INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Lung Cancer Detection by Multiple Feature Subset Extraction and Selection based on SVM-Weights and Genetic Algorithm-Neural Network

S S Ashwini[1]*, **M Z Kurian**[1], **M V Chidanandamurthy**[1], **M Nagaraja**[2]

1 Department of ECE, SSIT, Sri Siddhartha Academy of Higher Education, Tumakuru, 572105, India
2 Department of Physics, SSIT, Sri Siddhartha Academy of Higher Education, Tumakuru, 572105, India

OPEN ACCESS
Received: 08-05-2023
Accepted: 05-07-2023
Published: 05-08-2023

Citation: Ashwini SS, Kurian MZ, Chidanandamurthy MV, Nagaraja M (2023) Lung Cancer Detection by Multiple Feature Subset Extraction and Selection based on SVM-Weights and Genetic Algorithm-Neural Network. Indian Journal of Science and Technology 16(29): 2269-2275. https://doi.org/10.17485/IJST/v16i29.1093

* **Corresponding author**.

ashwini_249@yahoo.co.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop an optimum hybrid approach for lung cancer detection by multiple feature subset extraction and selection based on SVM-weights and Genetic Algorithm (GA-NN) in order to improve the performance measures such as accuracy, sensitivity and specificity. **Methods:** Initially in preprocessing phase, Computed Tomography (CT) lung images are de-noised using median filter and enhanced using contrast stretching. In the next phase, candidate patch extraction is formed and Gray Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP) features are extracted. This is followed by feature selection using Genetic Algorithm-Neural Network (GA-NN) with SVM weights. Finally, images are classified as cancerous and non-cancerous using multiple classifiers (SVM and KNN). For this research work, CT lung images are collected form LIDC dataset. Around 500 images are used out of which 70% is used for training and 30% is used for testing. **Findings:** From simulation results and comparative analysis, it is observed that GANN with SVM weights result in better predictive performance metrics with notable improvements. The suggested feature subset reduction outperforms current techniques for detection of lung cancer in CT images. The proposed method has resulted in improved accuracy, specificity and sensitivity by 95.8%, 91.3% and 93.5% respectively which is higher than the existing approaches. **Novelty:** This work presents a novel approach to detect the lung cancer by multiple feature subset extraction and selection based on SVM-Weights and Genetic Algorithm - Neural Network (GA-NN) with improved accuracy, sensitivity and specificity.

Keywords: Gray Level Cooccurrence Matrix (GLCM); Genetic AlgorithmNeural Network (GANN); KNearest Neighbor (KNN); Local Binary Pattern (LBP); Support Vector Machine (SVM)

1 Introduction

The rate of lung cancers among the population has hastily increased during the 20th century and became a large threat to the public particularly for passive and/or active smokers and those with higher exposure to radiation or chemicals in the workplace. The tumors in lung region at a certain stage are visible to experienced radiologists in different image modalities such as chest x-rays, CT scans, PET scans and MRI scans. At present, the survival rate for lung cancerous is around 10% and it is attributed as late detection [(1)].

Classification of lung cancer at the earlier stage is thus imperative to ensure the higher survival rates. However, it is very challenging task. Currently cancer classification is based on subjective interpretation of histopathological and clinical data. Clinical information may be inadequate at times and the wide classes of most tumors lack morphologic features which are indispensable for classification [(2–4)].

Many authors have worked on lung cancer detection using different approaches. Lung cancer detection and segmentation is presented in [(5)] using deep learning algorithm which achieves a sensitivity of 73%. This sensitivity can be enhanced by selecting efficient feature from the image for classification. As authors have chosen chest radiography as image modality, false-positive and false-negative are more when they are overlapped with normal anatomical structures, such as heart clavicle and ribs. Authors in [(6)] have discussed the lung cancer detection using SVM. Dataset consists of malignant, benign and pre-malignant images, where PCA has been used for feature selection strategies to cut down the dimensionality. An optimal features selection using feature co-relation followed by classification using SVM is carried out. Overall accuracy achieved in the proposed strategy is 87%, which could have been improved with appropriate feature selection. However, the proposed approach is not compared with the existing methods to reveal the efficiency of proposed method.

