Ensemble Based Medical Intelligent System for Discovering Cancerous Lung Nodules

Chandrakala P

{pckchandrakala@gmail.com}

Department of Computer Science, Goverment Arts College for Women, Nilakottai.

Abstract. Lung cancer is the one of the risky diseases which has the highest mortality rate. Hence earliest detection of cancer can upturn thechance of human survival rate. This paper aims to discover the cancerous lung nodules inComputed Tomography (CT) images through Ensemble Based Medical Intelligent System (EBMIS). The proposed system firstly embedsimage preprocessing mechanisms for enlightening the quality of CT images to improve the exactness of result. Then NoduleDetection System (NDS) is employed to segment the nodule candidates from lung region and set of intensity, geometric and texture features are extracted from the segmented nodules. Finally, distinguish the cancerous nodule from non-cancerous nodule is done by ensemble of classifiers encompassing Decision Trees, Naïve Bayes, K-NN and SVM Algorithms. The improvement in classificationin terms of reduced error rate and improved accuracy is obtained by performing clustering ensemble prior to classification which includes K-Means, FCM, DBSCAN and Hierarchical Clustering algorithms. Open Lung Image dataset from Lung Image Database Consortium (LIDC) is experimented for quantifying the efficacy of proposed system. The proposed approach outperforms existing art of approaches by achieving stability even in increasing the dataset size. The Experimental outcome is confirmedthat the proposed system is a promising tool in diagnosing cancerous lung nodules with accuracy of 99.63%.

Keywords: clustering, classification, ensemble.

1 Introduction

Lung Cancer is the one of top deadliest disease which is proved by the North American Association of Central Cancer Registries report [1] stated that 13% of women and 14% of men have been affected by lung cancer among which 154050 cases leading to death. Cancer patients will have the chance of longer survival if the disease is detected at earlier stage. The USAreport detailed that 17.4% of cancer affected people survive up to 5 years.

Mostly, affected people will have the following symptoms such as chest pain, memoryloss, weight loss, blood clots, bleeding, breath shortness etc. Usages of tobaccos or smokinghabits are the foremost reason behind the lung cancer whereas few people have been suffered because

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343288

of genetic factors. The abandonedgrowth of cells in lungs and nearby tissues called lung nodules must be examined for itsseverity. There are two sorts of lung nodulessuch as cancerous and non-cancerous. Theseverity of cancerous nodule is ranging from stage I to stage IV. At stage I, cancer is confinedto the lung. At stage II and III, cancer is confined to the chest. At stage IV, cancer has spread to other parts of the body [3].

The main intention of this work is automatic detection of lung cancer at an earlier stage in CT images with high precision. ThisPaper is structured as follows: Section II describes the related research work, Ensemble Based Medical Intelligent System (EBMIS)Architecture is explained in Section III, and Ensemble Culture is described in Section IV.Finally Experimental Results and Conclusion with Future Scope have been explained in Section V and Section VI respectively.

2 Related Works

CT image based medical diagnosis helpsto examine the lung cells carefully and detect the eccentricities in cells to predict the lung cancer. Image based diagnosis normally accomplished through the sequences of classic steps [4], [5]: Image Preprocessing, Region Segmentation, Feature Extraction and Classification. Onceimage is captured, there may be a chance ofhaving uncertainties which are in the form of shadow over the image, inexact gray levels and blurry images unable to separate the image from its background. These uncertainties must be removed to avoid incorrect diagnosis by enhancing the contrast of the images. Prior to this, noises in the image are unavoidable due to some environmental issues which corrupts the pixel of image. Normally, noises in CT images are of four types like Gaussian Noise, Salt and Pepper Noise (Impulse Noise), Poisson Noiseand Speckle Noise [6]. So, outliers and noises need to be detached for simplifying the further process through noise removal mechanism. Hence image preprocessing is a mixture of two methods such as noise removal and contrast enhancement.

Contrast enhancement techniques minimize the intensification of noise which are generally classified as spatial domain-based techniques such as Contrast Stretching and Histogram Equalization; Frequency domain-based techniques such as Discrete Fourier Transform and Discrete Wavelet Transform, and Spatial-Frequency domain-based techniques such as Homomorphic Filtering [8]. In Spatial domain, pixel intensity can be modified directly which is applicable when overall contrast of the image needs to be improved whereas Frequency domain firstly convert the image into frequency domain then applying filtering techniques whichis applicable when some specific information such as edge or other specific parts need to be improved. In [9], they conducted the survey among frequency domain techniques like Butterworth filter, Gaussian filter, Gabor filter, Fast Fourier transform and Discrete wavelet transform among which Gabor filter has been confirmed as a appropriate technique to extract more meaningful feature from CT images in lung cancer diagnosis.

Segmentation is of two kinds such as Layer based Segmentation and Block based Segmentation

[11]. Region based segmentation is a block-based technique where segmentation is performed based on features of the image.Here all pixels with same characteristics in termsof intensity levels, color components, texturesand edges are grouped into one region like clustering. Thorough Supervised or Unsupervised Learning methods, region-based segmentation is possible. Fuzzy C-MeansClustering algorithm is widely used soft partitioning algorithm in which the objects with similar characteristics are grouped into one cluster [12] and also every object has a membership degree to each cluster. As it takes Euclidean distance as a similarity measurement, it is not appropriate for non-Euclidean formationof images such as CT or MRI Images. To tackle this issue, Kannan et al. [13], proposed robust Kernelized Fuzzy C-Means (KFCM) method where it is suspected that it is failed to get improved precision from the segmented region. Then from the segmented regions, features are extracted to facilitate the classification steps further. Mostly, Features are fall under three categories namely Intensity Feature, Texture Feature and Geometric Feature [14]. Here, geometric and texture-based feature provides information about shape and color of the image respectively and Texture based feature plays vital role in correct nodule detection in lung cancer diagnosis.

