Research

An Automatic Leaf Disease Detection System for Legume Species

Sukhvir Kaur[1*], Shreelekha Pandey[1*♣], Shivani Goel[2]

[1] Department of CSE, Thapar University, Patiala, Punjab, India

[2] Department of CSE, Bennett University, Greater Noida, U.P., India

*Correspondence should be addressed to Sukhvir Kaur, Department of CSE, Thapar University, Patiala, Punjab, India; Tel: +919592640250; Fax: +919592640250; Email: sukhvir.kaur@thapar.edu.

♣Co-Correspondence should be addressed to Shreelekha Pandey, Department of CSE, Thapar University, Patiala, Punjab, India; Tel: +919993566727; Fax: +919993566727; Email: shreelekha.pandey@thapar.edu.

ABSTRACT

Legumes are crucial species which are used by the community worldwide. In this manuscript, a two stage approach to identify infected leaf region percentage in legumes (particularly Groundnut and Soybean) is proposed. First stage classifies between a healthy and a diseased leaf sample. Second stage detects the diseased region and identifies the proportion of leaf infected area. The two stage approach provides high accuracy and also, shows that texture features plays an important role for classification of healthy and diseased leaves. The experimental results obtained on a self-collected leaf image dataset show that the proposed approach accurately identifies the diseased region in legumes. The proposed methodology can also be used for the classification of different disease types.

Key words: Machine learning, Pattern recognition, Plant disease.

1. INTRODUCTION

Agriculture is one of the main sources which play a key role in economic development of any country. Damage to these plants leads to loss in production that ultimately affects the economy (1). Leaf is the most sensitive part which shows disease symptoms the earliest. Initially a time consuming technique, i.e. a manual measurement by plant pathologists is used for detection. In the past few years, a range of applications have utilized the concepts from machine vision and pattern recognition domains to develop intelligent systems. This fact motivates researchers to employ the technology in the field of agriculture for automatic detection of leaf diseases caused by various biotic and abiotic stresses (2, 3). This involves development of automatic and semi-automatic systems. Automatic systems do not require human intervention but semi-automatic does. Not only detection but measuring the disease severity is also important (4). In addition, a leaf may get infected by more than one type of disease (5, 6). But most of the existing works consider that the leaf contains only one type of disease, which is not always possible. It is also observed that plants are infected mainly by fungus (7). From a list of crops, this manuscript is focused on Groundnut and Soybean. Groundnut plays a major role in bridging the vegetable oil deficit in most of the countries and usually suffers from *cercospora* (8). Soybean is oilseed crop which is rich in protein and mainly suffers from fungus such as *rust*, *spots* and *powdery mildew* (9). Both the crops are important parts of a nutritious diet, thus their prevention from diseases as well as detection of diseases at an early stage (if infected) is very much required. Also, literature shows that these crops are less explored as compared to others like wheat, rice, corn, and cotton. This work thus presents an automatic leaf disease detection system for Groundnut and Soybean. The developed system assumes that the diseased region may contain symptoms of different diseases, which would further be used for disease classification. The rest of the manuscript is organized as follows. Section 2 presents a literature survey of systems for Groundnut and Soybean. A two stage system that classifies healthy and diseased leaves is explained in Section 3. The system also identifies the disease region and reports the percentage of leaf portion

infected. Results and comparative evaluations are included in Section 4 and lastly, the work is concluded in Section 5.

