Neural Processing Letters
https://doi.org/10.1007/s11063-020-10192-0

Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods

S. Shanthi[1,2] · **N. Rajkumar**[3]

Abstract

The symptoms of cancer normally appear only in the advanced stages, so it is very hard to detect resulting in a high mortality rate among the other types of cancers. Thus, there is a need for early prediction of lung cancer for the purpose of diagnosing and this can result in better chances of it being able to be treated successfully. Histopathology images of lung scan can be used for classification of lung cancer using image processing methods. The features from lung images are extracted and employed in the system for prediction. Grey level co-occurrence matrix along with the methods of Gabor filter feature extraction are employed in this investigation. Another important step in enhancing the classification is feature selection that tends to provide significant features that helps differentiating between various classes in an accurate and efficient manner. Thus, optimal feature subsets can significantly improve the performance of the classifiers. In this work, a novel algorithm of feature selection that is wrapper-based is proposed by employing the modified stochastic diffusion search (SDS) algorithm. The SDS, will benefit from the direct communication of agents in order to identify optimal feature subsets. The neural network, Naïve Bayes and the decision tree have been used for classification. The results of the experiment prove that the proposed method is capable of achieving better levels of performance compared to existing methods like minimum redundancy maximum relevance, and correlation-based feature selection.

Keywords Lung cancer · Small cell lung cancer (SCLC) · Non-small cell lung cancer (NSCLC) · Radiomic features · Gray level co-occurrence matrix (GLCM) · Gabor filter · Stochastic diffusion search (SDS) · Neural network (NN) · Naive Bayes and decision tree

✉ S. Shanthi
shan.sece@gmail.com

N. Rajkumar
drnrk42@gmail.com

[1] Anna University, Chennai 600025, India

[2] Department of CSE, Sri Eshwar College of Engineering, Coimbatore, India

[3] Hindusthan College of Engineering and Technology, Coimbatore, India

Published online: 08 January 2020

1 Introduction

Cancer has been identified as a dangerous disease resulting in death. According to the data obtained from the Global Burden Cancer, there have been 14.1 million cases of different types of cancer in the year 2012 where lung cancer stands first with 13%. The deaths due to cancer have been recorded as 8.2 million wherein lung cancer caused 19% of these deaths. The cancer cells spread from the lungs passing through the lymph fluid or the bloodstream. Normally, the cancer cells spread towards the centre of the chest because of the natural flow of the lymph. Metastasis will take place if this cancer spreads to the other organs and early detection helps in preventing this [1].

1.1 Types of Lung Cancer

Lung cancer is also called lung carcinoma. It is a tumour which is malignant and has uncontrolled growth of cell tissues. It is important to treat it early to avoid metastasis and most lung cancers are carcinomas. Tobacco smoking for a long period is the main factor that contributes to 85% of cancer in the lung [2]. Around 10–15% of such cases tend to occur among those people that have not ever smoked in their lives but have been exposed to air that is polluted, asbestos, passive smoking or radon gas. Both computer tomography (CT) and radiographs were the conventional methods to detect the present of such cancer.

Identified are two other types of cancer and these are: (1) The Small Cell Lung cancer (SCLC). (2) The non-small cell lung cancer (NSCLS). The actual size of the tumour will indicate the staging of cancer found in the lung nodes. There are however four different stages identified in lung cancer. The tumour, if identified in either the first or the second stage, chances of recovery are high. In case it is discovered in the third stage it becomes difficult. Identification in the fourth stage is extremely dangerous and chances of tumour removal is not possible. In order to treat lung cancer, the processes of CT-scan, biopsy, surgery, radiotherapy, and chemotherapy need to be done. Thus, we see the need for techniques of image processing that are extremely useful for this [3].

1.2 Radiomics Features

There is a data mining approach of high throughput known as radiomics, that can exploit the quantitative imaging features providing a comprehensive and detailed characterisation of the phenotype of the tumour. There have been various investigations that were conducted using several modalities of medical imaging such as the CT, ultra sound (US). The positron emission tomography (PET) and magnetic resonance imaging (MRI). It is noted that the radiomic features have been associated to various factors like the gene expression profiles, treatment response, metastasis and survival of the patients. All these associations are leveraged in order to build an effective and efficient model of prognostics. Thus, radiomics has been a very promising field which is cost-effective and non-invasive in personalized medicine [4].

