Lung Cancer Prediction and Classification Using Adaboost Data Mining Algorithm

P. Thamilselvan

***Abstract*—Currently, cancer has become a common disease which afflicts the survival of people. Lung cancer is one of the cancer types for which early intervention and detection is especially significant. If the lung cancer is detected in the initial stages, it is easy to save the lung cancer patients from the peril of death. In this way, numerous early detection or prediction techniques are being researched and utilized in the battle against the lung cancer. Cancer is the main cause of death for the people who belong to all age groups. Early identification of the lung cancer can be useful in curing the lung disease completely. So the prerequisite of techniques to detect the occurrence of cancer is increasing. Prior analysis of the lung cancer saves huge incidents of lives to avoid other extreme issues abruptly causing deadly ends. Its prediction and cure rate depend basically on the early detection and analysis of the disease. One of the most common types of medical malpractices internationally known is blundering in the pre-determination of the disease. Information discovery and data mining have tracked down various applications in business and scientific areas. Important information can be discovered from the application of data mining techniques in a healthcare framework. In this paper, Adaboost algorithm is proposed and used to predict the lung cancer, and to find the classification accuracy in Computer Tomography (CT) Lung Images.**

***Index Terms*—Adaboost, prediction, classification, CT images, data mining, detection, lung cancer.**

I. Introduction

Cancer is a potentially life-threatening disease across the world. Millions of people are being affected by cancer every day. According to the World Health Organization, 8.2 million people have been affected by this disease. Typically, the human organs such as lung, liver, skin, colon, and bosom of stomach are affected by cancer. Roughly 30% of cancer victims are affected by bad food habits [1]. Lung cancer has been, perhaps, the deadliest disease in the present decades. It has become one of the causes of death in both men and women. There are number of reasons causing lung cancer. Classification and prediction of early intervention of lung cancer is important. This paper to focuses on the various approaches that have been inferred lung cancer detection.

Lung cancer [2] is one of cancers that prompts destruction. The lung diseases are the lethal issues which adversely affect the lungs since the medical conditions are unpredictable, especially in India. Lung cancer constitutes 12.8% of all cancer types around the world. Additionally, it constitutes 17.8% of the cancer passing and it increases by 0.5% consistently around the world. The cause of cancer in people addresses 38.6% for males, and 5.2% for females. Nonetheless, it is proposed that 15% of lung cancer patients live 5 years more. On the other hand, they can survive for a longer time if it is ruled out in the early stages [3], [4]. The two significant kinds of lung cancer [4] are Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). The latter that develops and spreads in different ways causes eventual death. The patients who have manifestations of both the cancer types, are called blended little cell/huge cell cancer. Non-Small Cell Lung Adenocarcinoma (NSCLA) is more normal than SCLC and it commonly develops and spreads more slower than SCLC.

SCLC is connected with smoking elements and becomes develops quicker by shaping a huge cancer that can spread all through the body. Computer Aided Diagnosis (CAD) framework is extremely useful for doctors in detection and diagnosing anomalies prior and quicker [3] and it serves a second opinion and assessment for Medical Doctors (MD) prior to proposing a biopsy test. A few techniques are accessible for lung cancer detection. However, those techniques are costly, tedious and having less capability for detecting the lung tumor [4]. Hence, another prediction strategy is fundamental for predicting the lung cancer in its beginning stages. To diminish the preventable instances of cancer-related death and further develop the general robust framework, a more flexible and stronger innovation is required. Exactly, AI and data mining concepts can process robust data effectively. But the advanced data mining tools cannot be used without a proper training and expertise, which medical specialists don’t have. Therefore, the Adaboost method aims at making those methods accessible in a user-friendly manner. So this research aims at developing to predict lung cancer medical images in CT by using 1 to 10 data mining algorithms [5], namely, Adaboost.

