International Conference on Communication and Signal Processing, July 28 - 30, 2020, India

Lung Diseases Classification based on Machine Learning Algorithms and Performance Evaluation

Binila Mariyam Boban and Rajesh Kannan Megalingam

***Abstract*—Machine learning (ML) is a significant subset of Artificial Intelligence (AI) that plays a key role in medical diagnosis. The advantage of AI is they can automatically learn, extract and translate the features from data sets such as images, text or video, without introducing traditional hand-coded code or rules. This paper focuses on recognizing and classifying lung diseases by ML algorithms. It includes 400 lung disease images (i.e. CT scan images) including bronchitis, emphysema, pleural effusion, cancer, and normal. The input image is analyzed, categorized and classified using ML algorithms such as the MLP, KNN and SVM classifier. After feature extraction, the output is segmented and compares the classifier's accuracy. When a CT scan image was given to a classifier as an input, it contains irrelevant information. For the selection of the most relevant features (i.e. for extracting characteristics) here Gray Level Co-occurrence Matrix (GLCM) is used. For MLP, this classifier acquires 98% accuracy, for SVM accuracy is 70.45% and for KNN accuracy is 99.2%. These classifiers will help the doctors to prescribe the most effective treatment for a patient.**

***Index Terms*—Machine learning (ML), Artificial Intelligence (AI), Gray-Level Co-occurrence Matrix (GLCM), Multilayer perceptron (MLP), K-nearest neighbors (KNN), Support vector machine(SVM)**

I. Introduction

REDUCING the detection period of diseases and improving identification accuracy becomes the most important issue in creating a reliable and more efficient medical decision support systems (MDSS) to help the complicated decision process for diagnosis. A complex and fuzzy cognitive method is the medical diagnosis, soft computational methods such as ML algorithms like MLP, SVM, and KNN showed great promise in the design of MDSS for disease detection.

In medical diagnosis, computed tomography (CT) images are commonly used. Depending on their distinct gray scales, computed tomography images could be used to distinguish many body tissues. A medical diagnosis that can be performed using traditional X-rays provides multiple images within the body. The cross-sectional CT scan images provided a variety of body planes that can be generated in the 3D view. CT scans include high resolution pictures of lungs that can be viewed on a PC or printed on a film. Lungs are responsible for oxygen supply and carbon dioxide exhalation as well. Most individuals have smoking habit that leads to infection and biological disorders that cause pulmonary diseases.

This paper contains four disease types (i.e. bronchitis, emphy-sema, pleural effusion, and cancer) as well as a normal lung CT scan. The inflammation between the nose area and the lung tissue that surrounds the airways causes bronchitis. This causes pneumonia. Emphysema is a form of COPD (chronic obstructive lung disease) that causes damage to the lung air sacs when germs affect pleural space. Cigarette smoking triggered this. Pleural effusion is otherwise referred to as lung water. It is due to the accumulation of excess fluid between pleura layers. It will damage the inhalation and exhalation and reduce lung tissue growth. Lung cancer is uncontrollably caused by cell division in the lung and will affect breathing. When CT scan image itself is used as an input, we re-quire a large number of variables when handling data. The computing power and memory will be increased by large number of variables so extraction of features is used to reduce the information i.e. the number of variables used, for that in this paper the method used is GLCM (Gray Level Co-occurrence Matrix).The GLCM is a mathematical methodology for analysis of texture that provides the spatial ratio of pixels. The GLCM is the spatially dependent gray level matrix. This requires properties of texture and color. The properties of texture include contrast, correlation, energy and homogeneity. Mean, Standard Deviation, Entropy, and RMS are color properties. GLCM matrix includes these eight features. Classification was performed using MLP (Multilayer Perceptron), Support vector machine (SVM) and KNN (k-nearest neighbors), which are capable of handling complex data. After classification the performance of two classifiers are compared.

Binila Mariyam Boban is with Department Of Electronics and Communica-tion Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India. (E-mail: binilaboban@gmail.com).

Rajesh Kannan Megalingam is with Department Of Electronics and Com-munication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India. (E-mail: rajeshkannan@ieee.org).