1.3 Feature Selection

Another technique of image reduction used widely in image mining or knowledge discovery that permits the elimination of all redundant features and ensures retention of relevant features is Feature selection. It further removes all redundant and irrelevant features where the image recognition system's quality has to be improved to enhance the learning system and its performance. The choice of the image is done using the technique of feature selection like selecting, ranking or screening. Screening can remove all unwanted features and records. Ranking is used for sorting all remaining features on the basis of importance. Selecting will be used for the identification of the feature subsets by means of preserving the features that are critical and filtering the remaining ones. The technique of feature selection will thus screen, rank and select features which are important [5].

The different methods of feature selection had been grouped into three that are based on the manner in which they have been combined and these are the filter, the wrapper, and the embedded methods. Filter methods tend to assess the relevance of that of the features in the form of scores. These features get sorted by means of using their scores and also the ones that score low which are eventually removed. The wrapper methods, they embed an analysis model for the setup of which there is the future subset that is evaluated using the application of a model that has specific analysis to reduced data that makes use of the feature subset chosen. For the embedded methods, search is always for the optimal feature and their subsets built in the analysis algorithm. The filter methods were used as a fast, simple and independent method. They further permit these features to get prioritized and are quantified on the basis of scores which may be critical for their biological interpretation [6].

1.4 Motivation

Image processing techniques help in identification of cancer automatically using machine learning methods. A framework for identifying lung cancer helps medical professionals in processing large amount of lung images quickly for pre-screening. To improve the performance of the framework, feature selection method to obtain an optimal feature subset is essential. As feature selection is an NP-hard problem, it is proposed to use heuristic methods to achieve optimal feature subset. In this work, the use of heuristic algorithm for feature selection is investigated.

1.5 Contributions

The main contributions of this work are:

- The use of feature fusion of GLCM and Gabor filter features extracted.
- Use of symbolic data state, histology of lungs were obtained over maximum of 3 different time intervals and used for classification of the lung images.
- Novel feature selection method based on stochastic diffusion search (SDS) algorithm is proposed which achieves significantly improved performance in classifying the images.

1.6 Organization of the Paper

In this work, the algorithm known as the wrapper-based SDS feature selection for diagnosis of lung cancer has been proposed. The rest of the investigation has been organized thus. Section 2 explains all related work in literature. Section 3 explains the methods used. Section 4 discusses the results of the experiment and the conclusion is made in Sect. 5.

2 Related Works

Asuntha et al. [7] had made a discussion on the lung cancer formation and the system used for detection. The system was able to take any medical image using three different choices that consisted of the CT, the US, and the MRI images. The method employed the support vector machine (SVM), the genetic optimization and the particle swarm optimization (PSO). It was an extension of the image processing in the detection of lung cancer which produces results after segmentation. This system was formed based on any medical image that was within all of the three different choices that consist of the MRI, the US, and the CT. Once the image is pre-processed, there is a canny filter used for its edge detection. This work also proposed a new method for the detection of all cancerous cells from CT, the US, and the MRI. There was super pixel segmentation that was employed for the segmentation and the Gabor filter was employed for medical image de-noising.

The patients of lung cancer resulted in higher numbers of deaths with survival rates that were the lowest and for this, Silva et al. [8] had exploited another technique of deep learning combined with the genetic algorithm (GA) for the purpose of classification of lung nodules into either benign or malignant without having the need to be able to compute the texture features and also their shape. This methodology was duly tested on the basis of the CT images that were obtained from the lung image database consortium and the image database resource initiative (LIDC-IDRI) and this had a higher level of sensitivity of about 94.66%, a specificity of about 95.14% and an accuracy of about 94.78% along with the area under the ROC curve of about 0.949.

Veeramani and Muthusamy [9] had further employed another adaptive median filtering technique which was identified as a step of pre-processing. This image which is pre-processed had been extracted using a convoluted local tetra pattern, complete local binary patter (LBP), Haralick feature extraction and histogram of oriented gradient (HOG). These extracted features are later selected by means of applying the PSO along with the differential evolution (DE) feature selection. The final stage is during the time classifiers such as the relevance vector machine (RVM), and the multi-level RVM had been employed for the performance of classification of lung diseases. Diseases such as the respiratory distress syndrome (RDS), the transient tachypnea among new-borns, the meconium aspiration syndrome, pneumothorax, lung cancer, pneumonia and bronchiolitis were employed for both training and testing. An experimental analysis had exhibited a better level of fitness value, pixel count, specificity, sensitivity and finally accuracy.

