Lung Cancer Detection using Nearest Neighbour Classifier

R. Madana Mohana, R. Delshi Howsalya Devi, Anita Bai

Abstract: One of the most precarious diseases is lung cancer. Lung cancer detection is one of the main challenging dilemma nowadays. Most of the cancer cells are overlies with each other. It is tough to detect the cell but also important to identify the existence of cancer cells in the early stage unless unable to prevent. According to 2018 reports, 17 million new lung cancer cases are identified worldwide. The Computer Tomography can be used for diagnosis of cancer with image processing. In this research, we proposed two steps of process for diagnosing the presence of cancer either benign or malignant. In the first step, features are extracted by using GLCM. In the second step, the lung cancer cells are classified either benign or malignant by using Nearest Neighbour classifier. Experimental results demonstrated that the proposed approach performance is 98.76% classification accuracy for diagnosing the lung cancer data.

Index Terms: Lung Cancer, Computer Tomography, GLCM features, NN- Classifier.

I. INTRODUCTION

Lung cancer is also known as lung carcinoma. It is the most serious health problem worldwide. There is significant proof showing that the early detection of lung cancer will decrease mortality rate [24]. Lung cancer is cause due to uncontrolled growth of abnormal cells in one or both of the lungs. It is compulsory to treat this to avoid spreading its enlargement by metastasis to other parts of the body. Early finding of lung cancer is done by using many image processing techniques such as Sputum Cytology, Chest X-ray, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). But, most of these techniques are costly and time consuming. Hence, there is a great need of a new technology to diagnose the lung cancer in its early stages. Image processing method provides an excellent tool for cultivating the manual analysis. There are two classes of tumor (i) non-cancerous tumor (benign) and (ii) cancerous tumor (malignant).

One of the image processing tool is by using MATLAB. The input image is taken in Jpeg format. Pre-processing and segmentation are done thoroughly to segment the cancer affected parts. Classification is done to classify whether the image is normal or abnormal. NN classifier compares the given image with the database images; the tumor is identified by taking in count all the pixels. The feature extraction is a main process in recognition applications and classifications. Normally several texture based feature extraction classifications are used such as GLCM, LBP and SLBP.

Remainder of the paper is organized as follows. Section II summarized a literature review of the existing work for lung cancer. Section III focused on proposed methodology and implementation. Section IV describes the experimental results and performance measures of the proposed NN algorithm with existing classifiers. Section V discussed about applications of proposed research. The conclusion and a brief discussion of opportunities for future work are presented in Section VI.

II. RELATED WORK

A number of authors has developed and implemented diagnosing of the lung cancer by using various methods and algorithms of machine learning and image processing. Aggarwal et al [25] proposed a model that gives normal lung anatomy structure and nodules classification. Their model extracts statistical, geometrical and gray level characteristics. Linear discriminant analysis is used as classifier and minimal thresholding for segmentation. Observations show 84%, 97.14% and 53.33% accuracy, sensitivity and specificity.

Jin et. al [21] have implemented a convolution neural network as classifier using CAD system to detect the lung cancer. Their experimental results proved that accuracy of 84.6%, sensitivity of 82.5% and sensitivity of 86.7%. The main advantage of the proposed model is that it uses circular filter in Region of interest (ROI) extraction phase which can shrink the cost of training and recognition steps.

Sangamithraa and Govindaraju [20] have implemented a k-mean learning algorithm for clustering. During clustering, dataset are grouped according to certain characteristics. During classification this model implements back propagation network. To improve the accuracy, author used a median filter for image pre-processing to detect and remove noise.

Janee Alam et. al [2] have developed a automated system to detect the presence of cancer using Support Vector Machine (SVM) classifier. Their experimental result showed that SVM classifier achieved a classification accuracy of 95% for diagnosis of lung cancer.

Moffy Vas and Amita Desai [1] have implemented a methodology to remove the noise by median filter and then classify the cancer data into benign or malignant using feed forward artificial neural network classification. ANN classifier achieved a classification accuracy of 92% based on their experimental results for diagnosing the lung cancer data.