978-1-7281-4988-2/20/$31.00 ©2020 IEEE

The proposed method should be carried out in four phases, i.e. Pre-processing is done in the first phase with the use of median filters and morphological smoothening. The characteristics are derived from the pre-processed picture using GLCM (Gray-Level Co- occurrence Matrix) methodology. The second last phase of detection and separation of lung ailments is accomplished using the MLP (Multilayer Perceptron), Sup-port Vector Mechanism (SVM) and KNN (k-nearest neighbor) classifications. The final phase performance evaluation of the classifier. For implementing these algorithms software's such as MATLAB or python, can be used.

The rest of the paper is organized as below. Section II and Section III describes about the literature review and methodology of the work. The simulation results are discussed in Section IV. At last, Section V concludes the paper with conclusion of the work.

II. Literature Review

This paper [1] discusses the potential for medical diagnosis and prediction of osteoporosis by risk factor in the use of an artificial neural network (ANN). Artificial neural network (ANN) is developed in tandem with Probabilistic Neural Networks (PNN) based on MLPs with back propagation.In this paper [2], authors proposed a neural network focused on MLP backpropagation to predict heart disease.Here vari-ous multi-layer perceptron training functions are compared and the best training function is chosen for training. MLP with TRAINBR training algorithm gives 96.3% accuracy in heart disease prediction. In this paper [3], authors developed an artificial neural network with histogram based genomic gradient characteristics for predicting lung cancer. Together with histogram based gradient genomic features, this ANN network provides 95.90% percent accuracy and 0.0159 mean square error. In this paper [4], two forms of ANNs used to identify and diagnose Parkinson's disease were suggested by researchers. One is MLP (MultiLayer Perceptron) and the other is RBF (Radial Base Function). MLP is the best classifier with 93.22% percent accuracy based on the accuracy compar-ison. RBF classifier offers just 86.44% accuracy in classifying the same set of data. This can assist neurologists in their medical diagnosis. In this paper [5], researchers suggested a diagnostic method to assist doctors in the diagnosis of heart disease based on patient clinical conditions after translating it into numerical representation. Two classifiers were proposed: Multi-Layer Perceptron Neural Network (MLP) and Support machine vector (SVM). Here they considered the classification of two heart diseases and used the collected database to evaluate the performance of this classifier.

In this paper [6], authors proposed a Convolution Neural Network (CNN) for the classification of malignant or benign tumors in the lung. By using CNN as a classifier, the accuracy reached 96%, which is better than the traditional neural classifier accuracy. In this paper [7], authors focus on early lung cancer detection. This paper proposes a computational method, i.e. particle swarm optimization (PSO) with neural network. In this paper [8], authors concentrated on detection of lung cancer at early stage. For the identification, a non-parametric process, like genetic K-Nearest Neighbor (GKNN) algorithm is suggested. In this process K (50-100) are chosen for each iteration using genetic algorithms and performance tests in the exact range of 90%. Researchers introduced in this paper[9] a K-immediate neighbors classification to define and distinguish cancer into harmless or malignant pictures. In the classification of benign or malignant tumor, the overall classification acquired by the classifier is 97%. The learning time in this K nearest neighbor algorithm is 3 seconds and the nearest neighbor distance is 0.20889. Authors applied a SVM based description of diagnosis of lung cancer in this paper [10]. CLAHE Equalization technique improved the contrast of the CT scan graphic. After that, the method of walk segmentation was implemented. The writers in this paper [11] used median filters to minimize noise without affecting performance in pre-processing. After that feature, extraction has been done and the feature extracted has been selected by PSO (particle swarm optimization) algorithm method and lung disease classification has been done. In this paper [12] author proposed, a KNN based classifier together with the genetic algorithm for heart disease detection. Here values have been taken and recorded for different k values.

In this paper [13], features were derived from the GLCM method and the neural network back propagation algorithm was used for the classification of images. In the training stage, the classifier reaches 95% precision and 81.25% exactness in the evaluation level. This paper [14] explores the use of a neural network to diagnose various patterns of rubella, German measles and chickenpox signs, based on the pores and skin symptoms. The ANN will examine the signs and provides better predictions and credibility than a human doctor. Thus, patients can be monitored entirely based on the signs found for pores and skin problems. In this paper [15], a novel approach is suggested within order to achieve better rates of classification by integrating the predictive T-test and absolute ranking. Appropriate classification methods are also explored using linear SVM, proximal SVM and Newton SVM. Also presented is a descriptive study on the various extraction techniques. In this paper [16], they describe the image processing technique like fractal image compression and its properties and a method to improve the performance.