The detection of lung nodules becomes a critical aspect in the CAD systems and can result in generating several false positives (FPs). da Silva et al. [10] had proposed another methodology for reducing the FPs by employing the technique of deep learning which was in conjunction with other evolutionary techniques. PSO algorithms were used for optimizing the hyper parameters of the network within a convolutional neural network (CNN) for enhancing the performance of the network and for eliminating the manual search. This methodology had been tested using the CT scans from LIDC-IDRI that had the accuracy of about 97.62%, sensitivity of about 92.20% and the specificity of about 98.64%, with an area under a receiver operating characteristic (ROC) curve of about 0.955.

Generally, there is a measure for diagnosis in the early stage by using X-ray, MRI, and CT. D'Cruz et al. [11] had presented another task of medical image mining like the back propagation neural network (BPPN) that was able to classify digital X-rays, MRI and CT images as either normal or abnormal. A normal state will characterize the patient as healthy. An abnormal image is taken for feature analysis. For the optimized GA feature analysis, it extracts the features based on the fitness. The chosen features are then grouped based on whether they are grouped as cancerous or noncancerous. Thus, the system can help in drawing suitable decisions regarding the state of patients.

Naqi et al. [12] had presented a new method of automated detection and classification that facilitated radiologists in the process of diagnosis. There was a nodule detection with the classification that was proposed consisting of four different phases. The first was the lung region extraction that was performed on the basis of an optimal grey level threshold computed using the fractional-order Darwinian PSO. After this, there was yet another method of nodule candidate detection duly based on a geometric fit found within a parametric form that had incorporated the nodules and their properties that was proposed. The subsequent phase included a hybrid geometric texture feature descriptor which was created for being able to better represent all candidate nodules that had been a combination of 2D and 3D. Lastly, there had been a deep learning approach that was based on that of the stacked auto-encoder and the softmax which was used for both feature reduction and also its classification that had been applied for the purpose of reducing the FPs.

Zhang et al. [13] had proposed yet another automated system used for diagnosing lung cancer and this system was designed using a combination of two of the major methodologies which are the fuzzy base systems and evolutionary GAs, that were employed on the data of lung cancer for assisting the physicians in detecting it at the early stages. This hybrid algorithm which was known as the genetic-fuzzy algorithm, was able to produce some optimized diagnosis systems attaining a high performance of classification and the best six rule system had obtained an accuracy of 97.5 % with interpretive rules that were simple and a degree of confidence of about 93 % that did not have the need for reduction of dimensionality. The results based on the real data had indicated that the system was effective in diagnosing lung cancer and also used for their clinical applications.

Bhuvaneswari and Therese [14] had proposed a new genetic *K*-nearest neighbour (GKNN) algorithm to detect the methods that were identified to be non-parametric. This algorithm permits the physicians to be able to identify all the nodules which were present in these CT images of that of the lung in their early stage. As yet another manual interpretation of lung cancer in CT images were found to be very time-consuming, the GA method was combined along with the *K*-nearest neighbour (*K*-NN) algorithm that classifies cancer images effectively. In the case of a traditional *K*-NN algorithm, the distance between the testing samples and the training samples calculated and the K-neighbours that had greater distances which were used for classification. in the method proposed, using the GA can make K (50–100) sample numbers selected for every iteration and accuracy classification of about 90% which had been achieved as its fitness.

3 Methodology

In this work, symbolic data with radiomic features was employed. In the section, the GLCM with the method of Gabor filter feature extraction were employed and the SDS method of feature selection with decision tree classifiers, NN and Naïve Bayes was employed.

3.1 Dataset

140 normal and 130 abnormal images were used from the cancer genome atlas (TCGA) dataset was used in this investigation. For symbolic data state, histology of lungs were obtained over maximum of 3 different time intervals.

3.2 Radiomic Feature Extraction

Feature extraction refers to the process that acquires information on the texture, shape and colour. The features consisting of the relevant information used in image processing such as searching, retrieval and storing. Feature extraction includes the reduction of resources needed for describing large data. In radiomic feature extraction, quantitative features which can be extracted from images, such as texture, and shape are applied. In this system, in order to extract texture features vectors, the grey level co-occurrence matrix (GLCM) and Gabor filter for shape features is used [15].

