased. The significant features include information measure of correlation (IC) from G1, energy (ER) from G2 and maximum probability (MP) from G4 [15]. The corresponding mathematical representation is as follows

$$IC = \frac{ET - EI}{\max\{Ei, Ej\}} \tag{1}$$

$$ER = \sum_{i,j} M(i,j)^2 \tag{2}$$

$$MP = \max\{M(i,j)\} \tag{3}$$

where $$EI = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M(i,j)\cdot \log\{M(i)M(j)\} \tag{4}$$

and $$ET = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M(i,j)\cdot \log(M(i,j)) \tag{5}$$

$M(i)$, $M(j)$ are the i^{th} and j^{th} entry in the marginal probability matrix, Ei and Ej are the entropies of $M(i)$ and $M(j)$ respectively.

$M(i,j)$ represent the probability density function of gray-level pairs [14, 16]. These significant features help in segregation of diseased leaves from healthy.

C. Statistical analysis

The significance among the features are computed by paired t test. This compares the average of two pair of measurements, i.e the features that are extracted from the pre-processed leaf images. It checks the statistical mean difference between the two set of features that deviate from zero [17].

III. RESULTS AND DISCUSSION

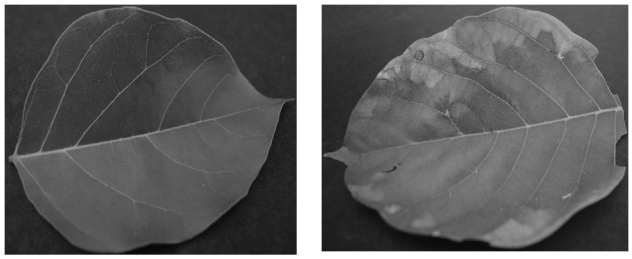
A representative of healthy and diseased leaf from the database is shown in Figure 2. The diseased leaf as seen from the figure 2 (b) has discoloration. This is due to lack of plant nutrients especially the Manganese deficiency. The leaf also has brown spots. They signify the leaf spot disease probably due to feedstuffs such as fungus or bacteria on the leaves. This also weakens the leaf and get torn. If the texture of such leaves are identified at an earlier time, necessary steps could have been taken to improve the nutrient of the plants.

(a) (b)

Figure 2. (a) Healthy and (b) diseased leaf

In order to identify the texture of the leaves the RGB color space has to be converted to gray color. Figure 3 (a) and (b)

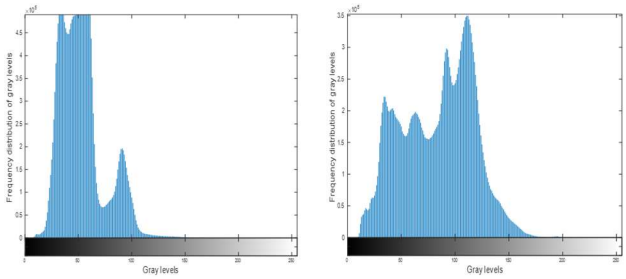
signifies the gray scale image of healthy and diseased leaf respectively.



(a) (b)

Figure 3. Gray scale image of (a) healthy and (b) diseased leaf

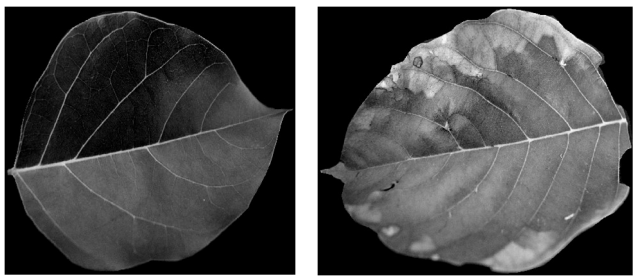
The histogram of the gray scale healthy and unhealthy leaf is shown in Figure 4 (a) and (b). Histogram of the healthy leaf shows two distinct peaks. One of the peak corresponds to the leaf shade and the other small peak corresponds to the background. Similarly, the histogram of unhealthy leaf shows significant peaks which corresponds to the different gray shades due to mal nutrient of the leaf. The histogram itself shows distinct difference between the healthy and the unhealthy leaf.



(a) (b)

Figure 4. Histogram of (a) healthy and (b) diseased leaf

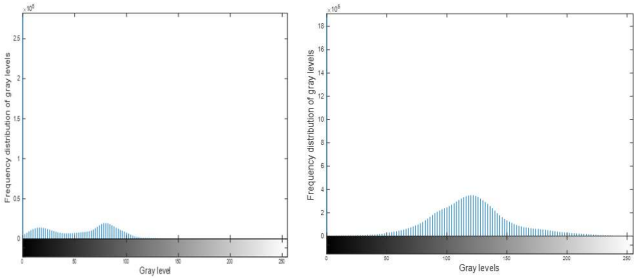
The discoloration and brown spots present in the unhealthy leaf are not clearly highlighted. So to highlight them the contrast of the image are improved. The corresponding images are shown in Figure 5 (a) and (b).