III. Methodology

A. Multilayer perceptron (MLP)

Neuron is a basic building block of a neural network (MLP) which is also known as artificial neurons that takes certain number of weighted input signals and bias and produce weighted output based on activation function as shown in Fig. 1. When a network has 5 inputs it will have 5 weights that can be adjusted in training section.

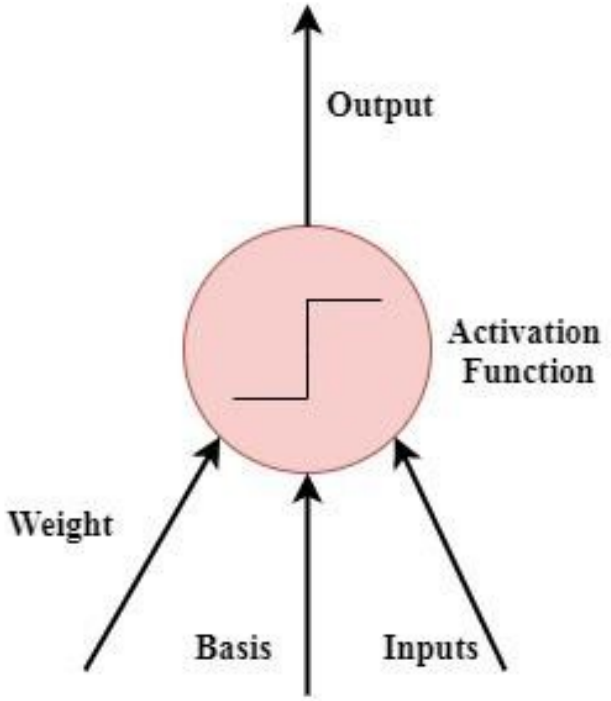

Fig. 1. A single neuron

The main purpose of training is changing the model's dimensions, or weights and biases to eliminate errors

$$\text{output} = £(\text{weight} \times \text{input}) + \text{bias} \quad (1)$$

Bias - It is an extra input to the neuron and its value is always one and it has its own weight when all the inputs are zero an activation in the neuron takes place.

It also allows the neuron to shift the decision boundary left or right i.e. it allow to decide the decision boundary.

Activation Function -It implies that the neural network is not linear. The triggering functions include tanH,sigmoid, RelU and SelU. Strength of output signal and the thresh-old at which neuron is activated was decided by the activation function.It is a mapping between the weighted input and output.

Input layer - The first layer which takes input values and no operations are apply on the input signals. Here there is no weight values and basis applied.

Hidden layer – Each layer collects knowledge from the data from the neurons (input line) and moves to the next step. When the no. of layer increasing it capture all the minute details. One hidden layer means one set of neurons that arrange vertically.It's called fully joined if all neurons in the secret layer are bound to each neuron in the next layer.

Output layer – The last layer to accept data from the most recent secret layer and output within the desired range.

Weights – It represents the strength of the connection between the nodes. They are initialized with some random values initially between 0–0.5.

Feed propagation – Forward movement of data from the input layer, where no process is done to next layer where process like multiplication, addition and activation process are take place and this repeated in coming layers until it reaches output layer. From output layer we get a predetermined value.

Back propagation - After forward propagation we get a predicted value at output side in order to find the error we compare the actual output value with these predicted one (loss function is usually used). Their difference is error, In order to reduce error we calculate the derivative of the error with respect to each and every weight in the network. Calculating the derivative gradients start from the last layer weights and move backwards until we reaches initial layer. Then subtract these gradient value from current weights and initialize the result as new weight. Then the input is given to check whether the error reduced. It will continue until the error reaches minimum value

B. K-nearest neighbors (KNN)

Algorithm for K neighbors (KNN) uses the similarity function to estimate values for the new data points, implying that a score will also be allocated to the current data points depending on how exactly they fit the training points.