2. MATERIALS AND METHODS

2.1. Related Work

To the best of our knowledge, research carried out to detect leaf diseases in Groundnut is limited to a very few studies (8, 10, 11). Utilizing the concepts of geometric moments, a work attempts to identify and analyze deficiencies in groundnut leaves (10). Similarly, another work explores the traditional thresholding operations to detect and determine the proportion of leaf infected by two fungi types: Cercosporidium personatum and Cercospora arachidicola (11). A work has also attempted to increase the number of diseases that can be detected by a single system (8). The system is based on back propagation neural network and used a large proportion of the dataset images (90%) for training (8). The scenario with Soybean crop is a bit different as it is much more focused in literature than Groundnut. A study has used only 30 infected leaf images of Soybean in a neural network based system and achieved an accuracy of 93% (4). A severity grading system is developed using k-means clustering to automatically detect three disease types (*bacterial leaf blight*, *septoria brown spot*, and *bean pod mottle virus*) (12). Efficacy of the system is evaluated by comparing the obtained results with those measured using a manual technique. Instead of texture, shape features are also used to differentiate healthy leaves from diseased leaves (13). In another system k-nearest neighbor method is used to classify between *frog eye* and *brown spot* with an average accuracy of 75% (14). One more system aims to study color distribution and pixel relationship at every stage of disease growth (9). Observations are made for 25 days using local as well as global features of *rust* infected leaf images. Percentage disease index (PDI) based on severity levels is computed to categorize a *rust* disease. Minimum PDI of 0.2 and maximum of 95.5 are observed on 6th and 25th day respectively. The study reveals that higher PDI indicates a decrease in spatial relationship between color and gray pixels due to lesser contribution of green color region. Most of the existing systems focus on determining the type of infection present in a leaf. A few of them attempted to increase the performance of a system in computing the percentage of infected leaf. Also, these studies have utilized datasets consisting of leaf images at each stage of a disease. Such image collections are actually difficult to obtain. Considering all the above points, this study focuses mainly on increasing the performance of a system in determining the disease region and finding the proportion of infected leaf in an image irrespective of disease stage. Because if an initial stage infection is present all over the leaf then it is equivalent to severe damage that may occur at later stages of a disease.

2.2. Proposed system

The system developed in this work makes use of a dataset collected manually from the field areas of agricultural university. This section presents the data collection phase followed by the background segmentation before explaining the actual modules of the developed system. The architecture of the system proposed in this work is shown in Figure 1. After background segmentation a SVM classifier is utilized to differentiate a healthy leaf from as infected leaf sample. If the leaf is infected then k-means segmentation is applied. The proposed method then used to determine an infected region and percentage of leaf area infected.

2.2.1. Data collection

The leaves collected from the field areas contain both healthy and fungus infected leaves for Groundnut and Soybean crops. The leaves are scanned with high resolution scanner. Dataset 1 is Groundnut dataset consisting of 150 leaves (74 healthy and 76 infected leaves). Dataset 2 contains 150 soybean leaves, out of which 74 are healthy and 76 are fungus infected leaves.

2.2.2. Background segmentation

During scanning, color variation (from white to gray) in the leaf background is observed and is required to be removed before any further processing. The procedure adopted for background elimination is shown in the first row of Figure 1.
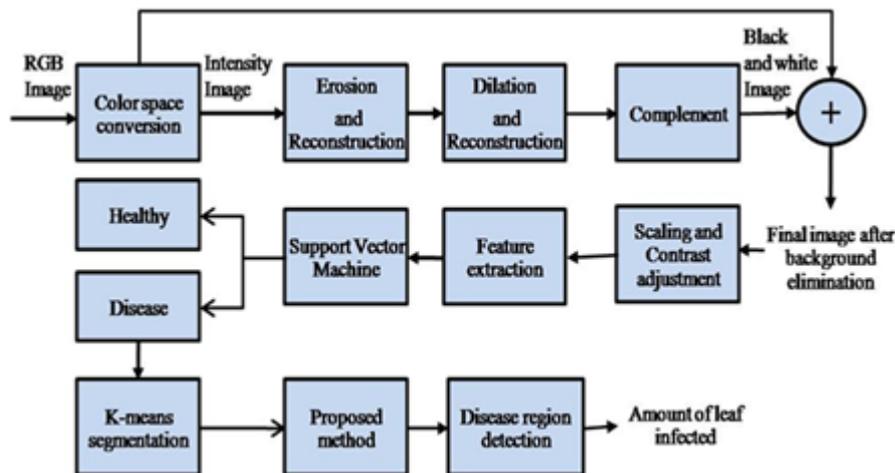
Figure 1. Workflow of the proposed plant disease detection system

The original RGB image is first converted into intensity image followed by opening-closing reconstruction operations. The resulting black and white image is finally overlapped with the original image to eliminate the background. The output for an example Soybean image along with all the intermediate module outputs is shown in Figure 2.
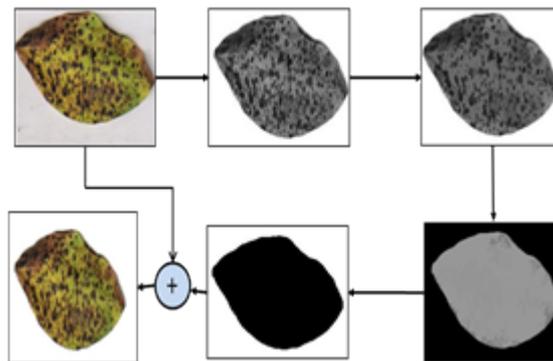
Figure 2. Background elimination operation on an infected Soybean leaf

The image can now be used for further processing in the developed system.

