CBIR System for Lung Nodule Retrieval in CT Scans

Ashwini Dasare, Harsha S

Abstract*: Lung cancer remains one of the fatal diseases with very high mortality rate in both men and women. Computer aided diagnostic systems have been contributing towards the enhancement of survival rate to a maximum extent. Most of such systems yield binary results, i.e. they classify whether a nodule is benign or malignant and they are computationally expensive. This paper proposes a methodology to build a Content Based Image Retrieval (CBIR) system that provides additional provision to the domain experts. Since the CBIR systems retrieve most similar images, this visual dimension will assist the budding and experience radiologist to assess the nodule information to greater detail. Nine visual and shape features are extracted for each nodule image collected from LIDC database and Minkowski distance measure is used for computing similarity. Experiments are conducted on 750 nodules out of which 375 are benign and 375 are malignant as identified by domain experts. Precision, recall and F measure metrics are considered to evaluate the methodology with achieved average values of 0.92, 0.82 and 0.86 respectively.

Keywords: **CBIR, nodule, Similarity measure.**

I. INTRODUCTION

Being the primary cause of cancer related deaths; lung cancer continues to haunt mankind. According to Globocon 2018, the lung cancer recorded 2.094 million fresh cases and 1.8 million deaths worldwide. At the outset the only way to enhance the survival rate is by early detection of pulmonary nodules as a nodule may be manifestation of lung cancer. A nodule appears as a rounded mass on a CT scan with a diameter up to 3cm [1]. Typically, a CT scan for a patient would consists of few hundreds of slices (images) and handling such a huge image set manually is tedious task with respect to time consumption and efficient interpretation. To assist the radiologists for the same CAD systems are designed yet they pose some limitations as they work like a black box. In other words, for a given case of CT scan, a very efficient CAD system would provide the decision as the patient being cancerous or not. Though such binary classified results are very important and relevant, little more can be expected by the domain experts. This gap is filled by advent of CBIR systems. A CBIR system provides diagnostic support to the radiologists by letting them to visualize numerous images/scans from the image database that visually most similar to the input query image.

Having a great diagnostic potential, a CBIR system operates in two modes. In offline mode, there exists an image repository and a feature base. The image repository contains large number of images of benign or malignant nodules. With the help of feature extraction process, multiple features are extracted and stored in the feature base. In online mode, the CBIR would accept an input query image of a nodule of benign or malignant nature through a HCI. Feature extraction process is carried out on this image and these features are compared with those in feature base. This comparison is aided by one of the various distance measures. Based on the threshold value visually most similar images are retrieved and displayed for the domain expert through a HCI. This paper proposes an end to end procedure for building a CBIR system for diagnosing lung cancer. This paper is further organized as follows. Section II provides related works in the field, section III illustrates proposed methodology including feature extraction and computation of similarity measure, section IV highlights the experimental results and discussion and lastly section V provides concluding remarks.

II. RELATED WORK

CBIR systems are gaining momentum in the domain of diagnostic radiology [2]. Though the literature contains huge number of research articles in the arena of generic CBIR, this section provides a brief review on contributions of various authors in CBIR systems specific to investigative radiology.

A typical CBIR system has 4phasessuch as submission of input query image, extraction of features, computation of similarity measure and a Human Computer Interface (HCI) to provide the visualization to end users. Feature extraction is key processing stage among all. Diverse features such as color feature, shape features and texture features are extracted during feature extraction phase, and the literature [3, 4, 5, and 6] highlights how these features are extracted. Computation of similarity measure is another defining stage in development of CBIR system. Among various measures of similarity, Euclidean distance, Minkowski distance, Manhattan distance, Cosine similarity, Jaccard similarity are widely used by the researchers. A detailed review about these can be found in [7].

Turning the attention towards role of CBIR in diagnostic radiology in general and CBIR for lung cancer diagnosis in particular, it has been evident that the diagnostic and interpretation accuracy is scaling high time to time with advent of CBIR systems. Since the CBIR system retrieves most similar images, the radiologist as the end user can derive appropriate conclusion by comparing the input query image and the retrieved set of images. The retrieved set can also be used study and reference purpose.

Revised Manuscript Received on December 13, 2019.