Kumar et. al [3] have developed a CAD system for detecting lung cancer.

Revised Version Manuscript Received on 16 September, 2019.

Dr. R. Madana Mohana, Professor, Computer Science & Engineering, Bharat Institute of Engineering and Technology, Hyderabad, Telangana. madanmohanr@biet.ac.in

Dr. R. Delshi Howsalya Devi, Associate Professor, Computer Science & Engineering, Bharat Institute of Engineering and Technology, Hyderabad, Telangana. delshi@biet.ac.in

Dr. Anita Bai Associate Professor, Computer Science & Engineering, Bharat Institute of Engineering and Technology, Hyderabad, Telangana. anitabai@biet.ac.in

They used a wavelet transform for pre processing and fragmentation. Their experimental results proved that, proposed algorithm achieved a classification accuracy of 86% for detecting lung cancer.

Joel George and Anitha Jeba Kumari [4] have proposed a algorithm for pre-processing that uses thresholding and histogram equalization. They used a K-means algorithm for clustering the dataset into number of clusters. Also applied a PSO method. This is used to measure PSNR and MSE [5].

Vesna Zeljkovic, and Milena Bojic [6] have developed a algorithm to find the abnormalities from radiography images by using similarity co-efficient. Rabia Almamlook [22] used Random Forest Based Decision tree algorithms to predict lung cancer with 85% accuracy.

From this survey, no one differentiate the affected part of lung from the original CT scans. Our proposed algorithm successfully reported the cancer affected region.

III. METHODOLOGY AND IMPLEMENTATION

Pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features for further processing. It is needed to minimizing the distortion effects identified as light fluctuation in imaging device, to remove blueness and in the same time pre-processing is required to remove unwanted areas from the images and some time it is used for enhancing the image features like lines, boundaries and textures of image so that we can easily divide the contents of images in two parts, wanted and un-wanted contents of image.

For removing noise from the image, many researchers use different filtering techniques which depend on type of noise. In medical imaging all types of filtering techniques may be used depending on noise present in image [1]. Details are given below:

(a) **Gaussian Noise**: Outside the Normal distribution values, usually we cannot see in the image.

(b) **Salt and Paper Noise**: Tiny white and black points randomly appear in the image.

(c) **Poisson Noise**: In Poisson distribution, mean and variance are equal. Noise is present due to Non-linear response of image detectors and recorders.

(d) **Impulse Noise**: Usually it appears in the result of electromagnetic interference, scratches on the recorded disks

(e) **Speckle Noise**: Appearance of waves which are found in many microscopic diffused reflections which create hurdles to understand the image components. This noise follow Gamma distribution found in ultrasound, CT scan and SAR (Synthetic Aperture Radar)images. De-Noising techniques categorized in two parts [2-3].

(i) Spatial Domain Filtering

(a) Linear Filter i.e. Wiener Filter or Mean Filter

(b) Non-Linear Filter i.e. Median Filter

(ii) Transform Domain Filtering i.e. Wavelet Transform.

The proposed work flow of execution is depicted in Fig. 1. Different phases are as follows:

A. Data collection
B. Pre-processing
C. Segmentation
D. Feature Extraction
E. Classification

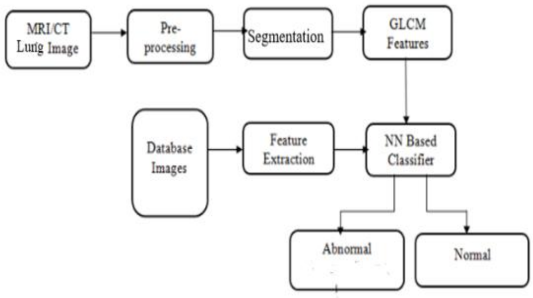

Fig.1. Flow of Execution

A.Data collection:

The input image is collected from kaggle.com website. During classification images are classified into normal or abnormal images which are showing in Fig. 2.