Finally, Machine learning algorithms especially classification algorithms are mostlyemployed with the extracted features as input to facilitate the medical diagnosis. Then there are several authors revealed their opinions about automatic lung cancer detection helps to acquire an idea to develop smart cancer prediction system. Ayushi Shukla et al. [15] and AnjaliKulkarni et al. [16] proposed SVM based Lung cancer detection in CT images and they achievedthe accuracy up to 90%. In [17], Kuruvilla and Gunavathi developed computer aided diagnosis based on Artificial Intelligence mechanism and 93.3% is a yielded accuracy. Song QZ et al. [18]have conducted survey among various deep learning approaches namely CNN, DNN and SAE in distinguishing cancerous nodules from non-cancerous nodules. Their experimentalresults said that CNN network achieved greatest performance with an accuracy of 84.15%. Shen et al. [19] utilized multilayer CNN and conducted the experiment on LIDC dataset and evidenced the accuracy as 86.84%. Even though numerous approaches have been developed, still misclassification is the crucial issue.

Ensemble learning is a recently opened branch in the field of machine learning [20], [21]which provides the desirable output to improve the accuracy of prediction through aggregating the decisions of multiple base classifiers. Thecore idea behind the ensemble is aggregatingseveral individual opinions will be better thanthe opinion of single one. It is a two-phase process. In first phase, different classifiers (base classifiers) or single classifier with different parameters are employed to form a creation of ensemble whereas in second phase, output fusion is performed to integrate the output ofbase classifiers [22]. There are two ways to fuse the outputs such as weighting methods and meta-learning methods. The former one is applicable when performance of base classifiers is comparable whereas the later one is applicablewhere base classifiers have differentperformances on different subspaces. Bagging, Boosting and Random Forest are the examplesof weighting methods and Stacking is theexample of meta-learning method. In Bagging,[24] voting is the simplest and effective method in which final prediction

is obtained by taking the base classifier output which has the majority of votes. In Ada Boosting, the main focus is on misclassified instances. Initially same weight is assigned to all instances. Then after every iteration, weights of misclassified instances are incremented and weights of correctly classified instances are decremented. In addition, with this, weight is assigned to all base classifiers basedon their performance. Finally, the average of weight is taken into account for creating the final prediction. In Random Forest, multiple decision trees are used as base classifiers. Here also weight is assigned based on theirperformance. Every iteration, trees with weakperformance are replaced by new trees to obtain final result. Stacking is mostly used meta-learning approach which involves learning from learners. Here, firstly, datasets are divided into two disjoint subsets. Then with many base classifiers, first subset is used for training phase and second subset is used for testing phase. At last, output of testing phase is fed as an input to yield final output. According to [23], the above-mentioned approaches involved in combining many weak learners into stronger one. It is hard to say that the certain ensemble approach is working well than others. Thus, selection of appropriate ensemble method is depending on the domain.

The perception behind the study is that many authors have involved themselves indeveloping the medical intelligent system for lung cancer diagnosis and compete with eachother to prove the high precision in classification. So, achieving greater accuracy is a major aspect in this domain becausemisclassification causes wrong diagnosis leadsto patient's death. The major contribution of thispaper is to discover the cancerous lung nodules in CT images at earlier stage through EBMIS. Inorder to reduce the misclassification rate, clustering ensemble is performed prior to classification for the first time in the lung cancermedical diagnosis [30]. In addition, with the greater accuracy, EBMIS attained reduced time complexity and stability even in increasing the dataset size.

3 EBIMS Archetype

The design of Proposed System is depictedin Fig. 1. It encompasses 3 stages namely Image Preprocessing Mechanism, Nodule Candidate Detection System and Ensemble Culture.

3.1 Image Processing Mechanism

Image Processing Mechanism is a two step process mainly focused on improving the visual appearance of the image to facilitate further steps effecvtively.

a. Noise Elimination Udin Median Filter

It is an denoising mechanism to diminish the amount of noise which is normally take place in the images during acquisition or transmission or compression phase. Salt and Pepper noise is mostly occurred noise in CT images which degarde the quality of image and loss of significant information leads to wrong medical diagnosis. Hence, noise elemination is considered as a prime factor in medical diagnosis. According to[7], Median filter is an excellent non liner filter at diminishing the Salt and Pepper noise with preservation of edgesd. It involves scanning of whole image using undersized matrix and median of all the values of adjoining pixels given in Eq.1.