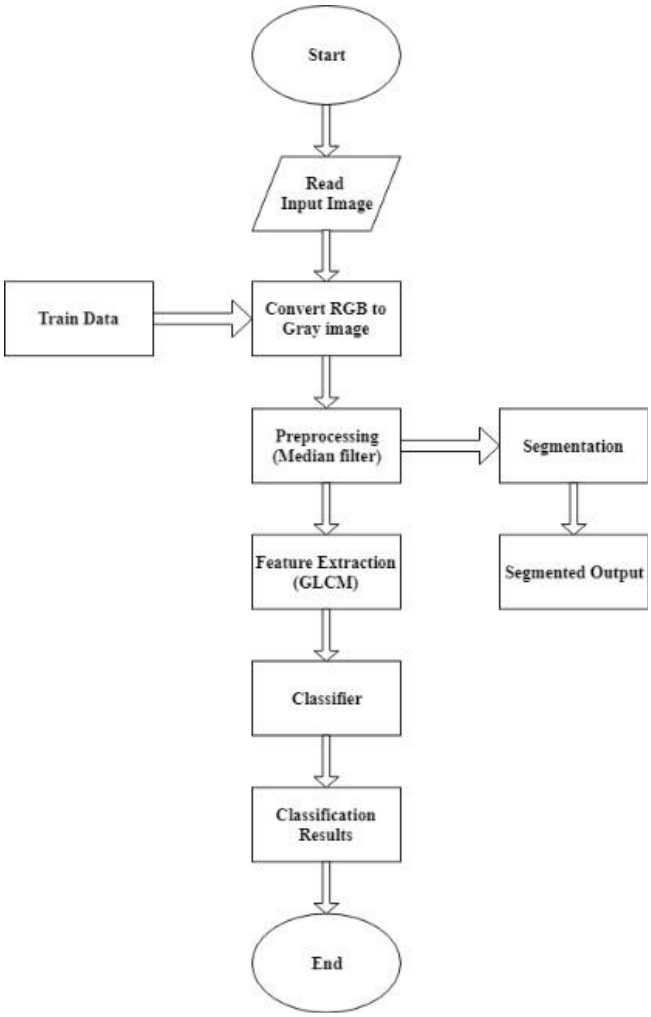

Fig. 2. Flow chart.

The following steps help us to understand its working:

Phase 1: For any algorithm, we need a data set. So during the first step of KNN, we will load the training and test data.

Phase 2: We will first select K (here k=3), i.e. the closest points of information.Difference between test data and each

training row is then calculated. The distance calculated in ascending order based on distance values is sorted from

Euclidean distance as distance metric.

Phase 3: Then get top k rows from the categorized list. The most common class is the real one.

C. Summer Vector Machine

Multi-class SVM attempts to allocate marking to instances of supporting vector machines that derive the mark from several elements in a finite range.The approach used here is to reduce the single multi-class problem to several binary classification problems via a one-to-all approach. The one-over-all approach is to create binary classifiers that differentiate one label from the rest.

From Fig. 2, first the input image (i.e. RGB image) is converted into grey format and applied to median filter to remove noises and for smoothening. Then the output image is now applied to GLCM so that certain parameters (Contrast, Correlation, Energy, Homogeneity, Mean, Standard deviation, Entropy, RMS) can be extracted. Then segmentation is done here we identifying the affected area. Finally images are passed to the classifier, where the classification takes place. After applying the classification techniques on the same dataset, it is found that KNN classifier is having higher accuracy than simple MLP and SVM classifier.

IV. Simulation Results

When a CT scan image is given as input first it is converting to gray image i.e. removing hue and saturation then given to median filter for removing noise and smoothing without affecting the sharpness of image. Fig. 3 shows when a CT scans image given as input: a-bronchitis, b-emphysema, c-pleural effusion, d-normal, e-lung cancer. first column corresponds to original image, then gray image and finally filtered image.

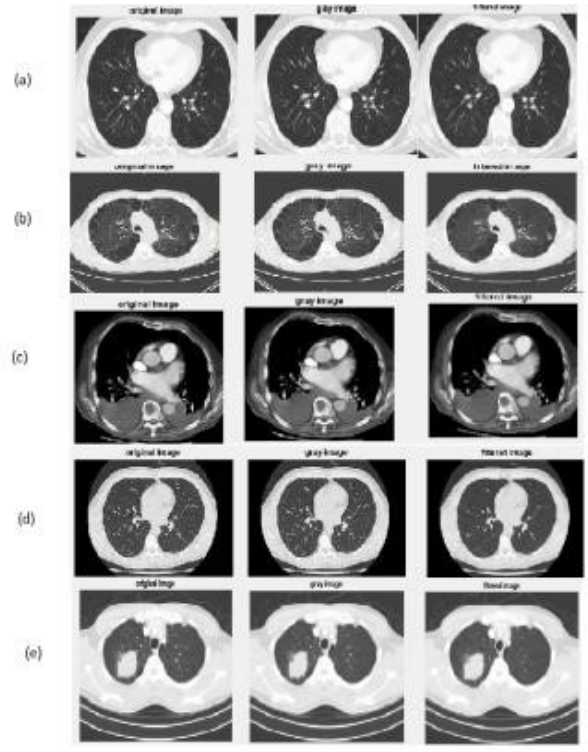

Fig. 3. CT scan image.

```
featuremat =
    0.1114    0.9904    0.2634    0.9522  152.7395   76.1374    6.4871   15.9684
```

Fig. 4. Feature matrix.