2.2.3. Pre-processing and feature extraction

In preprocessing phase, the image is first enhanced by proper adjustment of intensity values and then the conversion into gray scale is done. Leaf images have patterns and texture features are considered the best to distinguish among patterns. Following this fact, a gray-level co-occurrence matrix (GLCM) is used to measure a leaf texture. GLCM is a statistical approach that considers the pixel spatial relationship (15). Thirteen texture features, viz. correlation, contrast, homogeneity, energy, mean, entropy, standard deviation, root mean square, kurtosis, variance, skewness, smoothness, and inverse difference moment are extracted for each image. A short description of these parameters is included in Table 1.

Table 1. Texture features

Feature	Description	Equation
1.	Contrast	$C = \sum_{i,j} |i - j|^2 P(i,j)$
2.	Correlation	$Co = \sum_{i,j} \dfrac{(i - \mu_i)(j - \mu_j)P(i,j)}{\sigma_i \sigma_j}$
3.	Energy	$E = \sum_{i,j} P(i,j)^2$
4.	Homogeneity	$H = \sum_{i,j} \dfrac{P(i,j)}{1 + |i - j|}$
5.	Mean	$\mu = \dfrac{1}{mn} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} P(i,j)$
6.	Standard Deviation	$\sigma = \sqrt{\dfrac{1}{mn} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} (P(i,j) - \mu)^2}$
7.	Smoothness	$S = 1 - \dfrac{1}{1 + \sigma^2}$
8.	Entropy	$E = -\sum_{k=0}^{L-1} P_k (\log_2 P_k)$
9.	RMS	$R = \dfrac{\sqrt{\sum_{i=1}^{m} |\mu_{ij}|^2}}{m}$
10.	Variance	$V = \sqrt{\sigma}$
11.	Kurtosis	$K = \left\{ \dfrac{1}{mn} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} \left(\dfrac{P(i,j) - \mu}{\sigma} \right)^4 \right\} - 3$
12.	Skewness	$S = \dfrac{1}{mn} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} \left(\dfrac{P(i,j) - \mu}{\sigma} \right)^3$
13.	Inverse difference moment	$IDM = \sum_{\substack{0 \le i \le m-1 \\ 0 \le j \le n-1}} \dfrac{P(i,j)}{1 + (i - j)^2}$

Here, the size of an image is $m \times n$ with L different gray level intensity values. $P(i,j)$ is the intensity value of any pixel within an image and P_k is the probability of occurrence of the k-th gray level intensity. The obtained feature values are then used to train and test the classifier.

2.2.4. Training and testing

This work utilizes classifiers from supervised learning category. Literature supports the usage of support vector machines (SVM) in comparison to other traditional classifiers because of several advantages. SVM is less intensive computationally than neural network (16). In high dimensional space as well, SVM provides better performance and prevents over fitting. It works well with noisy data. Moreover its estimates are based on minimization of test error rather than training error. In SVM, a function $g(x)$ is obtained to find a hyperplane which suitably splits the data point inputs from two classes. The optimal hyperplane is obtained as in Eq. (1).

$$g(x) = \sum_{t \le n \le 1} z_n a_n (x_n + x) + c \qquad (1)$$

Where, z_n is a support vector with Lagrange multiplier having non-zero value (a_n), c is a threshold value, n is input data number, and x_n is input data point (17). Numerous advantages of SVM make this classifier straightforward and effective for various classification purposes. Hence SVM is selected in the proposed work for the identification and classification of healthy and infected leaves. After classifying a leaf image as an infected sample the system further identifies the fungus infections considered in this work. The procedure adopted is explained in the following section for disease region detection. In addition to classification the infected region is detected and presented in terms of percentages of region infected.