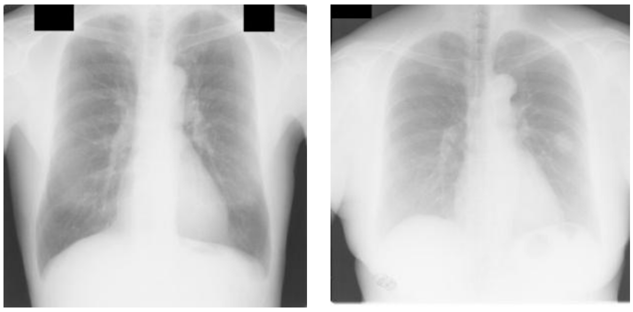

Fig.2. (a) Normal (b) Abnormal

B. Pre-processing:

In preprocessing, input images are taken in jpeg format. It is resized and then converted from RGB to grayscale. The salt and pepper noise from the images are identified by using preprocessing step. Then noises are removed by using median filter which is shown in Fig.3.

C. Segmentation:

During segmentation we performed following major steps:

(1) The image is converted into binary image by Thresholding.

(2) Complementing the image.

(3)Images are masked to fill the holes by super imposing with the original image.

(4)Images are segmented with many angles.

The process of segmentation is shown in Fig.4.

Retrieval Number: B14580982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1458.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

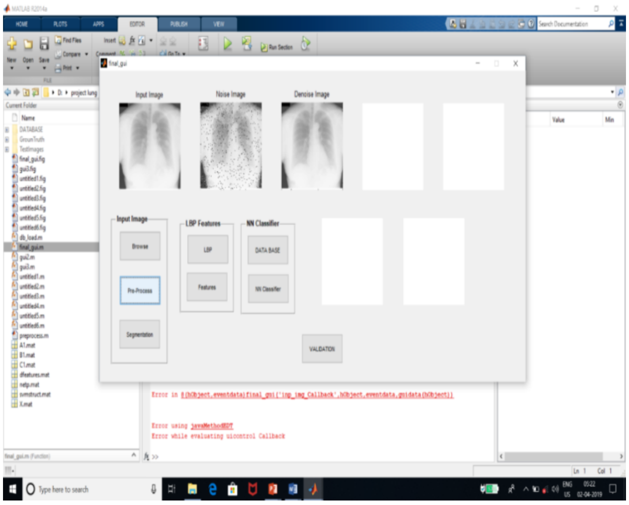

Fig.3. Pre-processing

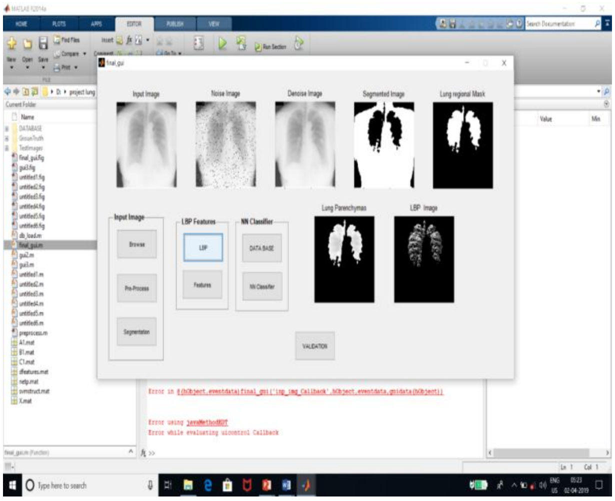

Fig.4. Segmentation

D.Feature Extraction:

The GLCM features are extracted using below mentioned equations. The images will be converted into Gray level matrix.

The p(i,j) represents the pixel values.

Energy= $\sum_i \sum_j p(i,j)^2$ (1)

Energy indicates the similarity level.

Correlation= $\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y}$ (2)

Correlation shows the linear dependency of the gray intensity values in the gray level co-occurrence matrix.

Variance= $\sum_i \sum_j (i-u)^2 p(i,j)$ (3)

Variance measures the spread of intensity values of GLCM pixels about mean.

Entropy= $-\sum_{i=0}^{N_g-1} P_{(x-y)}(i) \log(P_{(x-y)}(i))$ (4)

Contrast= $\sum_i \sum_j (i-j)^2 P(i,j)$ (5)

Contrast measures the intensity variations in the current pixel and its neighboring pixel.