Fig. 4. is the features extracted using GLCM function. Here we take only eight features and this is given to classifier for identifying the disease and for correctly classifying it.

Fig. 5. Result.

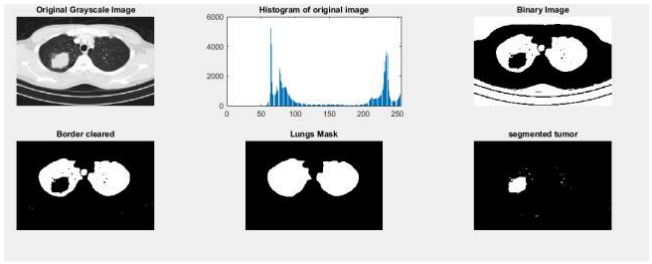

Fig. 6. Segmentation of tumor.

A message box is used to display the result if this is belong to bronchitis then ”predicted disease is bronchitis” such a message is displayed on the screen as Fig. 5. Similarly for other four classes here only one result is shown in this paper.

A. Segmentation

Fig. 6 shows the segmentation of tumor part by using a mask. Here when the predicted disease is cancer then segmentation take place and get the tumor out. Histogram graph shows the number of pixels in different intensity values. Method used for segmentation is binarization along with thresholding

B. Percentage of classification in MLP

Confusion matrix shown in Fig. 7 and confusion chart is used to evaluate the performance of classifiers. From chart or matrix we can find the percentage of correct and incorrect classifications in each class as shown in Fig. 8.

The accuracy get from MLP is 98.7%.Here the percentage is displayed on the command window by using this formula.

$$\text{Accuracy} = (t_p + t_n) / (t_p + t_n + f_p + f_n) \quad (2)$$

tp- True positive (The actual class is correctly predicted).

tn- True negative (The actual class is wrongly predicted).

fp- False positive (The wrong class is correctly predicted).

fn- False negative (The wrong class is wrongly predicted).

Fig. 7. MLP Confusion matrix

```
Percentage Correct Classification   : 98.715203%
Percentage Incorrect Classification : 1.284797%
```

Fig. 8. MLP classification percentage

```
ypred =

    0.0000
    0.0000
    1.0000
    0.0000
    0.0000
```

Fig. 9. Probability matrix

This is MLP output got from the output layer. The five values corresponds to the five classes i.e. bronchitis, emphysema, pleural effusion, cancer, and normal respectively. The output is get as probability as shown in Fig. 9. Here in this figure third row value is high i.e. the Ct scan image is belongs to cancer class.

Fig. 10. Accuracy of KNN

C. K nearest neighbors (KNN)

KNN is less complex than MLP because there is no activation function, weights and bias and it is mainly used for supervised machine learning as shown in Fig. 10. This paper is based on supervised ML. From the confusion matrix it see that incorrect classification made by KNN compared to MLP is less so accuracy will increase up to 99.6%

```
confusion matrix
    94     8     0     0    10
    30    55     0     8     4
     0     0    86     4     0
     0     0     0    70     0
    52     6     0    16    24

SVM (1-against-1):
accuracy = 70.45%
```

Fig. 11. Accuracy of SVM

D. Support Vector Machine (SVM)

Fig. 11 shows the SVM classifier accuracy when same set of dataset used. The classification accuracy is only 70.45% less than both MLP and KNN classifiers

V. Conclusion

In this project we are giving CT scan image of lungs in jpg format as an input to the program. After pre-processing i.e. converting to gray image and remove the noise then it is fed for feature extraction using GLCM. Here we get a matrix that contains only needed features; it helps to save time and memory i.e. to reduce the variables. After that matrix is given to successfully trained classifiers and compare the performances. Segmentation is done by using masking and thresholding. Comparing the performances shows that KNN (K nearest neighbor) is more accurate than MLP (Multi layer preceptron) and Support vector machine (SVM) classifiers. .