II. Literature Survey

Agarwal *et al.* developed algorithm on ensemble data mining concepts on Seer Data for lung cancer prediction [6]. Zhou and Jiang [7] developed Artificial Neural Network and Decision Trees algorithm for survivability examination of breast cancer. Lundin *et al.* [8] introduced ANN on SEER data to anticipate the breast cancer’s endurance. Delen *et al.* [9] observationally presented three algorithms that are decision tree, neural networks and calculated relapse for foreseeing 60 months-long breast cancer’s endurance. They found that the decision trees show 93.6% precision, which is trailed by neural networks.

Agrawal *et al.* [10] deliberated a clustering process utilized which was not exact in determining the breast cancer using

Manuscript received April 12, 2022; revised July 13, 2022.

P. Thamilselvan is with Dept. of Computer Science, Bishop Heber College (Autonomous) affiliated to Bharathidasan University, India (e-mail: thamilselvan1987@gmail.com).

DOI: 10.7763/IJCTE.2022.V14.1322

different data mining algorithms. This algorithm also adjusted and combined with Naive Bayes, reverse engineered artificial neural networks, C4.5, and decision tree for the breast cancer detection. Neural Network and Decision Trees show 86.7% and 86.5% classification accuracy in detecting the breast cancer.

Punithavathy *et al*. [11] applied neural networks and 3D DenseNet algorithm to identify the lung cancer using Computed Tomography Images. These algorithms are applied in entire lung 3D images.

Tekade *et al*. [12] elaborated two designs; the first, for classification and the second for finding the harm level of cancer. They have utilized straightforward thresholding, clear line, morphology disintegration and morphology shutting in preprocessing stages.

Asuntha *et al*. [13] proposed a profound learning method for distinguishing and classifying the cancerous tissues. Computed Tomography images were used for Histogram Adjustment and classification to improve the accuracy. This proposed method shows 95.62% precision accuracy, 95.89% recall accuracy and 96.23% classification accuracy.

Sasikala *et al*. [14] proposed Convolutional Neural Network-based methodology, which shows 69% sensitivity, 69% specificity, and 87% classification accuracy. The input images were classified into two groups, such as T1-T2 or T3-T4 using CNN.

Liu *et al*. [15] recommended a technique for developing the lung knob identification framework. They have used IoT (Internet of Things) concepts for detecting the lung cancer.

Togacar *et al*. [16] proposed Convolutional Neural Network to identify the lung cancer. They have used more than 100 images in connection with 69 various patients.

Dabeer *et al*. [17] implemented the Convolutional Neural Network algorithm to analyze cancer in histopathological images.

Zhang *et al*. [18] developed a deep learning method to detect cancer in lungs. Serj *et al*. [19] suggested a deep convolutional neural network method for detecting the lung tumor cells. It shows a better classification result compared with the previous CNN based models [20], [21] by Rosetto *et al*. and Dou *et al*.

Masood *et al*. [22] implemented their findings IoT and CNN based algorithms for recognizing the side effects of the lung cancer. Computed Tomography images were used in this research for detecting the cancer.

P. Thamilselvan *et al*. [23] developed algorithms on Advanced Classification and Regression Tree (ACART) for detecting the cancer and classifying benign and malignant cancer tissues and Principal Component Analysis (PCA) algorithm for preprocessing the image.

Chon *et al*. [24] recommended backpropagation neural network to recognize the cancer in its earlier stages. In the pre-processing stage, pixels in the CT images are changed into Hounsfield units for classification.

Alakwaa *et al*. [25] developed 3D based Convolutional Neural Network to detect the lung cancer. Classification, Sampling, Normalization, and Zero Centering were used during the image processing stage.

Teramoto *et al*. [26] implemented Deep-Based convolutional neural network to classify the lung cancer types. Gaussian algorithm and convolutional edges methods are used for systematic individual evaluation of health status of victim of cancer. The research has three layers: associated layers, convolutional layers and pooling layers. This proposed method shows 70% of classification accuracy by using Deep Convolutional Neural Network (DCNN).

III. Proposed Work

In this paper, to evaluate the performance of the Adaboost algorithm, lung cancer images and non-lung cancer images are applied. The dataset contains 100 patients' images which are gathered from different medical centers and hospitals in Tamilnadu. The whole dataset is now grouped as harmless or threatening, based on the opinions of the MDs. Fig. 1 shows an illustration of the test-gathered CT lung image dataset.