The features are all considered together for the classification. In Fig. 5 shows the features for the ten samples.

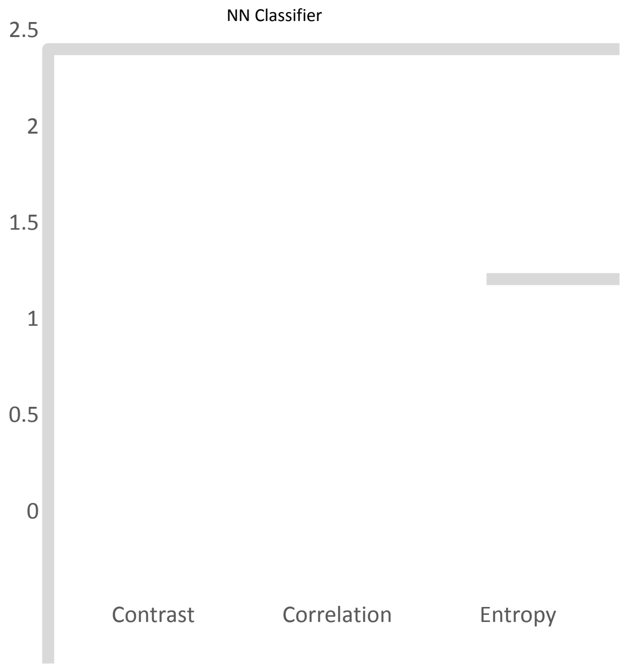

Fig.5. Features

E. Classification:

The fuzzy neural network model is used for classification. Fuzzy system is a learning machine that finds the parameters by exploiting approximation techniques from neural network. Initially the original dataset is trained using the features, then cross checked for the given input image and classifies it as normal or abnormal which is shown in the Fig. 6.

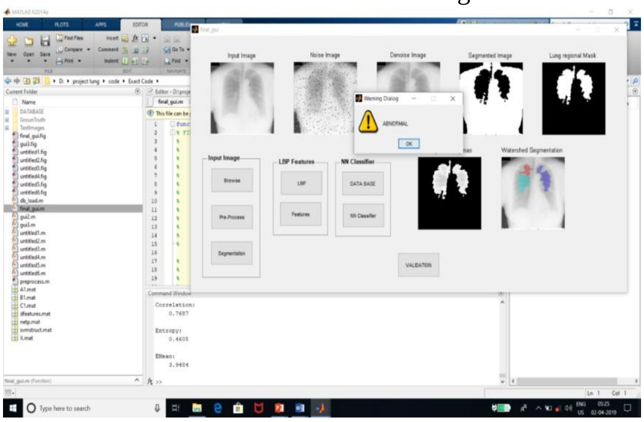

Fig.6. Classification

IV. EXPERIMENTAL RESULTS

The experimental results of diagnosing the lung cancer image dataset have been evaluated and analyzed using NN classifier with DWT segmentation for classification. Metrics used for evaluating the performance of the model are:

- Accuracy
- Sensitivity
- Training Time
- Recognition Time

The same dataset is run on the SVM and random forest classifier. The proposed algorithm achieved a classification accuracy of 98% where the SVM classifier achieved 95% and random forest achieved 94.2%. The following table 1 presents lung cancer diagnosis performance evaluation.

Table 1. Performance evaluation

	NN Classifier	SVM Classifier	Random Forest Classifier
% Correctly Classified Instances	98.76	95.00	94.21
% incorrectly classified instances	1.234	8.163	7.122
Total Training Time	8.6851	146.86	162.71
Total Recognition Time	0.468	0.284	0.356

V. APPLICATIONS

- Proposed methodology is widely used in many medical areas for early diagnosis of cancer. So the accurate treatment will be provided to the patient with benign or malignant.
- This proposed image processing technique can also be used to diagnose other cancer such as breast cancer and tumor with benign or malignant.

Table 2 shows different applications of existing algorithms.