$$f\hat{}(x,y)=median\ (s,t)\in Sxy\ \{g(s,t)\} \qquad (1)$$

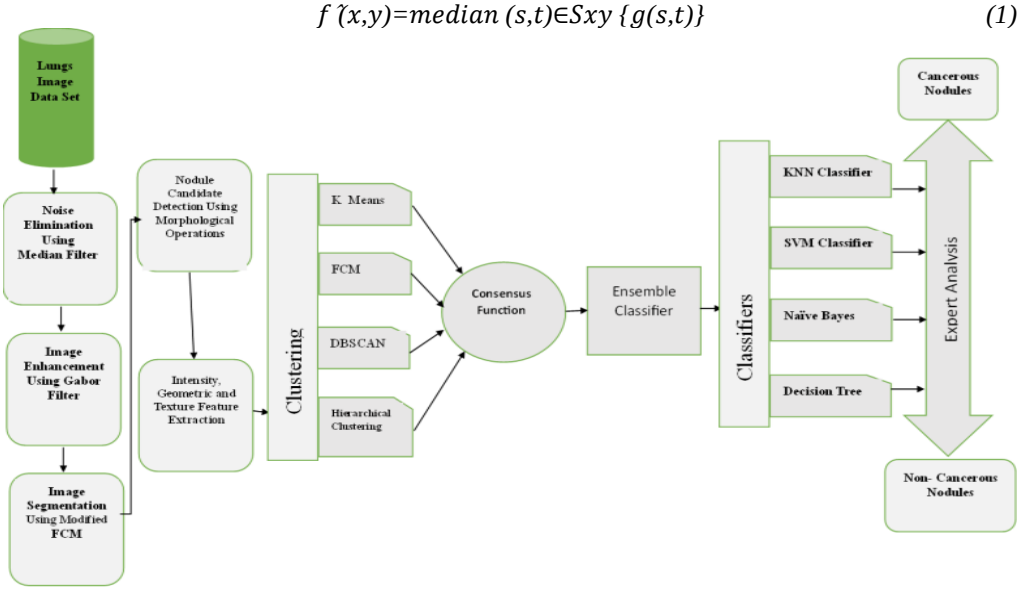

Fig. 1. EBMIS Archetype

b.Image Enhancement using Gabor Filter

It raising the contrast of images to get better clarity of minute matter of images. Gabor filter is a linear filter which considers the specific frequency content around the localized region. It yields optimum result because of excellent local and multi scale decomposition and its impulse response is found by multiplying Fourier Transform of Harmonic Function (H_f) and Gaussian Function (G_s) as shown in Eq. 2, Eq. 3 and Eq. 4. Where λ, σ, γ and Ψ signify wavelength, standard deviation, aspect ratio and offset respectively.

$$G_f(x,y)=exp[G_s(x,y)]*exp[i*H_f(x,y)] \qquad (2)$$

$$G_s(x,y)=-\frac{1}{2\sigma^2}x'^2+\gamma^2y'^2 \qquad (3)$$

$$H_f(x,y)=2\pi\frac{x\prime}{\lambda}+\Psi \qquad (4)$$

3.2 Nodule Candidate Detection System

It aims to dig out the lung region of image in orderto inspect the region of interest more accurately. It is begin with segmenting an image based on Modified FCM (M_FCM) followed

by detectingnodule candidates based on Morphological Operations.

a. Image Segmentation Using M_FCM isemployed in EBMIS to attain improvement in precision. It make use of kernel induced distanceinstead of euclidean distance and also distance and membership values of adjacent pixels are account into objective function [28] as shown in Algorithm 1.

b. Nodule Candidate Detection (NCD) Using Morphological Operations involvedetection of nodule candidates in segmented lung region. The name morphology implies processing the image based on shape. Firstly,morphological opening is applied to get rid of unnecessary objects occurred inside or outsideof segmented lung region. Then, morphological closing is applied further to augment the border of an image [3]. Finally, morphological filling is to fill holes of any detected nodules.

c. Intensity, Geometric and Texture Feature Extraction is performed on detectedlung nodule candidates to provide significant information for distinguishing cancerous nodules from non cancerous one. There are 15 features given in Table 1 such as Mean, Variance, SD, Skewness, Kurtosis, Roundness, Circularity, Compactness, Ellipticity, Eccentricity, Entropy, Energy, Contrast, Correlation and Homogeneity are taken into account for determining the intelligence ofEBMIS. Then the founded statistical features from each candidate are fed into the next stage for beginning ensemble culture.

4 Ensemble Culture

A modern medical discipline utilizes the culture of ensemble in disease prediction to improve its precision. The proposed EBMIS employs the framework of clustering ensemble prior to ensemble of classifier to improve the performance of classifier in terms of reduced error rate and improved speed in prediction.

4.1 Clustering Ensemble based on Iterative voting Mechanism (CE_IVM)

CE_IVM is a way of aggregating the Prtition of multiple base clustering algorithm $\pi = \{\varphi_1, \varphi_2, \varphi_3 \ldots \varphi_m\}$ into final partition (π^*). The first part of CE_IVM is generating the ensemble members using KMeans, FCM, DBSCAN and Hierarchical Clustering algorithms. There should be a high level of diversity among ensemble members can potentially improve the performance of ensemble [31]. The Second part is consensus function which combines the output of all ensemble members to provide a final clustering result.

a. K-Means Algorithm is a hard clustering algorithm where each data points belongs to exactly one cluster in which partitioning the given data points into K groups.

1. Initialize no. of cluster *(k)* and cluster centroids (c_t).
2. For every iterationi, Euclidean distance (μ_d) is found between all data points (x_i) and cluster centroids (c_t).
3. Then assign x_i to its closest c_t and calculate new c_t.