***** Correspondence Author

Ashwini Dasare *, Asst. Prof, Dept. of E&C, JSSATE, Bengaluru, India, ashwini.dasare@gmail.com

Dr. Harsha S, Assoc. Prof., Dept. of ISE, Jyothy Institute of Technology, Bengaluru, India, harsha.s@jyothyit.ac.in

Literatures with reference to CBIR systems for lung cancer are abundant. Semantic and visual similarities were used between query image and each image in the database by [8] for similar CT image retrieval. They evaluated their system on CT images of lung. A CBIR system to detect the malignancy levels of lung nodule was proposed by [9]. The features used by them are Nodule density level and nodule lesion density heterogeneity. Their experimentation was on CT images of lung collected from LIDC data set. 26 Haralick texture features were used to construct GLCM in four orientations by [10] Mahalanobis distance measure was used to compute the similarity between the query image of a nodule and reference nodule from the dataset. Margin sharpness descriptors were employed by [11] for retrieval of similar pulmonary nodules. For each nodule in CT section, margin sharpness vector of size 12 is extracted. Both 2D and 3D shape based, margin based and texture features were extracted and stored in the feature repository for CT scans from LIDC by [12] Euclidian, Manhatten and Chebyshev measures are used to compute the similarity. An enhanced watershed thresholding with probabilistic neural network method was proposed by [13] to diagnose lung cancer in their CBMIR system. Deep learning approach for retrieving pulmonary nodules from CT image repository was proposed by [14] and is used as an online assistance system for domain experts.

III. METHODOLOGY

Data Set

The dataset for the experiments is obtained from LIDC IDRI; It is a Lung Image Database Consortium which contains 1018 thoracic CT scans [15]. Each of these scans include images of CT scan and an additional XML file which contains the findings of two stage annotation carried out by four experienced radiologists. The annotation has the ground details of each of CT scans including coordinates of nodules, magnitude, calcification, sphericity, margin, lobulation, speculation and degree of malignancy for each of the nodules. An XML parser program is written and 750 nodules, both benign and malignant nodules 375 each are extracted from the CT scans. These nodule images are then cropped into 50 by 50 pixels and are stored in image base. These images are used for feature extraction and subsequent processing phases.

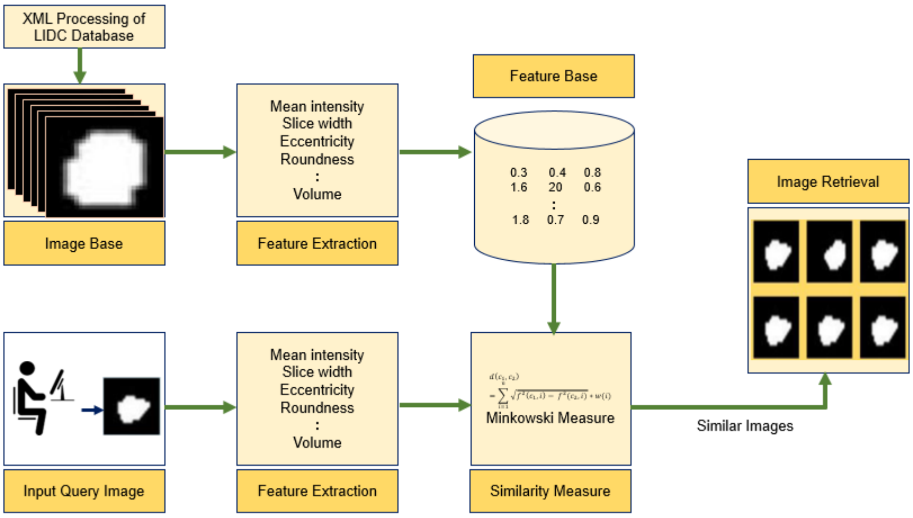


Fig. 1. Architecture of proposed CBIR system

Working Modules

The working principles of proposed approach are depicted in the Fig. 1. Initially a huge image base is created as described in the previous section. This image base contains 50 by 50 pixels of benign and malignant nodule images. For features are extracted in the feature extraction process whose details are given in forthcoming section. The extracted features are stored in the feature base. The steps discussed so far defines the offline mode of the CBIR system.