M. Sathya and et.al in [(7)] proposed a distinctive gene selection method which combines minimal Redundancy and Maximum Relevance ensemble (mRMRe) and genetic algorithm was used to increase classification accuracy for four micro array datasets while utilizing few numbers of selected genes. First stage of gene selection was done using mRMRe gene to identify necessary genes that have the least degree of redundancy. The main drawback of this approach was as number of genes increases; the accuracy tends to decrease. Early detection and classification of lung cancer has been investigated in [(8)] using geometric filtering as preprocessing, K-means for segmentation, ANN and KNN for classification, with an achieved accuracy of 90-92% with limited database. Syed Afsar Ali Shah Tirmzi et al [(9)] proposed a modified GA based technique for classification of brain tumor MRI images. This method is implemented on brain MRI datasets in which GLCM features are extracted from pre-processed image. Further GA is applied to extract optimal higher statistical order feature set. SVM is used for classification and attained higher performance measure. This technique can be applied for other type of cancer detection also. Article [(10)] is on innovative Wilcoxon Signed-Rank gain preprocessing for feature selection to reduce the computational time and search space using deep learning for lung cancer classification. The proposed method has achieved an accuracy of 87%. S. Akila Agnes et al [(11)] used convolution long short–term (ConvLSTM) method for classification of Lung CT image as malignant or benign with an accuracy of 92%. This method effectively captures spatial features in image in sequences and steadily compared to single dimensional LSTM based method. In case of early detection and prediction of lung cancer malignant tumors are classified as benign.

2 Methodology

Figure 1 represents proposed framework using GA-NN and SVM classifier for feature reduction. At first, preprocessing is done in order to enhance the image. In the next step, patches are separated and named as non-cancerous and cancerous, based on the information provided in the annotated dataset. From these patches, GLCM and LBP features are extracted. Further by employing the SVM weight and GA-NN technique, optimized feature subset is extracted. Classification algorithm such as SVM and KNN utilizes the optimized subset features to classify the images as cancerous or non-cancerous image.

2.1 Image Pre-Processing

As CT images are sensitive to noise, minimum contrast and non-uniform lighting, details are hardly appearing with frail contrast and minimum luminosity. Pre-processing steps are carried out in order to minimize these constraints and rendering the image acceptable for further processing and feature extraction. To minimize the noise, median filter of size [3×3] is used. Further filtered images are enhanced using contrast stretching. Figure 2 shows the noisy and preprocessed images.

2.2 Candidate Patch Extraction

A sample is said to be positive on the basis of reported cancerous by an expert. For every cancerous, 128×128 square window cancerous image is formed along with the annotated coordinates as middle. Non-cancerous patches have been haphazardly

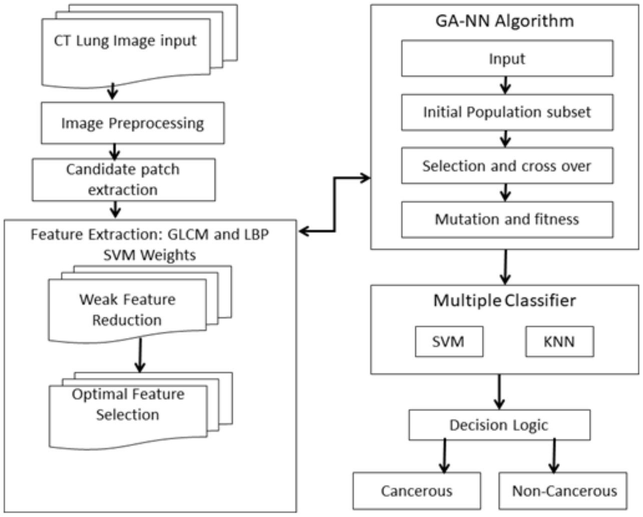

Fig 1. Proposed framework using GA-NN and SVM classifier for feature reduction

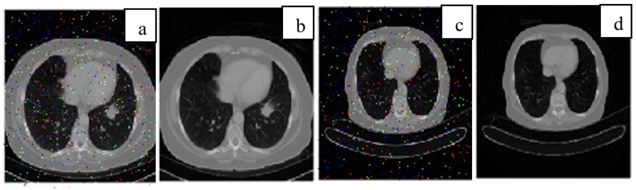

Fig 2. Lung CT Image shown in (a) is cancerous noisy image (b) cancerous preprocessed image (c) noncancerous noisy image and (d) noncancerous preprocessed image

obtained across the image (12).