4. Reiterate the above steps until convergence criterion *(β)* is met.

$$F_{obj} = \sum_{i=1}^{n} \sum_{t=1}^{k} \mu_{it}^{m} \, ||x(i) - c(t)||^2 \qquad (5)$$

$$\mu_{it} = \frac{1}{\sum_{l=1}^{k} \frac{||x(i)-c(t)||}{||x(i)-c(l)||}^{\frac{2}{m-1}}} \qquad (6)$$

$$c(t) = \frac{\sum_{i=1}^{n} \mu_{it}^{m} * x(i)}{\sum_{i=1}^{n} \mu_{it}^{m}} \qquad (7)$$

b. FCM Algorithm is a soft clustering algorithm where each data point belongs to more than one cluster. This algorithm mainly focused to minimize the objective function (F_{obj}) given in Eq. 5. As shown in Eq. 6 and Eq. 7, membership matrix (μ_{it}) and cluster centre (c(t)) values are updated. Finally, data point x(i) is reassigned to its closest c(t).

c. DBSCAN Algorithm is density basedspatial clustering of applications with noise. It begins with initializing two parameters such as

1. *Eps (Є)* used to find out the neighbors. If distance between two data points is less than or equal to Є then those points are considered as neighbors.
2. *Minpoints* used to specify the minimum no. of data points to form a denseregion.

The conception of this algorithm is grouping the data points that are close to each other based on the values of *Є* and *Minpoints* and mark some of the data points as outliers that are in low-density region.

Table 1a. Intensity Features

S.No.	Metrics	Formula
1	*Mean (m_n)*	$\frac{\sum_{x=1}^{w} \sum_{y=1}^{h} p_x^y}{w * h}$
2	*Variance(v_r)*	$\frac{(\sum_{x=0}^{w-1} \sum_{y=0}^{h-1} p_x^y - m_n)^2}{w * h}$
3	*Standard Deviation (s_d)*	$\sqrt{v_r}$
4	*Skewness(s_k)*	$(\frac{1}{N * s_d^3} (\sum_{x,y} (p_x^y - m_n))^2)^{\frac{1}{3}}$
5	*Kurtosis (k_u)*	$(\frac{1}{N * s_d^4} (\sum_{x,y} (p_x^y - m_n))^2)^{\frac{1}{4}}$

Table 1b. Geometric Features

S.No.	Metrics	Formula

1	*Roundness (r_n)*	$\frac{4a}{\pi d^2}$
2	*Circularity (c_r)*	$\frac{4\pi a}{p^2}$
3	*Compactness(c_p)*	$\frac{1}{c_r}$
4	*Ellipticity (e_l)*	$2 * (r_n^{-1})$
5	*Eccentricity (e_c)*	$\frac{major\_axis}{minor\_axis}$

In the above tables, $p^{y}{}_{x}$ denotes the value of pixel at (x,y), w and h represent width and height, a,d and p denote area, diameter and perimeter respectively.

d. Hierarchical Clustering Algorithm is of two types such as agglomerative clustering and divisive clustering. The proposed system utilizes the agglomerative clustering as an ensemble member which performs following steps.

1. Since it is a bottom-up approach, it treats all given points as a separate cluster.
2. Then calculate distance from one cluster to all other clusters to form a similarity matrix (s).
3. Then clusters with closest distance are merged together using average linkage function.
4. Repeat the process till single hierarchical tree is obtained. Finally, cut point is determined to create the number of partitions.

4.2 Consensus Function aggregates the partitions of multiple clustering algorithms using voting mechanism that is final clusters select its members who acquired the maximum number of votes. But the data point who is not acquiring majority votes may be placed to the closest cluster [32]. It violates the principle of voting mechanism and also random placement of data will degrade the quality of obtained clusters. To address this issue, the proposed system uses Iterative Voting Mechanism (IVM) which is explained in Algorithm 2 where voting is made for every iteration to create sub clusters. Here every iteration begins with the data points which are having equal votes in the previous iteration. This process is repeated till D* becomes empty. To facilitate voting, similarity matrix is created based on jaccard index as given in Eq. 8. Subsequently, relabeling is performed based on the maximum jaccard similarity coefficient.

$$J(\varphi_l, \varphi_m) = \frac{||\varphi l \cap \varphi m||}{||\varphi l \cup \varphi m||} \qquad (8)$$

The IVM is ended by employing FCM algorithm on yielded sub clusters to obtain final clusters. Now, each data point is labeled based on the final cluster result. Consider x is a data point and after applying IVM algorithm, appropriate cluster found for x is cluster3. So, the cluster label for data point x is set to 3. Likewise, all given data points are cluster labeled. This cluster label is act as a special feature in addition with those 15 extracted features. The process of adding such a new feature is reflected in the classifier's decision in prediction which is experimentally proved and shown in section 5.