GLCM method was proposed by Haralic in the year 1973 and continues to be a very popular method of texture analysis. The functions of the GLCM had characterised the texture by means of computing the frequency of the pairs of a pixel that have specific values within a spatial relationship that comes in an image. When working with the GLCM, the creation of GLCM, the specifying of offsets and extraction of statistical measures were the primary features. Texture features were important to identify the object of the region of interest (ROI).

A Gabor filter denotes linear filters that have an impulse replication defined using a harmonic function that was multiplied using a Gaussian function. Owing to the fact that it has a property of convolution, the Fourier transform of Gabor filters and their impulse replication denote convolution for the Fourier transform which is for the harmonic function and also the Fourier transform for the Gaussian function [16]. This will fundamentally analyse in case there has been a frequency content within the image and their localized regions. There was also another set of Gabor filters with various frequencies along with their orientations used for extraction of useful features from images. For a discrete domain, the 2D Gabor filters were given by (1),

$$\begin{aligned} G_c[i,j] &= Be^{\frac{i^2+j^2}{2\alpha^2}} \cos\left(2\prod f(i\cos\theta + j\sin\theta)\right) \\ G_s[i,j] &= Ce^{\frac{i^2+j^2}{2\alpha^2}} \sin\left(2\prod f(i\cos\theta + j\sin\theta)\right) \end{aligned} \tag{1}$$

wherein B and C denote all normalizing factors that need to be determined. The 2D Gabor filters tend to have certain rich applications in feature extraction for segmentation and texture analysis. f will define the frequency that is looked at for texture. By means of bringing about a variation to the θ, it may look out for the texture-oriented aspects in a certain direction. By means of varying the α, it will be able to ensure support based on the size of image regions that were analysed.

The extracted features are concatenated to form a feature subset. One drawback of radiomic features are the high dimension of the feature set. To improve the efficacy of the classification, it is thus required to optimize the feature subset using feature selection techniques.

3.3 Stochastic Diffusion Search (SDS) Algorithm Based Feature Selection

SDS for this method of feature selection will be using the communication for the performance of subset evaluation in an effective manner. In the initial stages, every agent will be assigned to combining the feature subset from their respective search spaces (all the possible combinations of these features). Every agent will now employ an independent and random split of the dataset for forming both training and testing subsets of 80% and 20% of a dataset. A hypothesis here denotes the binary string representing the feature subset this was well within its subset size. For the purpose of employing this string, if the bit was 1, its corresponding feature will be included and in case it is 0, it will not be the same [17].

In the case of the test phase, these activities of the agent had been determined on the basis of a classifier in their fitness function. This was further compared to the predictive accuracy in case the agent has an accuracy which is more than that of the accuracy of a random agent. The process will be repeated for the agents in order to determine their respective status. After this, the diffusion phase will ideally begin.

In the case of the diffusion phase, both the inactive and the active agents will pick other agents. In case the agent selected randomly is active, it will offset the hypothesis (the feature subset) and this will be shared with an inactive agent. In case this does not happen, the chosen agents will pick a new and random hypothesis (the feature subset) from its search space (all feature combinations in the subset size). For the purpose of offsetting, one of the features were randomly removed (which was by changing 1 to 0) with another randomly added one (by means of changing 0 to 1), and this will help in preserving the size of the subset. Furthermore, at the time the active agent picks another active agent maintaining a similar hypothesis (the feature subset) and the selecting agent will be set to inaction thus assigning them to a new and random hypothesis. This will now free up all the agents, and improve the diversity which increases the ability of various algorithms to be able to make a wide search throughout the entire search space. The cycle of this test and its diffusion gets repeated equal to the actual iterations that have been allowed.

SDS algorithm for feature selection as shown below:

```
#Initialisation phase
Assign agents to random hypotheses with inactive states, each agent represents a set of features
#Evaluation Phase
Evaluate the fitness value
Find the max imum fitness value
#Test Phase
if Agent's fitness > random agent's fitness then
Set agent as active else inactive
#Diffusion Phase
if agent is inactive then Select a random agent
if selected agent is active then Copy its hypothesis & offset it
Evaluate the fitness value
else Pick a random hypothesis
Evaluate the fitness value
Maximum iterations - Ter min ate
Optimal feature subset obtained
```