(a) (b)

Figure 5. Contrast enhanced image of (a) healthy and (b) diseased leaf

To further find the gray level distribution, histogram plots are done for the contrast enhanced leaves. The enhanced leaves presented in Figure 6 show significant gray level stretching. Also a sharp peak is noticed at the gray level zero. This represents the background of the image.



(a) (b)

Figure 6. Histogram of contrast enhanced (a) healthy and (b) diseased leaf

In order to extract the texture and tone, 29 features are extracted. The paired t test is conducted to pick out the significant features. The p values for the features whose value are less than 0.001 are to be considered as statistically significant.

Table 1. Mean ± standard deviation values of significant features

Category of leaf	Significant features		
	Energy	Maximum probability	Information measure of correlation
Healthy leaf	0.881 ± 0.071	0.397 ± 0.012	0.905 ± 0.009
Diseased leaf	0.538 ± 0.133	0. 971 ± 0.075	0.472 ± 0.021

The feature energy, maximum probability and information measure of correlation resulted in p values of 0.0002 , 0.0047 and 0.0073 respectively. The energy feature is extremely statistically significant, maximum probability is very statistically significant and information measure of correlation is considered to be very statistically significant. Table 1 shows the mean values of these features and their deviation from the mean for all the considered healthy and diseased leaves.

It is observed that the energy is high for healthy than the diseased. This is because the healthy leaf pixel intensities are more or less uniform and homogeneous. Table 1 shows the feature, maximum probability of diseased higher than healthy. The dominance in adjacent gray values is the reason for this. Since the gray level dependence between adjacent pixels are high, the information measure of correlation seems to be high.

The correlation measure among these features are studied, so as to identify the feature that is more relevant to categorize diseased from healthy. Figure 7 shows the correlation matrix of the significant features. The plot in the diagonal arm signifies the distribution of these features. It demonstrates that information measure of correlation has few outliers than energy and maximum probability.

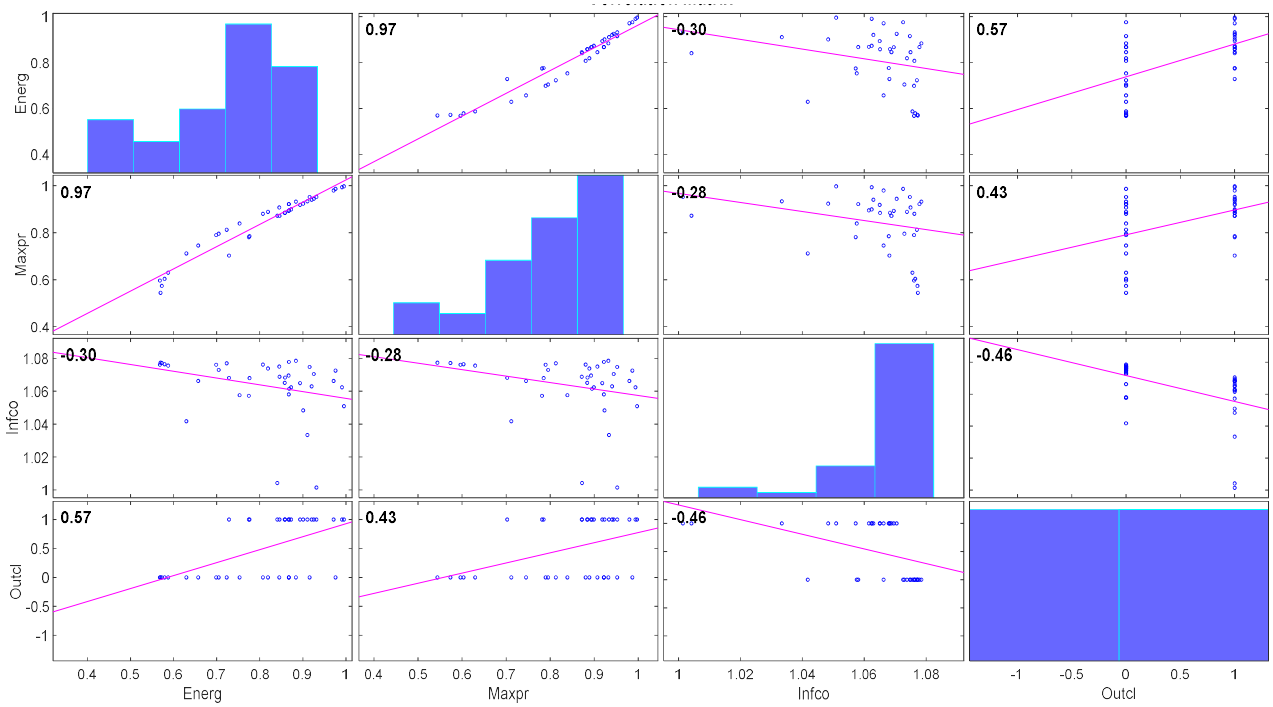


Figure 7. Correlation matrix of the statistically significant features

Table 2. Correlation values of the statistically significant features

Significant features & out class	Energy	Maximum probability	Information measure of correlation	Out class
Energy	1	0.969	-0.296	0.566
Maximum probability	0.969	1	-0.276	0.431
Information measure of correlation	-0.296	-0.276	1	-0.459
Out class	0.566	0.431	-0.459	1