In the online mode, the user submits an input query image, which can be an image of either benign or malignant nodule. For this image all ten features are extracted but are not stored. These features are then compared with the features of feature base using a similarity measure to determine visually most similar images. Murkowski's distance is sued to achieve this. The hence retrieved images are displayed on the HCI for the user/domain expert for visualization and analysis purpose.

Feature extraction:

The images of nodules stored in image repositories are subjected to feature extraction process during which the inherent feature representation of each of the nodules are extricated and are stored in feature repository with appropriate nomenclature.

LIDC scan number:

This information includes the LIDC case number and a slice number [16]. This will be used to correctly identify the specific CT scans after retrieval. This is as per the LIDC IDRI convention.

The Component number:

The components (nodules/non nodules) in the LIDC CT scans are identified with unique number, which helps in mapping the nodules.

Mean intensity

This value is computed for each and every nodule under observation. The intensity of a pixel is the value of that pixel. Say for an 8-bit gray scale image, the intensity values would vary from 0 to 255, which will be its intensity. In the proposed approach this feature is calculated by taking the ratio of sum of gray scale intensities of pixels of nodule to the aggregation of number of pixels that constitute the nodule as shown in (1).

$$\text{Mean Intensity} = MI(n) = \frac{1}{N}\sum_{i=1}^{x}\sum_{j=1}^{y} I(x,y) \tag{1}$$

Where $MI(n)$ =Mean intensity of nodule n under consideration.
N= Total number of pixels occupied by nodule n
I (x, y) =gray scale intensity

Slice width

The component, if it is a nodule, will span more than one Slices in the CT scan. This count will be very useful for similarity retrieval and is shown in (2).

$$\sum_{z=1}^{l} f(c,z) \tag{2}$$

$$f(c,z) = \begin{cases} 0, & if\ component\ c\ doesnotexistinslicez \\ 1, & otherwise \end{cases}$$

Volume of the nodule

The CT scans in LIDC repository contains an average of 250 Images (slices). The nodules whether benign or malignant span across multiples slices and this information contributes for efficient retrieval to a greater extent. Hence the feature ‘volume’ is considered to be summation of area of the nodule in all the slices of their presence and is shown in (3)

$$\text{Volume of the nodule} = V(n) = \sum_{z=0}^{l}\sum_{y=0}^{n}\sum_{x=0}^{m} n(x,y,z) \tag{3}$$

Where *l*= number of slices that the nodule spans across

Major axis length

This shape measure characterizes the appearance of the component under observation. The major axis length (mal) of the component c is the pixel distance between end points of the major axis and is shown in (4).

$$mal(c) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4}$$

Roundness

It is also called as circularity measure and is shown in (5), which excludes the local irregularity. This area to perimeter ratio equals to 1 if the component is circular and is less than 1 for those objects that deviates from being circular.

$$roundness = \frac{4 \times \pi \times area}{convex\ perimeter^{\ 2}} \tag{5}$$

Eccentricity

This shape descriptor measures the deviation of circularity of a component. Eccentricity as shown in (6) is calculated as a ratio of minor axis length to major axis length. This feature assumes the values between 0 and 1, while 0 for the object which is circular and 1 for a ling segment

$$eccentricity = \frac{Maj\_axis\_Len}{Min\ axix\ Len} \tag{6}$$

Degree of Malignancy

Since the image repository contains the images of benign and malignant nodules, the degree of malignancy (7) is also stored as a feature for those nodules. It is 0 for benign nodules and 1 for the malignant.

$$Deg.Malignancy = \begin{Bmatrix} 0\ if\ component\ c\ is\ benign \\ 1\ if\ component\ c\ is\ malignant \end{Bmatrix} \tag{7}$$

Similarity Measure

The primary goal of developing the proposed CBIR system is to attain visually most similar image retrieval. An input image benign/malignant is fed by the domain expert to the CBIR system. The features as discussed in the previous section are extracted for the input image. However, features of various images of benign and malignant nodules are stored in the feature repository, which is an off line process. Now the features extracted for the input image are compared with those in features repository to compute the similarity measure. Minkowski distance measure (8) is exercised in this approach. Certain threshold is determined, and if the similarity is above this threshold then those images are retrieved and displayed on a GUI as a resultant set of visually most similar images.