VI. CONCLUSION AND FUTURE WORK

The current preeminent model has no satisfactory result of accuracy and does not classify degree of cancer of detected nodules. Therefore we presented a new approach to diagnose lung cancer. Using the proposed approach the cancerous nodule from the lung CT scan image is detected. In this paper, removal of noise by using median filter and segmenting which gives the efficient results is achieved. The NN classification is also proved that it is better than the classifiers SVM and random forest. This can be still improved using the various combinations of feature selection and classification techniques for determination of relevant subset of features.

Table 2. Different applications of existing algorithms

Techniques	Applications
Gabor Filter	Optical character recognition
Image Processing and Classification	Remove Gaussian white noise
Weiner Filter	Noise reduction,Signal detection.
Layer Separation	Used to separate layer of image
Gray scale Image	Used to convert color in gray
Enhancement	Used to sharpen the image
Gabor filter	Feature extraction
Gabor Filters, Discrete Wavelet Transform and Auto Enhancement Algorithm	Identify Cancerous Cells
Fast Fourier Transform	Image reconstruction
Sparsity-based image modeling	Image Layer Separation
Edge detection-based methods	Lane edge detection
	Canny algorithm
Matching	Local matching
	3D Elastic matching
Classification	Cellular dependency
Support Vector Machine, Fuzzy C-Mean, Conventional Neural Networkand Computer Aided Design	Segmentation
Wiener filter	Image Restoration
Gray conversion	Histogram equalization
Image segmentation	Labeling
Thresholding	Deep leaming algorithms and convolutional networks
Region-based segmentation	Region growing
	Region splitting and merging
Clustering techniques	Seed Point Selection Algorithm
Morphological segmentation	Watershed algorithm
	Cell nuclei
Weibull multiplicative model	Image Segmentation
Marker-controller segmentation	Magnetic Resonance Imaging
	Watershed
Classification	Support Vector Machine
Classification	Supervised and Unsupervsed Tumor Characterization
Classification	Multi-label Classification

REFERENCES

[1]. Moffy Vas, Amita Desai, "Lung cancer detection system using lung CT image processing", IEEE, 2017.

[2]. Janee Alam, Sabrina Alam, AlamgirHossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-Class SVM Classifier", IEEE, 2018.

[3]. S. K. Kumar, J. Ramesh, P. T. Vanathi, K. Gunavathi, "Robust and automated lung nodule diagnosis from CT images based on fuzzy systems," International Conference On Process Automation Control and Computing, Coimbatore, India, IEEE, 2011.

[4]. Joel George R, Anitha Jeba Kumari D, "Segmentation and Analysis of Lung Cncer images using Optimization Techniques", IJEIT, 2014.

[5]. J. Marcello, F. Marques and F. Eugenio, "Evaluation of thresholding techniques applied to oceanographic remote sensing imagery," SPIE, 5573, pp. 96-103, 2004.

[6]. Vesna Zeljkovic, Milena Bojic, "Automatic Detection of Abnormalities in Lung Radiographs caused by planocellular Lung Cancer", IEEE, 2011.

[7]. N. Panpaliya, N. Tadas, S. Bobade, R. Aglawe, and A. Gudadhe, "A survey on early detection and prediction of lung cancer," International Journal of Computer Science and Mobile Computing, vol. 4, no. 1, pp. 175–184, 2015.

[8]. Asuntha, N. Singh, and A. Srinivasan, "PSO, genetic optimization and SVM algorithm used for lung cancer detection," Journal of Chemical and Pharmaceutical Research, vol. 8, no. 6, pp. 351–359, 2016.

[9]. P. Bhuvaneswari and A. Brintha Therese, "Detection of cancer in lung with K-NN classification using genetic algorithm," International Conference on Nano materials and Technologies, vol. 10, pp. 433–440, 2014.

[10]. Gindi, A. M., Al Attiatalla, T. A., & Sami, M.M. (2014) "A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Early stage Lung Cancer Diagnosis of chest x-ray images." Journal of American Science, 10(6): 13-22.

[11]. Anita Chaudhary and Sonit Sukhraj Singh, "Multi resolution Analysis Technique for Lung Cancer Detection in Computed Tomographic Images", International Journal of Research in Engineering & Applied Sciences, IJREAS Volume 2, Issue 2 January 2012.