4.3 Prediction based on Ensemble of Classifiers (PEC) is the final stage which exploits the supervised learning approaches to train the EBMIS in lung cancer prediction based on the extracted features including cluster label. The cluster label is considered as a major influencing factor since it diminishes error rate while perform classification. The PEC starts its execution by creating committee of base learners using Decision Trees, Naïve Bayes, K-NN and SVM classifiers. Finally, all base learners' opinions are combined to confer a final decision on a detected lung nodule.

a. Support Vector Machine (SVM) is a nonlinear classifier which was discovered by Boser et al. [33]. It follows the principle of statistical learning theory. The major conception of this algorithm is to find the right hyper plane that has the greatest distance with the closest data points in order to diminish the misclassification rate. To support nonlinear inputs, the function of hyper plane fhp(x) is defined with mapping function (ker) as shown in Eq. 9.

$$f_{hp}(x) = \beta + \sum_{s=1}^{N} \alpha_i z_i k_{er}(x_i, x) \qquad (9)$$

Where β and α designates bias and leagrange multiplier respectively and N is the number of Support Vectors. The mapping function ker is performed using Gaussian Kernel based Radial Basis Function [28] and width of Gaussian Kernel is determined by σ which is defined in the Eq. 10.

$$k_{er}(x_i, x_j) = e^{\left(-\frac{1}{2\sigma^2}||x_i - x_j||\right)^2} \qquad (10)$$

b. Naïve Bayes Classifiers (NBC) are quite simple learning algorithm to handle very large data sets. It assumes that the given features (attributes) are independent of each other and also all of them are treated equally [35]. These two properties lead to make this algorithm to learn rapidly. It starts with converting given data into frequency table which measures the count of attributes (β) in each class labels. Subsequently, likelihood table is formed by finding the probabilities of each feature and each class labels. Then posterior probability for each class label is calculated using naïve Bayesian equation which is given in the Eq. 11. Finally, class label (α) with highest posterior probability is chosen as the final prediction.

$$p(\alpha/\beta) = \frac{p\left(\frac{\beta}{\alpha}\right) p(\alpha)|}{p(\beta)} \qquad (11)$$

c. Decision Trees (DT) are tree like structure where internal nodes are act as decision nodes to conduct test on attributes and leaf nodes contain outcome of the decision that is class labels [34]. It is most preferable algorithm since it mimics human behavior while making decision. The proposed system utilizes an ID3 (Iterative Dichotomiser) algorithm to construct a decision tree. The core concept is to discover the finest splitting criterion that yields minimum Entropy (μ) or maximum Information Gain (σ). It begins with calculating μ for entire dataset (D) using the Eq.12 to measure how much uncertainty presents in the given dataset. Then, for each attribute (β), μ and σ values are calculated. The information gain is measured using the Eq. 13 to determine how much uncertainty in given dataset was reduced after applying splitting criterion. Finally, the feature or attribute with maximum σ is elected as a best splitting criterion to split

$$\mu(D) = \sum_{\forall \alpha} -p(\alpha) \log_2 p(\alpha) \qquad (12)$$

$$\sigma(\beta, D) = \mu(D) - \sum_{\forall s} p(s)\mu(s) \qquad (13)$$

the given data set. Then, the above-mentioned steps are performed repeatedly until desire tree is obtained.

Where α and s designates class labels and elements in subset (D') respectively and p(α) denotes the proportion to the number of elements in α to the number of elements in D. A subset D' is formed by splitting D by best splitting attribute β.

e. K-Nearest Neighbour Algorithm (KNN) is an effective non parametric algorithm to train the large datasets. It starts its execution by selecting K number of neighbours. To decide the class label or category for new data point, Euclidean distance is calculated between the new data point and all data points in each category. Then K nearest neighbours is selected based on the calculated distance and the category which has the maximum number of selected neighbours is chosen as the class label for new data point.

f. Ada Boost Ensemble Learner (AEBL) is employed in proposed system to make a final prediction on detected lung nodule. It improves the overall performance of EBMIS because of its effective boosting process [36]. Basically, decision trees are the weak learners of Ada Boost. But, in PEC, 4 kinds of classifiers are treated as base or weak learners. Each of them runs in sequentially and each subsequent learner try to correct the mistakes of its predecessor. To achieve this, in every iteration, the object (xi) which is misclassified carries higher weight than the correctly classified samples. The performances of base learners are determined by measuring the error rate (ε) and significance (α) among which α plays major role in final prediction.

The effectiveness of EBMIS is evaluated by conducting an experiment over 1097 LIDC lung images in which 681 images comprise cancerous nodules. To facilitate the Detection of such candidates, EBMIS has employed Image Processing mechanism (IPM) on LIDC dataset. The entire IPM operations from scratch to detection of nodule candidates are carried out with the help of python libraries include Matplotlib and Open CV which is shown in Fig. 2. In Fig. 2(e), the nodule on lung image is spotted and it is marked in red color.The significance of IPM in EBMIS is shown in Fig. 3.

An EBMIS without IPM could able to attain 68.18% of accuracy in prediction because diagnosis is based on analyzing the entire lung image rather than ROI. The efficacy of EBMIS is demonstrated at the stage of PEC where the classifier commences its execution with the help of Scikit-Learn Library. Moreover, statistical metrics which are listed in Table 2 have been used to conclude the diagnosis ability of EBMIS.