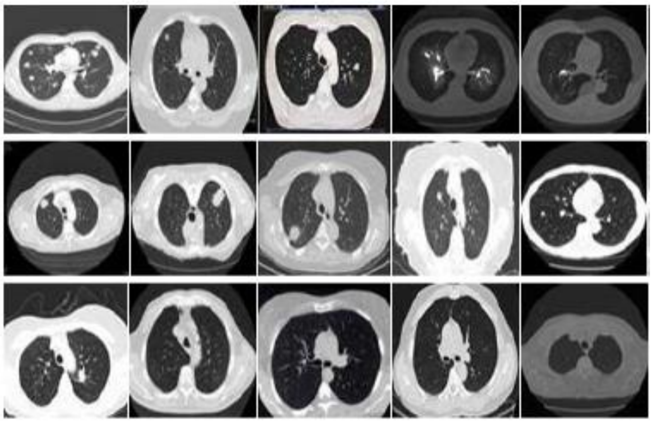


Fig. 1. Sample-collected CT image dataset.

This paper presents a technique for Computed Aided System supported framework to identify cancer tissues in CT images in the preprocessing stage. Profuse Clustering Technique (PCT) has been utilized to eliminate the unwanted noise in the cancer images by Thamilselvan *et al*. [27]. To recognize the cancer and group harmless or threatening images, Adaboost strategy was applied. Fig. 2 shows modules of the proposed framework.

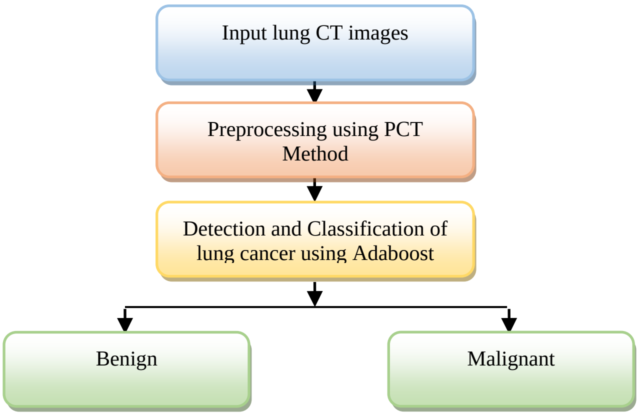

Input lung CT images
Preprocessing using PCT Method
Detection and Classification of lung cancer using Adaboost
Benign
Malignant

Fig. 2. Stages of proposed work.

This structure edifies the difference of the CT images done utilizing preprocessing procedure is finished by Profuse Clustering Technique for noise removal in the image. After the removal of noise it is applied in Adaboost method to detect and classify the cancerous images. The Adaboost method tested using lung CT images to detect and classify the lung cancer images has been implemented by using

Matlab10.

A. Proposed Adaboost Algorithm

AdaBoost likewise called Adaptive Boosting is a strategy in Machine Learning utilized as an Ensemble Method. It is a meta-calculation classifier, which prepares a few feeble classifiers and relegates a similar beginning load to each sample. After each round of preparing, the heaviness of sample will be changed according to a sample mistake rate. Expanding the heaviness of some unacceptable sample to stand out enough to be noticed in ensuring sample. As per the above interaction, k feeble learners are acquired through iterative preparation. At last, one performs a weighted blend to get a solid learner. The working process of Adaboost technique is displayed in Fig. 3.

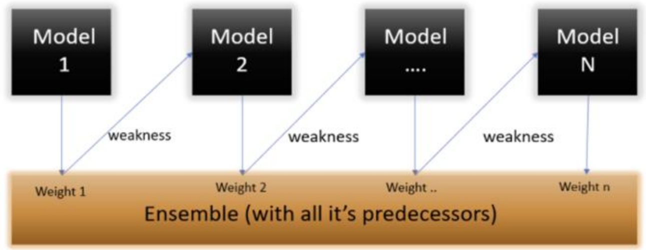


Fig. 3. Working structure of Adaboost.