3.4 Classifiers Used

3.4.1 Decision Tree Learning

Employs decision trees to be the predictive model that maps the observations regarding an item for conclusions on the target value of the item. The decision tree algorithm was a new technique of data mining induction which will now be able to recursively partition the record dataset that makes use of the depth-first greedy approach and sometimes the breadth-first approach until such time all of these data items are part of the class. There is a new decision tree structure which is made of the root, the internal and the leaf nodes. This is taken to be a tree structure such as a flow chart in which each internal node indicates the test condition on the attribute and every branch indicates the results of the test condition along with every leaf node (terminal node) will be assigned to a class label. There is a topmost node that is a root node. This decision tree was constructed in a new approach of divide and conquer. Every path in the decision tree will form the decision rule. Normally, it employs the greedy approach that is used from the top to the bottom [18].

The classification of a decision tree gets performed by using two different phases: the tree building and the tree pruning. The former was performed by using a top-down approach. In this phase, the tree gets partitioned recursively until such time all data items are part of the same class label. This will be very intensive computationally since this training dataset has been repeatedly traversed. Tree pruning, on the other hand, will be used in a new manner in which it was bottom-up. This was employed in order to improve the classification accuracy along with its prediction by means of minimizing the problem of over-fitting which can lead to errors being misclassified.

3.4.2 Naïve Bayes

The technique known as the Naive Bayes Classifier was based on the Bayesian theorem which was well-suited at the time the input dimensionality is high. It considers all attributes or features that contribute independently to the probability of these classes. The classifier also is trained to work on the method of supervised learning and works well in various real-world situations that are complex. In spite of its simplicity, the Naïve Bayes was able to outperform many of the other sophisticated methods of classification. This Bayesian approach permits the scientists to be able to combine all new data along with their current expertise or knowledge. By making use of the training dataset, Bayesian classifiers are able to determine the chances of being associated to some of these classes in some of the instances with the values of all predictor variables. The Naive Bayes classifier also performs in a way that is equivalent to all other techniques of machine learning that have a low effort of computation and high levels of speed [19].

3.4.3 Neural Network (NN)

Refers to a network of neurons that are interconnected transmitting patterns of electrical signals. The artificial neural network (ANN) denotes a learning neuron cluster that is based on the biological NNs (the human brain). Normally, the NN contains 100 billion neurons and every neuron will be connected to about 10,000 other neurons. The ANNs are normally presented as the systems "the neurons" and these are interconnected for exchanging messages among various neurons. These connections will have some numeric weights which adapt on the basis of the experience thus creating the neural nets that are adaptive to the inputs and are capable of learning. The main advantage of the ANN was that they were well-suited to solve the problems that are complex which may be solved using some conventional techniques, or it could be hard to find the algorithmic solutions. There was also another layer that was for the input variables and one more for its output. These layers include the following [20]:

- The input layer—this includes the input units that indicate all unrefined information that is provided for a network.

Table 1 Summary of results

	Classification accuracy	Recall for normal	Recall for AD	Precision for normal	Precision for AD
MRMR—decision tree	83.33	0.9071	0.7538	0.7987	0.8829
CFS-decision tree	82.22	0.9214	0.7154	0.7771	0.8942
SDS-decision tree	87.41	0.95	0.7923	0.8313	0.9364
MRMR-Naïve Bayes	81.85	0.9286	0.7	0.7692	0.901
CFS-Naïve Bayes	81.48	0.9143	0.7077	0.7711	0.8846
SDS-Naïve Bayes	88.52	0.9571	0.8077	0.8428	0.9459
MRMR-neural network	87.04	0.9571	0.7769	0.8221	0.9439
CFS-neural network	85.19	0.95	0.7462	0.8012	0.9327
SDS-neural network	89.63	0.9571	0.8308	0.859	0.9474