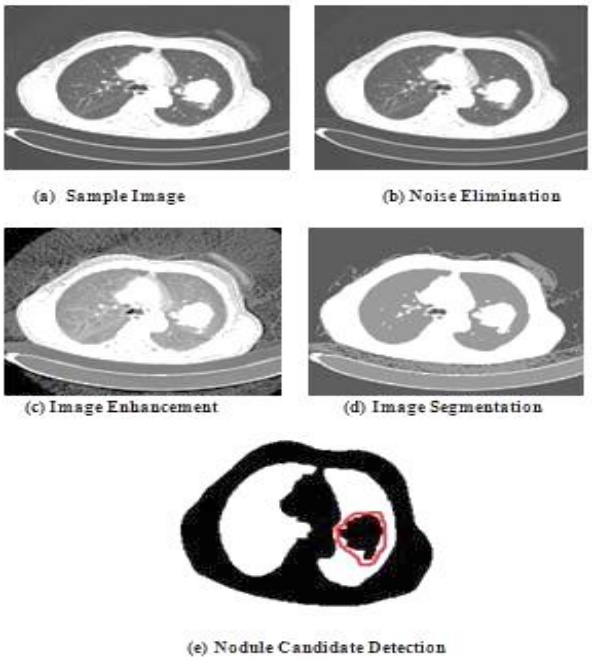

Fig. 2. IPM over Cancerous Lung ImageTable 2. Performance Metrics

The ability of EBMIS is evaluated through train- test split approach against all base classifiers namely SVM, NBC, DT and KNN and also with existing ensemble approaches such as DAELGNN and MSKFCM. The experiment is carried out repeatedly by dividing the samples in LIDC dataset into three splits such as 50%-50%, 60%-40% and 70%- 30% for training and testing. The output of proposed system on each split is given in Table 3.

Table 2. Performance Metrics

S.No.	Metrics	Formula	Description
1	Accuracy (Acc)	$\frac{t^+ + t^-}{N}$	t^+ - *true positives* t^- - *true negatives* f^+ - *false positives* f^- - *false negatives* *N – total no of samples*
2	Sensitivity (S_n) (or) Recall (R_e)	$\frac{t^+}{t^+ + f^-}$	
3	Specificity (S_p)	$\frac{t^-}{t^- + f^+}$	
4	Precision (P_r)	$\frac{t^+}{t^+ + f^+}$	
5	Fmeasure (F_m)	$2*(\frac{Pr * Re}{Pr + Re})$	

6	Gmean(G_m)	$\sqrt{S_n * S_p}$	

Table 3. Performance of EBMIS over 3different Train-Test Splits

EBMIS	Acc	Sn	Sp	Pr	Fm	Gm
70%-30%	0.995	0.996	0.993	0.996	0.996	0.994
60%-40%	0.996	0.997	0.995	0.997	0.997	0.996
50%-50%	0.998	0.999	0.998	0.999	0.999	0.998

The comparison results are populated in Table 4 by taking the average output of different splits which says that the proposed EBMIS is outperforming (99.63% accuracy) the rest of the approaches. The accuracy of former approach is 97.26% and later approach has made an improvement in accuracy to reach up to 99.63%.

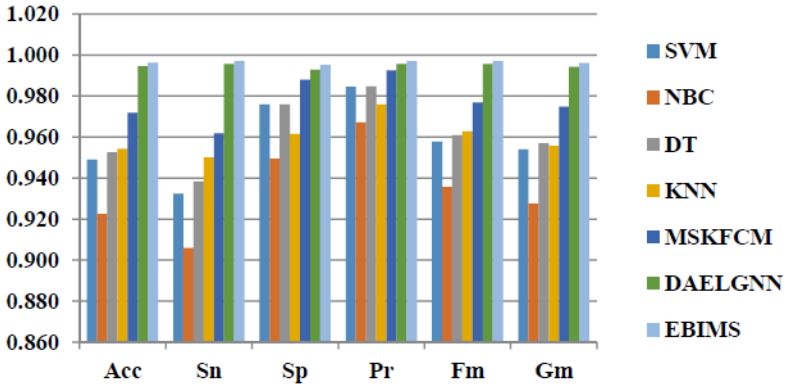

Fig. 3. Evaluation on Efficacy of EBMIS

The process of CE_IVM gives the cluster label for each input image which is considered as a special feature in addition with those 15 extracted features. Such phenomenon enhances the diagnosis ability of proposed system with 0.36% of error rate. A confusion matrix is a tabular representation which depicts the performance of classifiers in terms of actual classification and predicted classification done by EBMIS_CE_IVM_PEC and EBMIS_PEC as shown in Table 5a and 5b respectively.

Table 4. Evaluation on Efficacy of EBMIS

Methods	Acc	Sn	Sp	Pr	Fm	Gm
SVM	94.895	0.932	0.976	0.984	0.958	0.954
NBC	92.252	0.906	0.950	0.967	0.936	0.928
DT	95.260	0.938	0.976	0.985	0.961	0.957
KNN	95.442	0.950	0.962	0.976	0.963	0.956
MSKFCM	97.174	0.962	0.988	0.992	0.977	0.975
DAELGNN	99.453	0.996	0.993	0.996	0.996	0.994
EBMIS	**99.635**	**0.997**	**0.995**	**0.997**	**0.997**	**0.996**

Table 5a. EBMIS_CE_IVM_PEC

EBMIS_CE_IVM_PEC		Predicted	
		c+	c-
Actual	c+	61.90 %	0.18 %
	c-	0.18 %	37.74 %

Table 5b. EBMIS_ PEC

EBMIS_PEC		Predicted	
		c+	c-
Actual	c+	59.98 %	2.10 %
	c-	0.64 %	37.28 %

In addition, with accuracy, execution time of proposed system is analyzed which is shown in Fig.5. Since it involves execution of both clustering and classification, comparison is made against EBMIS_CE_IVM_PEC and EBMIS_ PEC and also execution time is measured for different train-test splits. An integration of CE_IVM with PEC is around 1.8 times slower than EBMIS_PEC. However, clustering and classification combo has reduced error rate of 0.36% in prediction which in turn reduces the mortality rate. If the clustering is not combined with classification, EBMIS has misclassified a greater number of samples leads to high mortality rate (Up to 2.7%). As more effort in terms of time and money is needed in medical diagnosis to save humans life, the time complexity induced by proposed system is substantial.