2.2.5. Disease region detection

The leaf image identified as an infected sample is gone through the clustering module that utilizes the concept of k-means algorithm. Clusters are formed on the basis of colors as it is observed that the color of an infected leaf region varies with the differences in attacks by various fungus types. The k-means algorithm generates three clusters. First is the background cluster (with dark spots disease region), second is the diseased cluster (with lighter color spots), and third is the leaf cluster (with green portion of a leaf). Each of the three clusters is then converted into a gray scale image before extracting the infected leaf region. The step by step procedure for extraction is explained as follows:

Let, $I_{backgroundremoved}$ is the image obtained after the background removal but before the preprocessing step. $cluster_k$, where $k = 1:3$, is the gray scale image generated by k-means algorithm. $max(cluster_k(:))$ gives the maximum gray level value in the cluster color image $cluster_k$. Thus,

$Max_A = max(cluster_1(:))$
$Max_B = max(cluster_2(:))$
$Max_C = max(cluster_3(:))$
$if \ (Max_A < Max_B) \ \&\& \ (Max_A < Max_C)$
$O = cluster_1$
$else \ if \ (Max_B < Max_A) \ \&\& \ (Max_B < Max_C)$
$O = cluster_2$
$else \qquad O = cluster_3$
$I_{infected} = I_{backgroundremoved} - O$

The final infected cluster image is stored and here it represented as $I_{infected}$. Figure 3 (a) – (c) shows the three clusters obtained using k-means clustering algorithm and the final image stored in $I_{infected}$ is shown in Figure 3 (d).
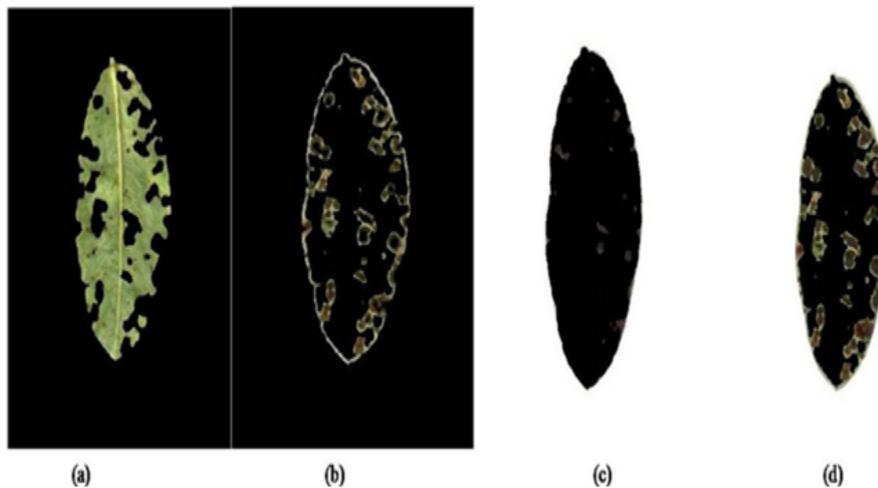
Figure 3. The output of the k-means clustering algorithm (a) leaf cluster, (b) diseased cluster, (c) background cluster, and (d) finally selected infected cluster

2.2.6. Estimation of infection in percentage

After determining the maximum infected cluster as explained in the previous section, the percentage of leaf infection is calculated using the formula shown in Eq. (2).

$$A = 1 - \frac{A1}{A2} \times 100 \qquad (2)$$

Where, A_1 are black pixels after disease region detection

and A_2 are white pixels in the leaf area of the leaf. Results for a few images from the dataset considered in this work are shown in Table 2.

Table 2. Results of the proposed system

Leaf image after background elimination	Diseased region after k-means clustering	Percentage of leaf infected
		95%
		5%
		89%

It can easily be observed that the infected leaf area percentage obtained automatically by the proposed system matches with the actual visible results.

3. RESULTS AND DISCUSSION

Both the datasets used in this work consists of healthy and diseased leaves. The infected images of Groundnut and Soybean crops belong approximately to the single stage of a disease. Three experimental model datasets are created by varying the number of training and testing images. Model 1 contains 50% training and 50% testing images. Model 2 contains 60% training and 40% testing images. Model 3 contains 70% training and 30% testing images. Table 3 shows the classification results obtained with three models for each of the two crops. It is clearly visible that the accuracy of the system increases with an increase in number of training images.