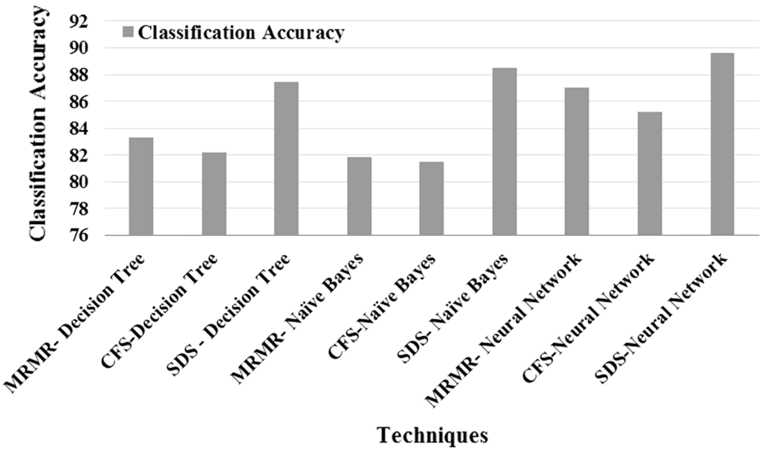

Fig. 1 Classification accuracy for SDS-NN

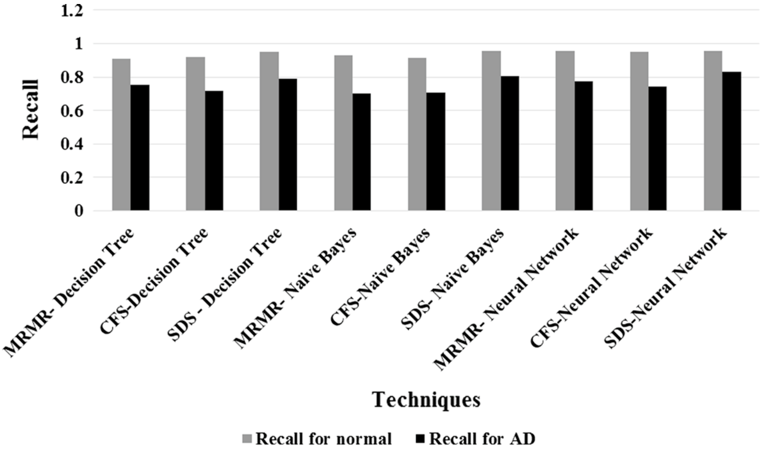

Fig. 2 Recall for SDS-NN

- The hidden layer—this includes all hidden units on the basis of the behaviour of the input unit and their weighted neurons that are connected to the input in hidden units.
- The output layer—this is based on hidden units and their specificity and its weighted neuron.

4 Results and Discussion

In this section, the results of the experiments carried out are presented. 140 normal and 130 abnormal images were used. For symbolic data state, histology of lungs were obtained over maximum of three different time intervals are considered. The features are extracted for the different time interval and concatenated which is then used for classifying the images.

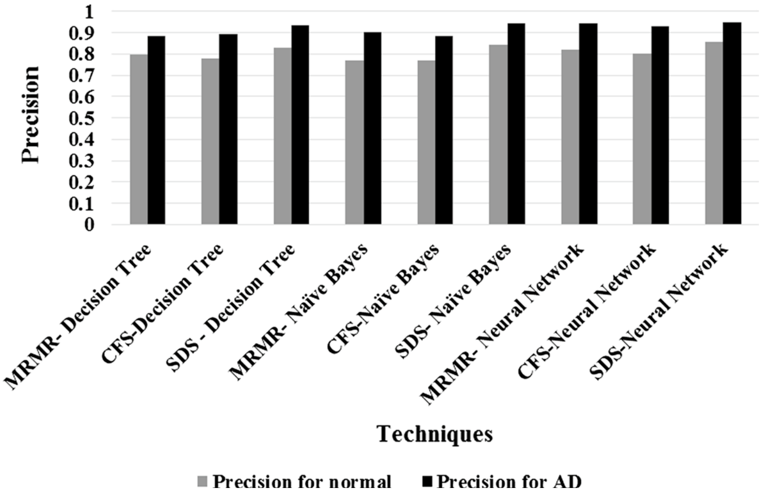

Fig. 3 Precision for SDS-NN

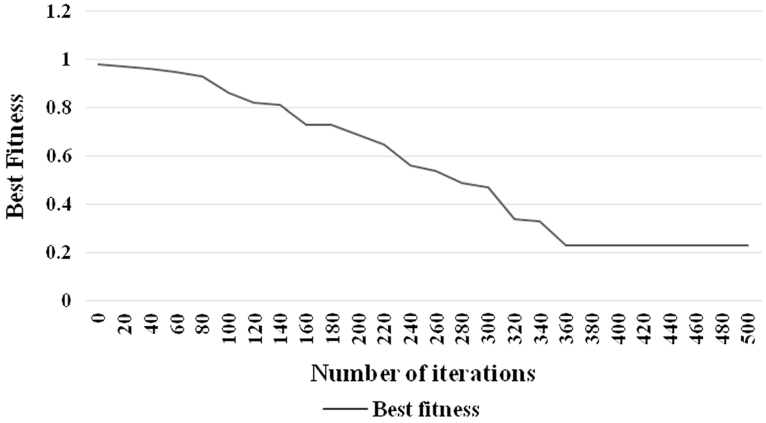

Fig. 4 Best fitness

The proposed SDS method is compared with minimum redundancy maximum relevance (MRMR), and the correlation-based feature selection (CFS) techniques.