It settles on 'n' number of choice trees during the data preparing period. As the principal choice tree/model is made, the inaccurately grouped record in the primary model is given need. Just these records are sent as contribution for the subsequent model. The interaction continues until we determine various base model that need to make. Keep in mind, redundancy of records is permitted with all the helping methods.

This Fig. 3 shows how the principal model is made and the mistakes from the main model are noted by the calculation. The record which is mistakenly grouped as a contribution to the following model. This interaction is rehashed until the predefined condition is met. As you can find in the figure, there are 'n' number of models made by taking the mistakes from the past model. This are the means by which supporting can work. The models 1,2, 3…, N are individual models that can be known as choice trees. A wide range of supporting models work on a similar rule.

```
Given($x_1;y_1$),…($x_m,y_m$), $x_i€y_i\sum Y=\{-1,1\}$.
Initialize $D_1(i)=1/m$
For t=1….T:
1. Train the weak classifier using $D_i$
2. Get weak hypothesis $h_t:X\rightarrow\{-1,1\}$error
   $\sum kh_t(x_i)\neq y_iD_t(x_i)$
3. Choose $\dot{\alpha}_t=1/2\log(1-E_t/e_i)$
4. Update
   $D_{t+1}(i) =D_t(i)/Z_t$
5. Output $H(x)= \text{sign}(\sum T_{t=1}\dot{\alpha}_th_t(x))$.
```

Code 1: Pseudocode for boosting algorithm Adaboost

IV. Results and Discussion

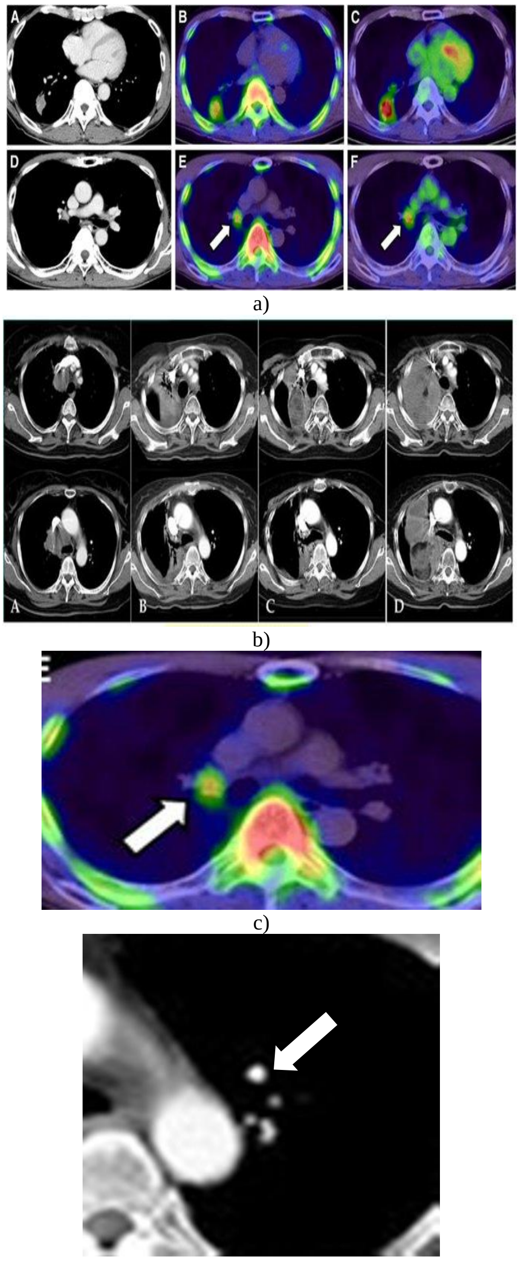
In this work, an improved Adaboost Algorithm has been developed for identifying non-small lung cancer cells and small scale lung cancer cells. The performance of Adaboost Algorithm is tried more than 100 of cancer and non-cancer CT images and it is shown in Fig. 1, Figs. 4.a. to Fig. 4.d. The insightful cycle is accurately performed by following exactness formula.

$$\% \text{ Accuracy} = [(TP+TN) / (TP+TN+FP+FN)] \times 100\%$$

where,

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

In this work, the result of detecting the cancer is magnificently improved using the Adaboost Algorithm. The Fig. 4a) to 4d). shows detected cancerous cells in CT images by using the proposed Adaboost method.



a)

b)

c)

d)

Fig. 4. a) Non-small cell lung cancer; b) Small cell lung cancer; c) Non-small cell lung cancer; d) Small cell lung cancer.