[12]. Q. Song, L. Zhao, X. Luo and X. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images", Journal of Healthcare Engineering, vol. 2017, pp. 1-7, 2017.

[13]. Valente, P. Cortez, E. Neto, J. Soares, V. de Albuquerque and J. Tavares, "Automatic 3D pulmonary nodule detection in CT images: A survey", Computer Methods and Programs in Biomedicine, vol. 124, pp. 91-107, 2016.

[14]. Larkins DB, Harvey W. "Introductory computational science using MATLAB and image processing". Procedia Computer Science. 2010; 1(1):913-9.

[15]. Dubey AK, Gupta U, Jain S. "Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis". Chinese Journal of Cancer. 2016; 35(1):71.

[16]. Laccetti AL, Pruitt SL, Xuan L, Halm EA, Gerber DE. "Prior cancer does not adversely affect survival in locally advanced lung cancer: A national SEER medicare analysis". Lung Cancer. 2016.

[17]. Al-Tarawneh, M.S., "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, Volume 11, No. 21, pp. 147-58, 2012.

[18]. Dr. N. Ganeshan, Dr. k. Venkatesh, Dr. Rama, A. Malathi Palani, "Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data", International Journal of Computer Applications, vol- 1,2010 .

[19]. Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI),DOI: 10.1109/ICACCI.2015.7275773.

[20]. Sangamithraa, P., & Govinda raju, S. (2016) "Lung tumour detection and classification using EK-Mean clustering," 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.

[21]. Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design

[22]. Rabia Almamlook, "Lung Cancer Survival Prediction Using Random Forest Based Decision Tree Algorithms", Proceedings of the International Conference on Industrial Engineering and Operations Management Washington, DC, USA, September 27-29, 2018.

[23]. R.Delshi Howsalya Devi" Effective Diagnosis of Heart Disease using Inter Quartile Range Filter and Decision Tree Classifier", Middle-East Journal of Scientific Research, June 2019.

[24]. Madana Mohana R, Rama Mohan Reddy A, "Machine Learning and Data Mining Techniques for Sign Language Recognition and Retrieval System", International Journal of Engineering Computational Research and Technology (IJECRT), Volume 1, Issue 1, December 2016.

[25]. Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." International Conference On Advances In Computing, Communications and Informatics

AUTHORS PROFILE

Dr. R. Madana Mohana is currently working as Professor in the Department of the Computer Science and Engineering at Bharat Institute of Engineering and Technology, Ibrahimpatnam - 501 510, Hyderabad, Telangana. Received Ph.D in Computer Science & Engineering, Sri Venkateswara University, Tirupati. His research interests include Machine Learning, Data Mining, Bigdata, Information Retrieval and Computational Intelligence. He is a Life Member of ISTE and Member of CSI & ACM. He published 34 papers in journals and conferences of repute, 8 patents were filed and 3 patents are published.

Dr. R. Delshi Howsalya Devi received her B.E in Computer Science and Engineering from the Madurai Kamaraj University, Madurai, in 2004, ME in Computer Science and Engineering from the Anna University, Chennai, in 2008 and PhD in Information and Communication Engineering from the Anna University Chennai in 2018. She is an Associate Professor at Computer Science and Engineering, Bharat Institute of Engineering and Technology, Hyderabad, Telangana, India. She has 13 years of teaching experience and has approximately 23 conference publications and 16 international journal publications. Her research interests include data mining, outlier mining and big data analytics. She published 1 patent and 2 patents were filed.

Dr. Anita Bai received the Ph.D. from Visvesvaraya National Institute of Technology, Nagpur, India and the M.Tech. degree from National Institute of Technology, Rourkela, India. She is working as Associate professor with the Computer Science and Engineering at Bharat Institute of Engineering and Technology, Hyderabad, Telangana, India. She has co-authored a number of research articles in various journals, conferences, and book chapters. She is a member of IEEE and IAENG. Her research interests include data mining, soft computing and machine learning. She published 1 patent related to cyber crime detection and control.