Table 3. Classification results obtained by the proposed system

Models	Healthy or Disease leaves	Number of images		Performance measure (Accuracy)
		Training images	Testing images	
Groundnut				
Model 1	Healthy	37	37	70%
	Disease	38	38	65%
Model 2	Healthy	44	30	94%
	Disease	45	31	90%
Model 3	Healthy	51	23	100%
	Disease	53	23	100%
Soybean				
Model 1	Healthy	37	37	73%
	Disease	38	38	54%
Model 2	Healthy	44	30	93%
	Disease	45	31	89%
Model 3	Healthy	51	23	98%
	Disease	53	23	97%

As a result the maximum classification accuracy of greater than equals to 97% is reported by Model 3. Also minimum accuracy is recorded for Model 1. On closer observation it can be said that Model 2, based on 3:2 ratio of training and testing images, is performing as good as Model 3. The leaf images identified as infected are processed further to determine the diseased region and percentage of infected region in the leaf.

3.1. Comparison with other works

The proposed system is first compared with a method presented by Barbedo to measure early and late leaf spots (11). They applied threshold on yellow and magenta channel after converting the RGB image into CMYK color

model. During thresholding the results are presented only in white and black pixels to measure the amount of leaf infected area. Moreover, this system had considered only diseased leaves. Using the existing methodology it's difficult to view the diseased region in color image. The same is shown in Figure 4 for an image from the collected dataset where the method in (11) does not accurately identify the diseased region.

Figure 4. (a) Original image, (b) Result obtained using (11), (c) Result obtained using the proposed method

These results thus support the usage of the proposed method over for further stages of disease classification. In comparison to, the proposed method first classifies between healthy and diseased leaf, then identifies the diseased region accurately. The identified region can further be used in disease classification and measuring the diseased leaf region (11). In another system shape features are extracted for classification between healthy and diseased leaves (13). The accuracy obtained using is less as compared to the proposed method which uses texture features. The comparative results for Model 2 and Model 3 are shown in Figure 5.
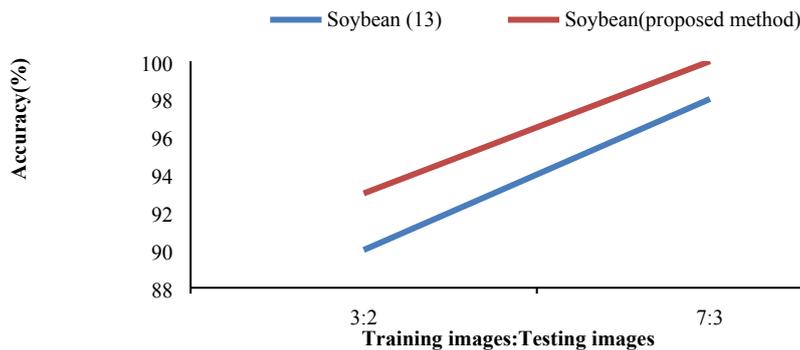
Figure 5. Performance comparison (in term of accuracy) with the existing system in (13)

These results also indicate that texture features always supersede other features in disease identification systems for plants.

4. CONCLUSION

This research presents an approach to identify the infected leaf region and also reports the amount of leaf region infected in percentages. Diseases such as leaf spots caused by different types of fungus are considered in this work. This approach can further be used for classifying various types of diseases. The proposed work is found to work better than other exiting approaches. However, the system at its present status is not 100% accurate as some discrepancies are observed with images showing a less infected leaf. In future, these results can be improved and also, the leaves with complex background can be taken into account with different environmental conditions.

Moreover, GLCM feature vector consumes vast computation cost. Thus in future an attempt to reduce the length of feature vector can also be made. Lastly, identification of diseases in real environmental conditions is a challenging task. That can also be focused in the near future.

FUNDING/SUPPORT

AUTHORS CONTRIBUTION

This work was carried out in collaboration among all authors.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this paper.

REFERENCES

1. Savary S, Ficke A, Aubertot J-N, Hollier C. Crop losses due to diseases and their implications for global food production losses and food security. Food Security. 2012;4(4):519-37.

2. Martinelli F, Scalenghe R, Davino S, Panno S, Scuderi G, Ruisi P, et al. Advanced methods of plant disease detection. A review. Agronomy for Sustainable Development. 2015;35(1):1-25.

3. Wahabzada M, Mahlein A-K, Bauckhage C, Steiner U, Oerke E-C, Kersting K. Plant phenotyping using probabilistic topic models: uncovering the hyperspectral language of plants. Scientific reports. 2016;6.