Acknowledgment

I am thankful for the wonderful opportunity offered me by Amrita University and Humanitarian Lab to transform my thoughts on a specific project. I am thankful for the space and

infrastructures required for completing this project to Dr Rajesh Kannan Megalingam. I appreciate everyone who helped me get this project done in good time.

REFERENCES

[1] Dimitrios H. Mantzaris, George C. Anastassopoulos , Dimitrios K. Lymberopoulos, "Medicaldisease prediction using artificial neural networks", 2008 8th IEEE International Conference on BioInformatics and BioEngineering,Oct. 2008, DOI: 10.1109/BIBE.2008.4696782.

[2] Durairaj M, Revathi V, "Prediction of heart disease using back propagation mlpalgorithm",International Journal of Scientific Technology Research volume4, issue 08, August 2015.

[3] Emmanuel Adetiba, Oludayo O. Olugbara," "Lung cancer prediction using neural networkensemble with histogram of oriented gradient genomic features",TheScientific World Journal, Volume 2015, Article ID 786013,http://dx.doi.org/10.1155/2015/786013.

[4] Farhad Soleimanian Gharehchopogh ,Peyman Mohammadi, "A case study of parkinson'sdisease diagnosis using artificial neural networks",International Journal ofComputer Applications (0975 – 8887), vol. 73– No.19, July 2013.

[5] Tabreer T. Hasan,Manal H. Jasim, Ivan A. Hashim, "Heart disease diagnosis systembased on multi-layer perceptron neural network and support vector machine",International Journal of Current Engineering and Technology, vol. 7, oct 2017.

[6] S. Sasikala, M. Bharathi, B. R. Sowmiya,"Lung cancer detection and classificationusing deep cnn",International Journal of Innovative Technology andExploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-2S,December, 2018.

[7] Dr. S. Senthil,B. Ayshwarya,"Lung cancer prediction using feed forward back propagation neural networks with optimal features",International Journal ofApplied Engineering Research ISSN 0973-4562, vol. 13, Number 1 pp.318-325, 2018.

[8] P . Bhuvaneswari , Dr. A. Brintha Therese ,"Detection of cancer in lung with k-nnclassification using genetic algorithm",2nd International Conference onNanomaterials and Technologies, 2014.

[9] P. Thamilselvan, Dr. J. G. R. Sathiaseelan,"An enhanced k nearest neighbor method todetecting and classifying mri lung cancer images for large amount data",International Journal of Applied Engineering Research ISSN 0973-4562, vol.11, Number 6 pp 4223-4229, 2016.

[10] R Sathishkumar,Kalaiarasan K,Prabakaran A, Aravind M,"Detection of lung cancer usingsvm classifier and knn algorithm",International Journal of ScientificResearch and Review, Volume 8, Issue 3, 2019.

[11] Tejinder Kaur,Neelakshi Gupta,"A new algorithm for classification of lung diseases",International Journal of Advances in Electronics and Computer Science,ISSN: 2393-2835, Volume-2, Issue-9, Sept.-2015.

[12] M.Akhil jabbar, B.L Deekshatulua Priti Chandra ,"Classification of heart disease using k- nearestneighbor and genetic algorithm",International Conference on ComputationalIntelligence: Modeling Techniques and Applications (CIMTA), 2013.

[13] Kusworo Adi, Catur Edi Widodo, Aris Puji Widodo, Rahmat Gernowo, Adi Pamungkas, Rizky Ayomi Syifa ,"Detectionlung cancer using gray level co-occurrence matrix (glcm) and backpropagation neural network classification",JOURNAL OF EngineeringScience and Technology Review, March 2018.

[14] Monisha M; Suresh A; Rashmi M R, "Artificial Intelligence Based Skin Classification Using GMM", Journal of Medical Systems, vol. 43, no. 1, p. 3, 2018.

[15] Arunkumar Chinnaswamy , Ramakrishnan S, "Two Step Feature Extraction Method for Microarray Cancer Data using Support Vector Machines", International Journal of Computer Applications, vol. 85, no. 8, pp. 34-42, 2014.

[16] Loganathan D, Amudha J; Mehata K.M,"Classification and feature vector techniques to improve fractal image coding",IEEE Region 10 Annual International Conference, Proceedings/TENCON, Volume 4, Bangalore, p.1503-1507 (2003).