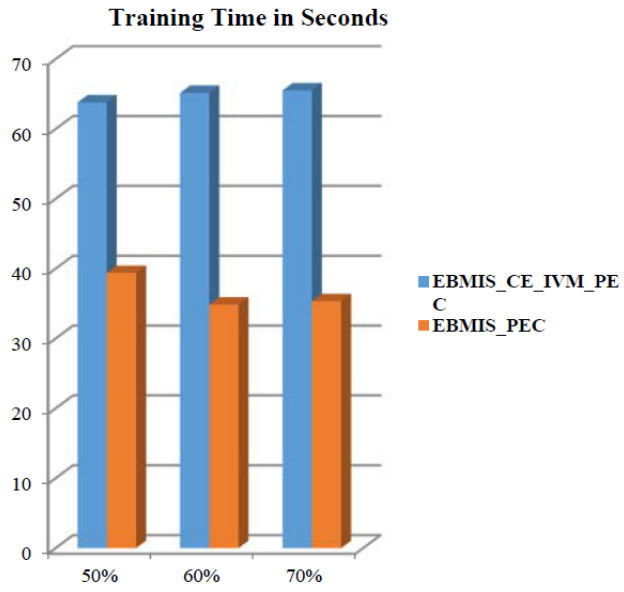

Fig. 5. Training Time of EBMIS for three different splits

In Fig. 5, it shows the training time taken by proposed system for three different splits to identify the scalability nature of EBMIS. The proposed system has taken 63.81s, 65.17s and 65.53s for the sample size 50%, 60% and 70% respectively. From this, it has been confirmed that considerable amount of time is taken even increase in the data set size.

6 Conclusion

This paper introduced an ensemble based medical intelligent system for discovering cancerous lung nodules. It has three vital stages namely image preprocessing mechanisms, nodule candidate detection system and ensemble culture. At stage 1, appearance of the raw image has been improved with the help of Median and Gabor filters. In the next stage, segmentation has been performed through MFCM and nodule candidates are detected based on morphological operations. Subsequently, geometric, intensity and texture features are been extracted from the detected candidates. At last stage, ensemble culture is employed for making decision on detected nodules. An ensemble culture has combined both clustering and classification process with IVM and ABEL process to confer the strong decision.

References

[1] P. Mohamed Shakeel, Amr Tolba, Zafer Al- Makhadmeh, Mustafa Musa Jaber (2018), "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks", Neural Computing and Applications, https://doi.org/10.1007/s00521-018-03972-2

[2] May Phu Paing , Kazuhiko Hamamoto, Supan Tungjitkusolmun and Chuchart Pintavirooj (2020), "Elimination of Noise in CT Images of Lung Cancer using Image Preprocessing Filtering Techniques", International Journal of Advanced Science andTechnology, Vol. 29, No. 4s, pp. 1823-1832.

[3] Khin Mya Mya Tun, Aung Soe Khaing (2014),"Feature Extraction and Classification of Lung Cancer Nodule using Image ProcessingTechniques", International Journal of Engineering Research & Technology (IJERT), Vol. 3, No. 3, ISSN: 2278-0181.

[4] Lee HY et al (2015) "Differential expression of micro RNAs and their target genes in non-small-cell lung cancer", Mol Med Rep 11:2034–2040.

[5] Manogaran G, Shakeel PM, Hassanein AS, Priyan MK, Gokulnath C (2018)," Machine-learning approach based gamma distribution for brain abnormalities detection and data sample imbalance analysis", IEEE Access. https://doi.org/10.1109/ACCESS.2018.2878276.

[6] Madhura J, Dr .Ramesh Babu D R (2017), "A Surveyon Noise Reduction Techniques for Lung Cancer Detection", International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017).

[7] P. Muthamil Selvi, Dr.B. Ashadevi(2020),"Elimination of Noise in CT Images of Lung Cancer using Image Preprocessing Filtering Techniques", International Journal of Advanced Science and Technology, Vol. 29, No. 4s, pp. 1823- 1832.

[8] S.Ziyad, Dr.V.Radha, Dr.V.Thavavel (2019), "Performance Evaluation of Contrast Enhancement Techniques in Computed Tomography of Lung Images", *2019 5th International Conference for Convergence in Technology (I2CT), Pune, India.*

[9] T. Rajasenbagam, S. Jeyanthi (2020), "Image Enhancement Filtering Techniques to Enhance CT Images of Lung Cancer", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9, No. 4.

[10] N.R. Pal, S.K. Pal (1993), "A review on image segmentation techniques", Pattern Recognit., Vol. 26, pp. 1277–1294.

[11] Nida M Zaitouna, Musbah J Aqelb (2015), International Conference on Communication, Management and Information Technology, Elsevier, doi:10.1016/j.procs.2015.09.027.