4. Gharge S, Singh P. Image Processing for Soybean Disease Classification and Severity Estimation. Emerging Research in Computing, Information, Communication and Applications: Springer; 2016. p. 493-500.

5. Husin ZB, Shakaff AYBM, Aziz AHBA, Farook RBSM, editors. Feasibility study on plant chili disease detection using image processing techniques. Intelligent Systems, Modelling and Simulation (ISMS), 2012 Third International Conference on; 2012: IEEE.

6. Lu J, Cui D, Jiang H. Discrimination of tomato yellow leaf curl disease using hyperspectral imaging. ASABE Paper. 2013(131597646).

7. Mendgen K, Hahn M. Plant infection and the establishment of fungal biotrophy. Trends in plant science. 2002;7(8):352-6.

8. Ramakrishnan M, editor Groundnut leaf disease detection and classification by using back probagation algorithm. Communications and Signal Processing (ICCSP), 2015 International Conference on; 2015: IEEE.

9. Pujari JD, Siddarammayya Yakkundimath R, Jahagirdar S, Byadgi A. Quantitative detection of soybean rust using image processing techniques. Journal of Crop Protection. 2015;5(1):75-87.

10. Nisale SS, Bharambe CJ, More VN. Detection and Analysis of Deficiencies in Groundnut Plant using Geometric Moments. World Academy of Science, Engineering and Technology. 2011;58.

11. Barbedo JGA, editor Automatically measuring early and late leaf spot lesions in peanut plants using digital image processing. CONGRESSO BRASILEIRO DE AGROINFORMÁTICA; 2013.

12. Jadhav SB, Patil SB. Grading of Soybean Leaf Disease Based on Segmented Image Using K-means Clustering. IAES International Journal of Artificial Intelligence (IJ-AI). 2016;4(1).

13. Dandawate Y, Kokare R. An automated approach for classification of plant diseases towards development of futuristic Decision Support System in Indian perspective. In: Int. Conf. on Advances in Computing, Communications and Informatics. IEEE; 2015, p. 794-799.

14. Shrivastava S, Hooda DS. Automatic brown spot and frog eye detection from the image captured in the field. American Journal of Intelligent Systems. 2014;4(4):131-4.

15. Gavhale KR, Gawande U. An Overview of the Research on Plant Leaves Disease detection using Image Processing Techniques. IOSR Journal of Computer Engineering (IOSR-JCE). 2014;16(1):10-6.

16. Amsaveni V, Singh NA, Dheeba J, editors. Application of Support Vector Machine Classifier for Computer Aided Diagnosis of Brain Tumor from MRI. International Conference on Swarm, Evolutionary, and Memetic Computing; 2014: Springer.

17. Yu H, Kim S. Svm tutorial—classification, regression and ranking. Handbook of Natural computing: Springer; 2012. p. 479-506.

Role of IT in Bioinformatics and Neuroinformatics
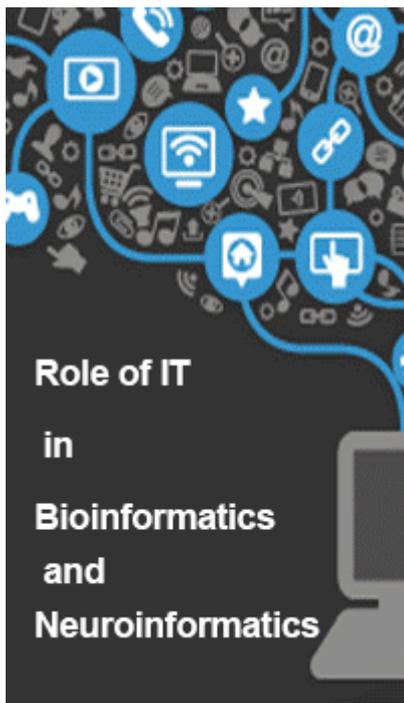
Special Issue on:
'Role of IT in Bioinformatics and Neuro-informatics'

Guest Editors:

Prof. Deepshikha Bhargava
Amity University Rajasthan – India
dbhargava1@jpr.amity.edu

Dr. Ramesh C. Poonia
Amity University Rajasthan – India
rameshcpoonia@gmail.com

Dr. Swapnesh Taterh
Amity University Rajasthan – India
staterh@jpr.amity.edu