The Minkowski distance measure $d(c_1,c_2)$, to compute the similarity between two components c_1 and c_2 is given by

$$d(c_1,c_2) = \sum_{i=1}^{n}\sqrt{f^2(c_1,i) - f^2(c_2,i)} * w(i) \tag{8}$$

Where,n = Number of features,
$f(c,i)$ =Normalizedi^{th} feature of component c,
$w(i)$ =Weight of i^{th} feature

Based on the distance$d(c_1,c_2)$, the degree similarity is decided according to set threshold T by using (9)

$$Degree\ of\ similairty = f(x) = \begin{cases} 0, if\ d(c_1,c_2) > T \\ 1, otherwise \end{cases} \tag{9}$$

IV EXPERIMENTAL RESULTS AND DISCUSSION

The proposed CBIR system is evaluated by conducting series of experiments. To assist the radiologist a Human Computer Interface (HCI) is designed and sample screenshots are as shown in Fig 2 and Fig 3. When a query image (image of benign or malignant nodule) is submitted through the HCI through browsing, the sequence of processing steps including feature extraction, computation of similarity measure as discussed in previous sections would retrieve visually most similar images on the output screen. These retrieved images are labeled with their corresponding LIDC number and the component number. Number of similar images retrieved is also displayed on the output screen.

The performance of the proposed methodology is analyzed and evaluated by considering the metrics: precision, recall and F measure which is the harmonic mean of recall and precision (10, 11, and 12)

$$precision = \frac{numberofsimilarimagesretrieved}{totalnumberofretreivedimages} \tag{10}$$

$$recall = \frac{numberofsimilarimagesretreived}{numbersimilarimagesintherepostory} \tag{11}$$

$$F - measure = 2 \times \frac{precision\ \times recall}{precesion + recall} \tag{12}$$

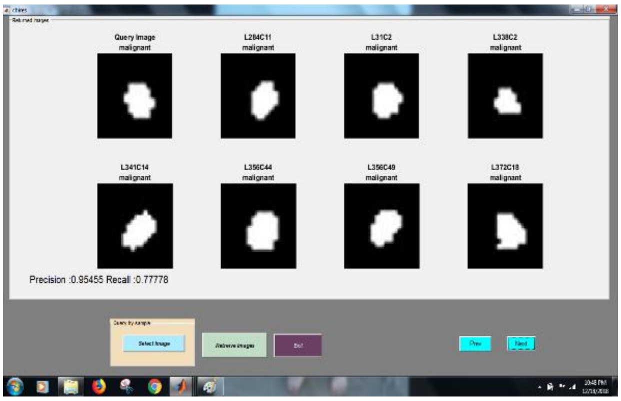
The table I summarizes the average precision, recall and F – Measure values calculated for 10 experiments conducted on various LIDC case numbers.

DOI: 10.35940/ijitee.B1063.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

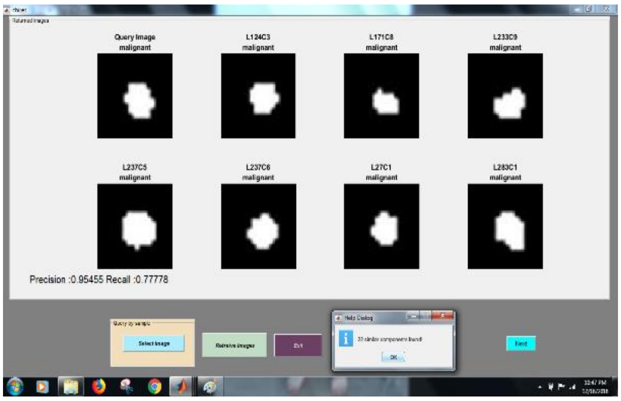
The obtained average values are precision= 0.92, recall= 0.82 and F- measure =0.86 Fig. 4 provides the graphical representation of the same.

The results shown in the Table I are preliminary in nature. The proposed system includes a regression mechanism to account for the errors that occur in the early detection.



Precision :0.95455 Recall :0.77778

Fig. 2 Screenshot of malignant nodule retrieval



Precision :0.95455 Recall :0.77778

32 similar components found!