2.3 Feature Extraction

Local Binary Pattern (LBP) and Grey Level Co-occurrence Matrix (GLCM) are employed for texture feature extraction. Comparisons of gray scale and local spatial variations are collected effectively by LBP descriptors as expressed in equation 1 and 2 (12).

$$LBP(N,L) = \sum_{n=0}^{N-1} F\left(I_n - I_c\right) \cdot 2^n \tag{1}$$

$$I(y) = \begin{cases} 1, & \text{if } y \geq 0 \\ 0, & \text{if } y < 0 \end{cases} \tag{2}$$

I_C represents current pixel intensity, I_n represents adjacent pixel intensity and N is the number of adjacent pixels picked at a distance L. Value of N and L is set as 8 and 1 respectively. Totally 72-D function vectors of which 13 GLCM and 59 LBP features have been extracted. GLCM is used to compute second-order statistical texture characteristics. Extracted characteristics like: Correlation, Homogeneity, Entropy, Variance, Standard deviation, Mean, Kurtosis, Skewness, Smoothness and Inverse Difference Moment normalized (IDM).

2.4 Feature Subset Selection

After feature extraction, feature subsets are derived using SVM weights for different thresholds (0.1 and 0.2), further optimum feature subset are extracted by utilizing the GA-NN algorithm. Individual attribute weights are calculated from the derived

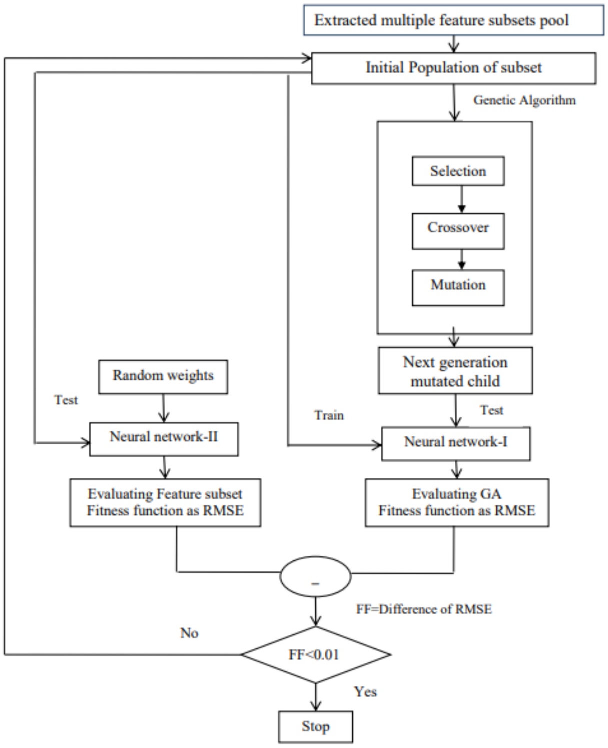
Extracted multiple feature subsets pool
Initial Population of subset
Genetic Algorithm
Selection
Crossover
Mutation
Next generation mutated child
Test
Train
Neural network-I
Random weights
Test
Neural network-II
Evaluating Feature subset Fitness function as RMSE
Evaluating GA Fitness function as RMSE
-
FF=Difference of RMSE
No
FF<0.01
Yes
Stop

Fig 3. Flow chart of Genetic Algorithm-Neural Network

feature set and trained using linear SVM. Figure 3 represents the flow chart of GA-NN algorithm. The algorithm steps are as follows:

Step 1: Population creates an existing generation solutions subset. By utilizing the Roulette Wheel Selection process, parents are selected. Crossover with the Pc = 1, the two-point crossover technique is applied, that means for every iteration it performs a recombination with a mutation likelihood of Pm=0.2 using a Gaussian mutation operator.

Step 2: The selected feature subset is utilized to train the Neural Network-1 (NN-1). Based on Levenberg-marquardt optimization, neural network training mechanism updates the bias and weight. Back propagation algorithm is applied for NN-1 along with the Multi-Layer Perceptron (MLP) algorithm and sigmoid activation function. By utilizing the attributes subset, one hidden layer, five neurons and one output, the neural network has been trained. Fitness function is evaluated through Root Mean Square Error (RMSE) as expressed in equation 3.

$$RMSE = \sqrt{\frac{\sum (x_{ct} - x_{est})^2}{N}} \tag{3}$$

Where x_{ct} is the target class, x_{est} is the predicted output and N is sample number

Step 3: Untrained Neural Network-2 (NN-2) is initialized with irregular bias and weights. By utilizing the subset feature, testing is done. Depending upon the RMSE of the NN-2s output, its fitness function is evaluated.