The MRMR-decision tree, CFS-decision tree, SDS-decision tree, MRMR-Naïve Bayes, CFS-Naïve Bayes, SDS-Naïve Bayes, MRMR-NN, CFS-NN and SDS-NN methods are investigated. Table 1 shows the summary of results. The classification accuracy, recall for normal and AD and precision for normal and AD as shown in Figs. 1, 2 and 3. Table 1 and Fig. 4 shows the best fitness.

Classification accuracy is defined as the number of correctly classified images to the total number of images, in this work, it is given in percentage. From the Fig. 1, it can be observed that the SDS-NN has higher classification accuracy by 7.28% for MRMR-decision tree, by 8.62% for CFS-decision tree, by 2.51% for SDS-decision tree, by 9.07% for MRMR-Naive Bayes, by 9.52% for CFS-Naive Bayes, by 1.24% for SDS-Naive Bayes, by

Table 2 Best fitness

Number of iterations	Best fitness
0	0.98
20	0.97
40	0.96
60	0.95
80	0.93
100	0.86
120	0.82
140	0.81
160	0.73
180	0.73
200	0.69
220	0.65
240	0.56
260	0.54
280	0.49
300	0.47
320	0.34
340	0.33
360	0.23
380	0.23
400	0.23
420	0.23
440	0.23
460	0.23
480	0.23
500	0.23

2.93% for MRMR-NN and by 5.07% for CFS-NN respectively. It is observed that the feature selection technique does significantly improve the efficacy of all the classifiers. The neural networks are more efficient in classifying the images (Table 2).

Recall is also known as sensitivity, is the fraction of the total number of relevant images that are correctly classified. From the Fig. 2, it can be observed that the SDS-NN has higher recall for normal by 5.36% for MRMR-decision tree, by 3.8% for CFS-decision tree, by 0.74% for SDS-decision tree, by 3.02% for MRMR-Naive Bayes, by 4.57% for CFS-Naive Bayes, by same value for SDS-Naive Bayes & MRMR-NN and by 0.74% for CFS-NN respectively. The SDS-NN has higher recall for AD by 9.71% for MRMR-decision tree, by 14.92% for CFS-decision tree, by 4.74% for SDS-decision tree, by 17.08% for MRMR-Naive Bayes, by 16% for CFS-Naive Bayes, by 2.82% for SDS-Naive Bayes, by 6.7% for MRMR-NN and by 10.72% for CFS-NN respectively. It is seen that the proposed SDS feature selection helps improve the recall as optimal feature subset is used as input for the classifiers.

From the Fig. 3, it can be observed that the SDS-NN has higher precision for normal by 7.27% for MRMR-decision tree, by 10.01% for CFS-decision tree, by 3.27% for SDS-decision tree, by 11.03% for MRMR-Naive Bayes, by 10.78% for CFS-Naive Bayes, by 1.9% for SDS-Naive Bayes, by 4.39% for MRMR-NN and by 6.96% for CFS-NN respectively.

The SDS-NN has higher precision for AD by 7.04% for MRMR-decision tree, by 5.77% for CFS-decision tree, by 1.16% for SDS-decision tree, by 5.02% for MRMR-Naive Bayes, by 6.85% for CFS-Naive Bayes, by 0.15% for SDS-Naive Bayes, by 0.37% for MRMR-NN and by 1.56% for CFS-NN respectively.

From the Fig. 4, it can be observed that the convergence is achieved in about 350th iteration.