Fig. 3 Screenshot of malignant nodule retrieval

Table I summary of the experimental values

Exp. No.	LIDC Cases	Precision	Recall	F-Measure
1	LIDC 6	0.96	0.84	0.88
2	LIDC 8	0.90	0.78	0.83
3	LIDC 14	0.92	0.8	0.85
4	LIDC 17	0.92	0.76	0.82
5	LIDC 31	0.93	0.91	0.91
6	LIDC 40	0.90	0.80	0.84
7	LIDC 73	0.92	0.77	0.82
8	LIDC 146	0.92	0.83	0.86
9	LIDC 181	0.91	0.88	0.89
10	LIDC 278	0.96	0.90	0.92
Average		**0.92**	**0.82**	**0.86**

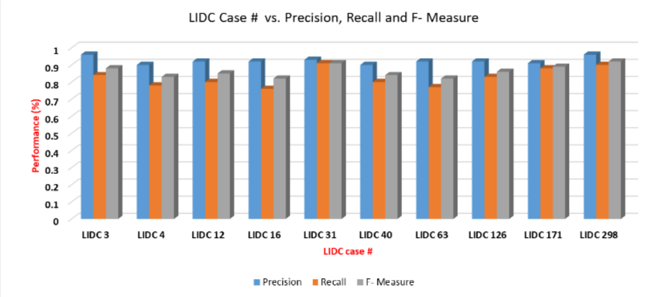


Fig. 4 Graphical representation of performance evaluation

The same is shown in Table II with the graph of the original data and the data after regression in Fig 5. It is evident from the table and the figure that the proposed system has an increased accuracy compared to the systems that have been in existence till date and post regression, it also performs with an enhanced stability

Table II Performance behavior after regression

Exp. No.	Precision	Recall	F- Measure	Precision after regression	Recall after regression	F-measure after regression
1	0.96	0.84	0.88	0.94	0.86	0.90
2	0.9	0.78	0.83	0.92	0.80	0.85
3	0.92	0.8	0.85	0.94	0.82	0.87
4	0.92	0.76	0.82	0.94	0.78	0.84
5	0.93	0.91	0.91	0.95	0.93	0.89
6	0.9	0.8	0.84	0.92	0.82	0.86
7	0.92	0.77	0.82	0.94	0.79	0.84
8	0.92	0.83	0.86	0.94	0.85	0.88
9	0.91	0.88	0.89	0.93	0.90	0.87
10	0.96	0.9	0.92	0.94	0.92	0.90
11	0.97	0.89	0.92	0.95	0.91	0.90
12	0.93	0.86	0.89	0.95	0.88	0.87
13	0.94	0.78	0.85	0.92	0.80	0.87
14	0.96	0.8	0.87	0.94	0.82	0.89
15	0.92	0.81	0.86	0.94	0.83	0.88
16	0.94	0.88	0.9	0.92	0.90	0.88
Average	0.93125	0.830625	0.869375	0.94	0.85	0.87

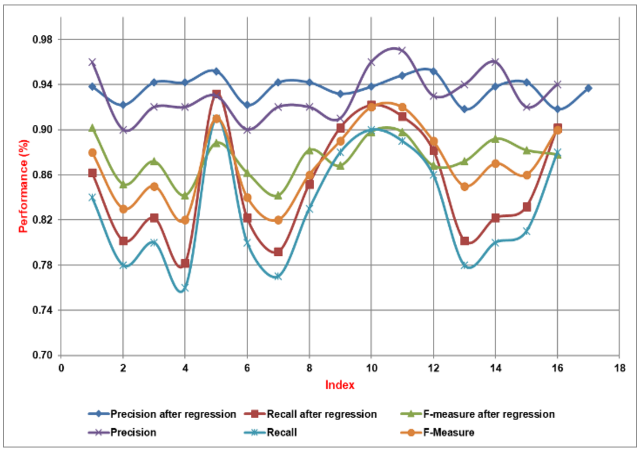


Fig. 5 Representation of performance behavior before and after regression

IV. CONCLUSION

This paper proposed an approach to build a CBIR system for lung nodule retrieval and analysis. The experiments were conducted on nodule images LIDC database which are retrieved through an XML parser. Nine visual and shape based features were extracted and Minkowski distance was used to compute the similarity between the features of input query image and those of the feature base to retrieve similar images. Experimental analysis yielded an average of 0.92, 0.82 and 0.86 for precision, recall and F- measure respectively.