[12] D.Q. Zhang, S.C. Chen (2004), "A novel kernelized fuzzy c-means algorithm with application in medical image segmentation", Artif. Intell. Med., vol. 32, pp. 37–50.

[13] S.R. Kannan, S. Ramathilagam, A. Sathya, R. Pandiyarajan (2010), "Effective fuzzy c-means based kernel function in segmenting medical images", Comput. Biol. Med., vol. 40, pp. 572–579.

[14] May Phu Paing , Kazuhiko Hamamoto, Supan Tungjitkusolmun and Chuchart Pintavirooj (2019),

"Automatic Detection and Staging of Lung Tumors using Locational Features and Double-Staged Classifications", Appl. Sci., doi:10.3390/app9112329.
[15] Ayushi Shukla, Chinmay Parab, Pratik Patil, Prof. Savita Sangam(2018), " Lung Cancer Detection using Image Processing Techniques", International Research Journal of Engineering and Technology (IRJET), vol. 5,No. 4, pp. 2517-2521.
[16] Anjali Kulkarni ,Anagha Panditrao (2014), "Classification of Lung Cancer Stages on CT Scan Images Using Image Processing", IEEE International Conference on Advanced Connnunication Control and Computing Teclmologies (lCACCCT), pp. 1384-1388.
[17] J. Kuruvilla and K. Gunavathi (2014), "Lung cancer classification using neural networks for CT images.," Comput. Methods Programs Biomed., vol. 113, no. 1, pp. 202–211.
[18] Song QZ, Zhao L, Luo XK, Dou XC (2017), "Using deep learning for classification of lung nodules on computed tomography images" J HealthcEng,https://doi.org/10.1155/2017/ 8314740.
[19] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian(2015), "Multi-scale convolutional neural networks for lung nodule classification," in Proceedings of 24th International Conference on Information Processing in Medical Imaging, pp. 588–599.
[20] L.I. Kuncheva (2004), "Combining Pattern Classifiers: Methods and Algorithms", John Wiley & Sons.
[21] R. Polikar," Ensemble based systems in decision making",Circuits Syst. Mag., IEEE, vol. 6, pp. 21–45.
[22] Omer Sagi, Lior Rokach (2018), "Ensemble learning: Asurvey", WIREs Data Mining Knowl Discov., https://doi.org/10.1002/widm.1249
[23] Shaohua Wan, Hua Yang (2013), "Comparison among Methods of Ensemble Learning", doi: 10.1109/ISBAST.2013.50.
[24] R. Polikar (2006), "Ensemble Based Systems inDecision Making," Circuits Syst. Mag. IEEE, vol. 6, no. 3, pp. 21–45.
[25] M.C. Lee, L. Boroczky, K. Sungur-Stasik, A.D. Cann, A.C. Borczuk, S.M. Kawut, C. A. Powell (2010), "Computer-aided diagnosis of pulmonary nodules usinga two-step approach for feature selection and classifier ensemble construction",Artif. Intell. Med., vol. 50, pp. 43– 53.
[26] F.V. Farahani, A. Ahmadi, M.F. Zarandi (2015), " Lung nodule diagnosis from CT images based on ensemble learning", Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE Conference, pp. 1–7.
[27] Alam J, Alam S, Hossan A (2018), "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier", In Proceedings of the InternationalConference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, pp. 1–4.
[28] Farzad Vasheghani Farahani, Abbas Ahmadi, Mohammad Hossein Fazel Zarandi (2018), "Hybrid intelligent approach for diagnosis of the lung nodule from CT images using spatial kernelized fuzzy c-meansand ensemble learning", Math. Comput. Simulation, https://doi.org/10.1016/j.matcom.2018.02.001.
[29] P. Mohamed Shakeel, M. A. Burhanuddin, Mohammad Ishak (2020), "Automatic lung cancer detection from CT image using improved deep neural network andensemble classifier", Springer-Verlag London Ltd., partof Springer Nature , https://doi.org/10.1007/s00521- 020-04842-6.
[30] Tanmoy Chakraborty (2014), "EC3: Combining Clustering and Classification for Ensemble Learning", Journal of Latex Class Files, Vol. 13, No. 9.
[31] Tahani Alqurashi, Wenjia Wang (2017), "Clustering ensemble method", International Journal of MachineLearning and Cybernetics,https://doi.org/10.1007/s13042-017-0756-7.
[32] Khedairia Soufiane, Mohamed Tarek Khadir (2019), "Amultiple clustering combination approach based on iterative voting process", Journal of KingSaudUniversity – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.09.013.
[33] B.E. Boser, I.M. Guyon, V.N. Vapnik (1992), "A training algorithm for optimal margin classifiers", Proc.5th Annu. ACM Work. Comput. Learn. Theory, pp.144–152.
[34] Mr. Chintan Shah, Dr. Anjali G. Jivani (2013), "Comparison of Data Mining Classification Algorithmsfor Breast Cancer Prediction ", IEEE 4th ICCCNT, Tiruchengode, India.
[35] George Dimitoglou, James A Adams, and Carol M Jim (2012), "Comparison of the C4.5 and a Naïve

Bayes Classifier for the Prediction of Lung Cancer Survivability", arXiv, eprint:1206.1121.
[36] Balázs Kég (2013), "The return of AdaBoost.MH: multi-class Hamming trees", arXiv, eprint:1312.6086.