Step 4: Stopping Criteria: This step aims at minimizing RMSE. If the RMSE difference of two neural networks is less than 0.01, then the iteration stops, otherwise step1 repeats.

2.5 Classification

Two classifiers such as KNN and SVM are used for classifying image patches as cancerous or non-cancerous images. To evaluate the performance, simulation is carried out using optimal feature extraction through GA-NN and SVM weights. These classifiers quantify the performance of the analyzed features with respect to accuracy, sensitivity and specificity[13].

3 Results and Discussion

Around 500 CT lung images are collected from LIDC dataset among them 70% are used for training and remaining 30% are used for testing. GA-NN algorithm and SVM weights are used for optimal feature selection which has yielded fruitful results when compared with existing approachesin detecting the lung cancer. Experimentation is carried out using MATLAB 2020a version.

Reduced features based on two SVM weights are illustrated in the Table 1. With SVM weight equal to 0.1, GLCM features are reduced by 39.5% and with weight equal to 0.2, it is reduced by 61.5%. Similarly, LBP features are reduced by 59.3% and 64.4% with SVM weights equal to 0.1 and 0.2 respectively. Moreover, when both the features LBP and GLCM are combined, features are reduced by 63.8% (SVM weights = 0.1) and 72.2% (SVM weights = 0.2). Since the SVM with weight = 0.2 has resulted in reduced feature size, the performance metrics like sensitivity, specificity and accuracy are evaluated using the results obtained for SVM with weight equal to 0.2 in the further steps.

Table 1. Feature subset based on SVM weights

Various Features	No. Original Features	No. Reduced Features	
		SVM Weights=0.1	SVM Weights=0.2
GLCM	13	8	5
LBP	59	24	21
GLCM and LBP	72	26	20

3.1 Performance Metrics without GA-NN and with GA-NN Feature Subset

The performance metrics such as accuracy, sensitivity, specificity are computed using the equations 4, 5 and 6 respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

Where TP- True Positive, TN- True Negative, FP- False Positive and FN-False Negative.

The performance measures of reduced features using SVM-weights equal to 0.2 are illustrated in Table 2 which reveals that SVM classifier gives better results when compared to KNN classifier.

Table 2. Performance metrics without GA-NN For various features

Feature Set with Different Classifier	Performance Metrics in %		
	Accuracy	Sensitivity	Specificity
KNN_LBP	78.3	73.5	70.6
KNN_GLCM	81.8	75.7	73.4
KNN_GLCM+LBP	83.3	80.9	79.7
SVM _LBP	86.8	84.6	83.1
SVM_GLCM	89.7	87.5	85.3
SVM _GLCM+LBP	92.5	89.6	87.2

Table 3 gives the performance metrics of feature subset after selection of optimum feature using GA-NN filter feature selection technique. This technique helps in increasing the performance measures and reduces the redundancy. Also, it reduces the inappropriate feature selection from fusion feature extraction. Table 3 depicts that, the proposed feature subset selection with GA-NN for SVM classifier has resulted in better performance metrics.

Table 3. Performance metrics of proposed method with GA-NN for various features

Feature Set with Different Classifier	**Performance Metrics in %**		
	Accuracy	**Sensitivity**	**Specificity**
KNN_LBP	82.6	75.7	73.2
KNN_GLCM	84.3	79.5	76.3
KNN_GLCM+LBP (Proposed)	85.3	83.8	81.6
SVM_LBP	90.7	87.3	85.1
SVM_GLCM	93.6	89.2	86.8
SVM_GLCM+LBP (Proposed)	95.8	93.5	91.3

3.2 Performance Comparison with Existing Methods

The proposed GA-NN with SVM method is compared with existing approaches as indicated in Table 4. Different approaches like Weight Optimized Neural Networks with Maximum Likelihood Boosting (WONN-MLB), Convolution Neural Network (CNN), Wilcoxon Signed Generative Deep Learning (WSGDL) and SVM are considered for the comparative analysis. The proposed method has resulted in improved accuracy by 8.8%, 9.8%, 12.47% and 13.8% when compared to the approaches presented in (4,6,10,14) respectively.