5 Conclusion

Lung cancer has been observed to be a very dangerous disease which is wide-spread which is also the most common cause of death. The object for this is the prediction of early detection by means of employing classifiers that have optimal features. Feature selection has been used for the identification of predictive subsets of the cancer cells in a database that can bring down the cancer cells. There is better performance when some features are discarded. The method also introduced the SDS for choosing all relevant subsets for the task of classification. For the purpose of this algorithm, the SDS was adapted to choose a feature subset that was suitable. The techniques of classification were able to handle large volumes of data and their processing. The Naive Bayes classifier was a simple classifier which assumes the variables that contribute to the classification that has been correlated mutually. The NNs will deal with other problems where the neurons were trained and also tested with the given database. An evaluation of the performance of the method proposed has shown some effective results that indicate the NN which was used effectively for the diagnosis of lung cancer for helping oncologists. The results prove that an SDS- NN had an accuracy of classification by about 2.51% for the SDS-decision tree and further by about 1.25% for the SDS-Naive Bayes respectively. It is seen that feature selection improves the classification of images, further investigation to optimize the classifier needs to be explored. This work focused on feature selection, some pre-processing methods such as noise removal, optimal classifiers can be further investigated.

References

1. Zhang G, Jiang S, Yang Z, Gong L, Ma X, Zhou Z, Bao C, Liu Q (2018) Automatic nodule detection for lung cancer in CT images: a review. Comput Biol Med 103:287–300
2. Senthil Kumar K, Venkatalakshmi K, Karthikeyan K (2019) Lung cancer detection using image segmentation by means of various evolutionary algorithms. Comput Math Methods Med 2019:4909846. https://doi.org/10.1155/2019/4909846
3. Sevani A, Modi H, Patel S, Patel H (2018) Implementation of image processing techniques for identifying different stages of lung cancer. Int J Appl Eng Res 13(8):6493–6499
4. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, Madabhushi A (2018) Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer 115:34–41
5. Kishore MR (2015) An effective and efficient feature selection method for lung cancer detection. Int J Comput Sci Inf Technol (IJCSIT) 7(4):135–141
6. Narayanan BN, Hardie RC, Kebede TM, Sprague MJ (2019) Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities. Pattern Anal Appl 22(2):559–571
7. Asuntha A, Singh N, Srinivasan A (2016) PSO, genetic optimization and SVM algorithm used for lung cancer detection. J Chem Pharm Res 8(6):351–359
8. da Silva GL, da Silva Neto OP, Silva AC, de Paiva AC, Gattass M (2017) Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed Tools Appl 76(18):19039–19055

9. Veeramani SK, Muthusamy E (2016) Detection of abnormalities in ultrasound lung image using multi-level RVM classification. J Matern Fetal Neonatal Med 29(11):1844–1852
10. da Silva GLF, Valente TLA, Silva AC, de Paiva AC, Gattass M (2018) Convolutional neural network-based PSO for lung nodule false positive reduction on CT images. Comput Methods Programs Biomed 162:109–118
11. D’Cruz J, Jadhav A, Dighe A, Chavan V, Chaudhari J (2016) Detection of lung cancer using back-propagation neural networks and genetic algorithm. Comput Technol Appl 6(5):823–827
12. Naqi SM, Sharif M, Jaffar A (2018) Lung nodule detection and classification based on geometric fit in parametric form and deep learning. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3773-x
13. Zhang G, Jiang S, Yang Z, Gong L, Ma X, Zhou Z, Bao C, Liu Q (2018) Automatic nodule detection for lung cancer in CT images: a review. Comput Biol Med 103:287–300
14. Bhuvaneswari P, Therese AB (2015) Detection of cancer in lung with K-NN classification using genetic algorithm. Procedia Mater Sci 10:433–440
15. Kohad R, Ahire V (2015) Application of machine learning techniques for the diagnosis of lung cancer with ANT colony optimization. Int J Comput Appl 113(18):34–41
16. Johora FT, Jony MH, Khatun P, Rana HK (2018) Early detection of lung cancer from CT scan images using binarization technique (No. 545). EasyChair
17. Alhakbani H, al-Rifaie MM (2017) Feature selection using stochastic diffusion search. In: Proceedings of the genetic and evolutionary computation conference. ACM, pp 385–392
18. Jadhav SD, Channe HP (2016) Comparative study of K-NN, Naive Bayes and decision tree classification techniques. Int J Sci Res 5(1):1842–1845
19. Hosseinzadeh F, KayvanJoo AH, Ebrahimi M, Goliaei B (2013) Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus 2(1):238
20. Senthil S, Ayshwarya B (2018) Lung cancer prediction using feed forward back propagation neural networks with optimal features. Int J Appl Eng Res 13(1):318–325

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.