As a future work a huge image base and feature base can be built by considering more nodules from LIDC, JSRT and ECLAP databases. Machine learning techniques are seldom used for CBIR in diagnostic radiology, which opens up the additional future direction.

REFERENCES

1. Bhavanishankar, K., and M. V. Sudhamani. Techniques for detection of solitary pulmonary nodules in human lung and their classifications-a survey. Int. J. Cybern. Inf.(IJCI) 4, no. 1 (2015): 27-40.
2. Akgül, CeyhunBurak, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and BurakAcar, Content-based image retrieval in radiology: current status and future directions, Journal of Digital Imaging vol.24, no. 2, pp.208-222, 2011
3. Akgül, CeyhunBurak, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and BurakAcar, Content-based image retrieval in radiology: current status and future directions, Journal of Digital Imaging vol.24, no. 2, pp.208-222, 2011.
4. G. Schaefer, An introduction to content-based image retrieval, 8th International Conference on Digital Information Management (ICDIM 2013), Islamabad, pp. 4-6, 2013.
5. M. Kaur and N. Sohi, A novel technique for content-based image retrieval using color, texture and edge features, International Conference on Communication and Electronics Systems (ICCES), Coimbatore, pp. 1-7, 2016.
6. Rafiee, Gholamreza, Satnam Singh Dlay, and WaiLok Woo, A review of content based image retrieval, Communication Systems Networks and Digital Signal Processing (CSNDSP), 7th International Symposium on. IEEE, 2010.
7. Cha, Sung-Hyuk, Comprehensive survey on distance/similarity measures between probability density functions, City 1, no. 2, 2007.
8. Ma, Ling, et al., A new method of content based medical image retrieval and its applications to CT imaging sign retrieval, Journal of biomedical informatics 66, pp. 148-158, 2017.
9. Wei, Guohui, et al., A content-based image retrieval scheme for identifying lung nodule malignancy levels, Control and Decision Conference (CCDC), 29th Chinese, IEEE, 2017.
10. Wei, Guohui, He Ma, Wei Qian, Hongyang Jiang, and Xinzhuo Zhao. Content-based retrieval for lung nodule diagnosis using learned distance metric. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3910-3913. IEEE, 2017.
11. Junior, José Raniery Ferreira, and Marcelo Costa Oliveira. Evaluating margin sharpness analysis on similar pulmonary nodule retrieval. In 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, pp. 60-65. IEEE, 2015.]
12. Dhara, Ashis Kumar, SudiptaMukhopadhyay, Anirvan Dutta, Mandeep Garg, and Niranjan Khandelwal. Content-based image retrieval system for pulmonary nodules: Assisting radiologists in self-learning and diagnosis of lung cancer. Journal of digital imaging 30, no. 1 (2017): 63-77.]
13. Shakeel, P. Mohamed, Mohamad IshakDesa, and M. A. Burhanuddin. Improved watershed histogram thresholding with probabilistic neural networks for lung cancer diagnosis for CBMIR systems. Multimedia Tools and Applications (2019): 1-19.]
14. Ibanez, Daniel Perez, et al., Deep Learning for Pulmonary Nodule CT Image Retrieval—An Online Assistance System for Novice Radiologists, Data Mining Workshops (ICDMW), International Conference on. IEEE, 2017.
15. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
16. Bhavanishankar, K., and M. V. Sudhamani. Filter Based Approach for Automated Detection of Candidate Lung Nodules in 3D Computed Tomography Images, In International Conference on Cognitive Computing and Information Processing, pp. 63-70. Springer, Singapore, 2017.

AUTHORS PROFILE

Ashwini Dasare, is working as an Assistant Professor in the Dept. of Electronics and Communications Engineering at JSSATE, Bengaluru. She is an M. Tech degree holder in Digital Electronics and Communication. Her research interest includes Image Processing and Signal Processing

Dr. Harsha S is Associate Professor in Department of Information Science and Engineering, Jyothy Institute of Technology, Bengaluru. He has received Ph.D. from Visvesvaraya Technological University in the field of Computer Science and Engineering. His research field encompasses Cryptography, Networking, Machine Learning and Artificial Intelligence.