Table 4. Performance comparison of proposed approach with existing literature

Approaches	**Accuracy (%)**	**% Improvement**	**[Ref]**
WONN-MLB	82	13.8	(14)
CNN	83.33	12.47	(4)
WSGDL	86	9.8	(10)
SVM	87	8.8	(6)
SVM_GLCM+LBP (Proposed)	95.8	—	—

WONN-MLB: Weight Optimized Neural Networks with Maximum Likelihood Boosting, CNN: Convolution Neural Network, WSGDL: Wilcoxon Signed Generative Deep Learning, SVM: Support Vector Machine

4 Conclusion

The proposed research work uses feature subset extraction using 0.2 SVM weights. Optimal features are selected from feature subset using GA-NN technique to enhance the classification accuracy. This is compared with other standard features like GLCM and LBP. Proposed work gives better classification accuracy viz.95.8% when compared with existing approaches. As a future scope, geometric features can be extracted from Region of Interest (ROI) image to classify the different stages of lung cancer. Also, various preprocessing and segmentation methods can be applied to increase the classification and segmentation accuracy.

References

1) World Cancer Research Fund International. . Available from: https://www.wcrf.org/cancer-trends/worldwide-cancer-data.
2) Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*. 2022;23(1):40–55. Available from: https://doi.org/10.1038/s41580-021-00407-0.
3) Bushara AR, S VKR. Deep Learning-based Lung Cancer Classification of CT Imagesusing Augmented Convolutional Neural Networks. 2022. Available from: https://doi.org/10.5565/rev/elcvia.1490.
4) Pradhan A, Sarma B, Dey BK. Lung Cancer Detection using 3D Convolutional Neural Networks. 2020. Available from: https://doi.org/10.1109/ComPE49325.2020.9200176.
5) Shimazaki A, Ueda D, Choppin A, Yamamoto A, Honjo T, Shimahara Y, et al. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Scientific Reports*. 2022;12(1):727. Available from: https://doi.org/10.1038/s41598-021-04667-w.
6) Manju BR, Athira V, Rajendran A. Efficient multi-level lung cancer prediction model using support vector machine classifier. *IOP Conference Series: Materials Science and Engineering*. 2021;1012(1):012034. Available from: https://doi.org/10.1088/1757-899X/1012/1/012034.
7) Sathya M, Jeyaselvi M, Joshi S, Pandey E, Pareek PK, Jamal SS, et al. Cancer Categorization Using Genetic Algorithm to Identify Biomarker Genes. *Journal of Healthcare Engineering*. 2022;2022:1–12. Available from: https://doi.org/10.1155/2022/5821938.
8) Sharmila N, Arunkumar G, Kumar A, Shivlal B, M, Kumar S, et al. Lung Cancer Classification and Predictionusing Machine Learning an Image Processing. *BioMed Research International Journal*. 2022;2:1–8. Available from: https://doi.org/10.1155/2022/1755460.
9) Tirmzi SAAS, Umar AI, Shirazi SH, Khokhar MAH, Younes I. Modified genetic algorithm for optimal classification of abnormal MRI tissues using hybrid model with discriminative learning approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2022;10(1):14–21.

Available from: https://doi.org/10.1080/21681163.2021.1956371.

10) Obulesu O, Kallam S, Dhiman G, Patan R, Kadiyala R, Raparthi Y, et al. Adaptive Diagnosis of Lung Cancer by Deep Learning Classification Using Wilcoxon Gain and Generator. *Journal of Healthcare Engineering*. 2021;2021:1–13. Available from: https://doi.org/10.1155/2021/5912051.
11) Agnes A, Pandian SIA, Anitha S, Solomon JA. Classification of Lung nodules using Convolutional long short-term Neural Network. 2021. Available from: https://doi.org/10.1109/ICCMC51019.2021.9418319.
12) Jebran M, Gupta PS. Microaneurysm detection by multiple feature subset extraction and selection based on SVM-weights and Genetic Algorithm-Neural Network. *International Conference on Advanced Computing & Communication Systems (ICACCS)*. 2021;p. 129–134. Available from: https://doi.org/10.1109/ICACCS51430.2021.9441746.
13) Angayarkanni D, Jayasimman L. Recognition of Disease in Leaves Using Genetic Algorithm and Neural Network Based Feature Selection. *Indian Journal Of Science And Technology*. 2023;16(19):1444–1452. Available from: https://doi.org/10.17485/IJST/v16i19.218.
14) Alzubi JA, Bharathikannan B, Tanwar S, Manikandan R, Khanna A, Thaventhiran C. Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing*. 2019;80:579–591. Available from: https://doi.org/10.1016/j.asoc.2019.04.031.