A better representation of the correlation values is presented in Table 2. It is observed that the feature energy shows a high correlation and the corresponding out class seems to be 0.566 . This shows that energy is so significant in categorizing the leaves into diseased from the healthy.

IV. CONCLUSION

Given the limitation of agricultural land bank in India, productivity enhancement is a major factor and mainly depend on disease detection that too at an early stage. Manual detection of plant disease needs expert knowledge which has its own limitation, hence automation of early plant disease is envisaged through this project by way of image processing techniques such as automated vision application [6].

Statistical feature-based analysis is performed over the pre-processed leaves taken from Mendeley Data set. The pre-processing procedure includes conversion of RGB to gray scale and contrast enhancement technique. The feature, energy outclassed in categorizing the leaves as healthy and unhealthy. A correlation of is observed for the energy with the out class when compared with the rest of the features. This seems as a step towards automating the procedure using computer vision. As a future work, aritificial intelligence-based machine learning or deep learning methods could be involved for classification of the classes.

REFERENCES

[1] Spiertz, J. H. J., & Ewert, F. (2009). Crop production and resource use to meet the growing demand for food, feed and fuel: opportunities and constraints. NJAS: Wageningen Journal of Life Sciences, 56(4), 281-300.

[2] Prasanna Mohanty, S., Hughes, D., & Salathe, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. arXiv e-prints, arXiv-1604.

[3] Hassan, S. M., Maji, A. K., Jasiński, M., Leonowicz, Z., & Jasińska, E. (2021). Identification of plant-leaf diseases using CNN and transfer-learning approach. Electronics, 10(12), 1388.

[4] Reddy, S. R., Varma, G. S., & Davuluri, R. L. (2021). Optimized convolutional neural network model for plant species identification from leaf images using computer vision. International Journal of Speech Technology, 1-28.

[5] Kiani, E., & Mamedov, T. (2017). Identification of plant disease infection using soft-computing: Application to modern botany. Procedia computer science, 120, 893-900.

[6] Rout, R., & Parida, P. (2019, September). A Review on Leaf Disease Detection Using Computer Vision Approach. In International Conference on Innovation in Modern Science and Technology (pp. 863-871). Springer, Cham.

[7] Wijerathne S. D., Paheerathan, V. & Sivakanesan, R. (2020). Evaluate the effectiveness of Pongamia pinnata seed powder on Pityriasis versicolor (Themal). IAR Journal of Medical Sciences, 1(3), 169-175.

[8] Chouhan, S. S., Singh, U. P., Kaul, A., & Jain, S. (2019, November). A data repository of leaf images: Practice towards plant conservation with plant pathology. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 700-707). IEEE.

[9] Savithri, C. N., & Priya, E. (2019). Statistical analysis of EMG-based features for different hand movements. In Smart Intelligent Computing and Applications (pp. 71-79). Springer, Singapore.

[10] Shanthakumari, G., & Priya, E. (2022). Interpretation of Lung Sounds Using Spectrogram-Based Statistical Features. In Futuristic Communication and Network Technologies (pp. 815-823). Springer, Singapore.

[11] Chitra, R., & Priya, E. (2021). Digital Filter Implementation for Removal of Baseline Wander in ECG Signals. In Advances in

Automation, Signal Processing, Instrumentation, and Control (pp. 2711-2718). Springer, Singapore.

[12] Velvizhi, V. A., & Priya, E. (2022). A Preprocessing Techniques for Seismocardiogram Signals in Removing Artifacts. In Futuristic Communication and Network Technologies (pp. 845-853). Springer, Singapore.

[13] Geetha, R., & Priya, E. (2021). Optimization-based multilevel threshold image segmentation for identifying ischemic stroke lesion in brain MR images. In Handbook of Decision Support Systems for Neurological Disorders (pp. 223-244). Academic Press.

[14] Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, (6), 610-621.

[15] Priya, E., Srinivasan, S., & Ramakrishnan, S. (2012). Differentiation of digital TB images using texture analysis and RBF classifier. Biomedical sciences instrumentation, 48, 516-523.

[16] Nikoo, H., Talebi, H., & Mirzaei, A. (2011, November). A supervised method for determining displacement of gray level co-occurrence matrix. In 2011 7th Iranian conference on machine vision and image processing (pp. 1-5). IEEE.

[17] Kent State University Libraries. (2017). SPSS tutorials: Independent samples t test.