Figs. 4a) to Figs. 4d) show sample detected lung cancers in CT images as small cancer (benign) and non-small (malignant) cancer by using the Adaboost method. The outcome of this work shows that the CT image dataset that have benign or malignant lung that are classified by the adaboost classifier. The results of the Adaboost method show 98.46% efficiency.

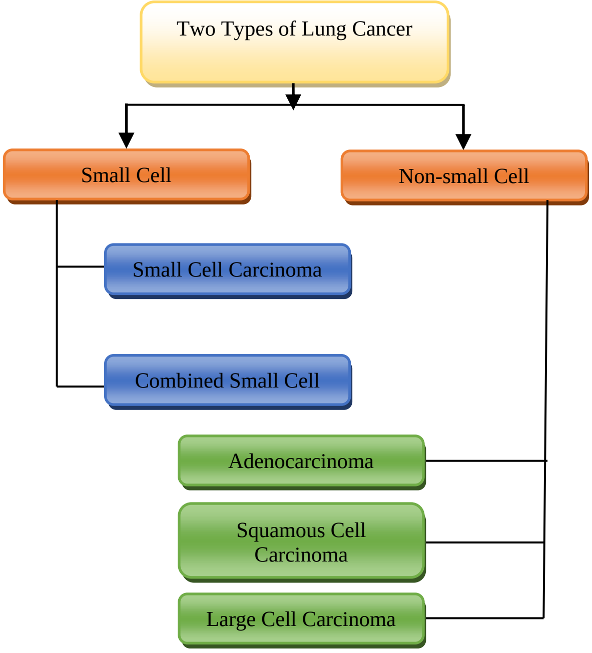


Fig. 5. Types of lung cancer.

Fig. 6 show a result of the proposed Adaboost method. Fig. 6a) shows training dataset, Fig. 6b) shows training data with classification, Fig. 6c) shows classification errors and number of weak classifications, Fig. 6d) shows test data result. In Fig. 6d). the red color cells shows the malignant cells and blue colour cells show the benign cells with classification result.

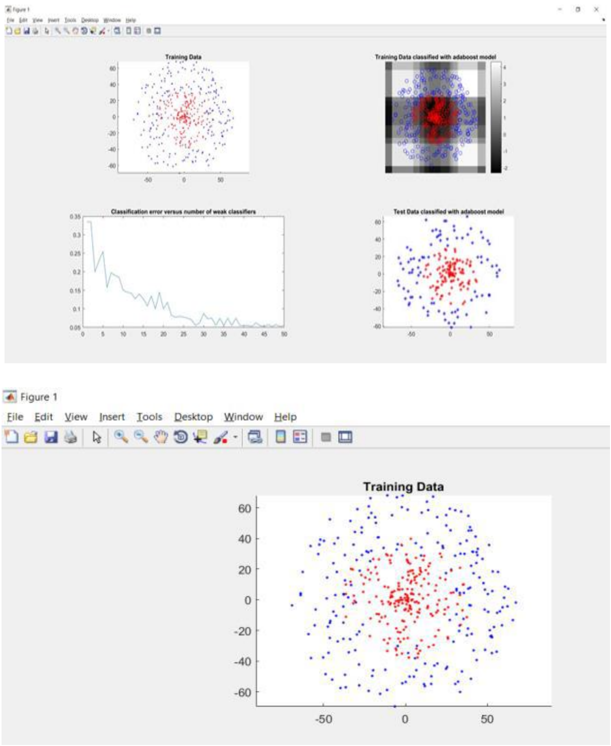
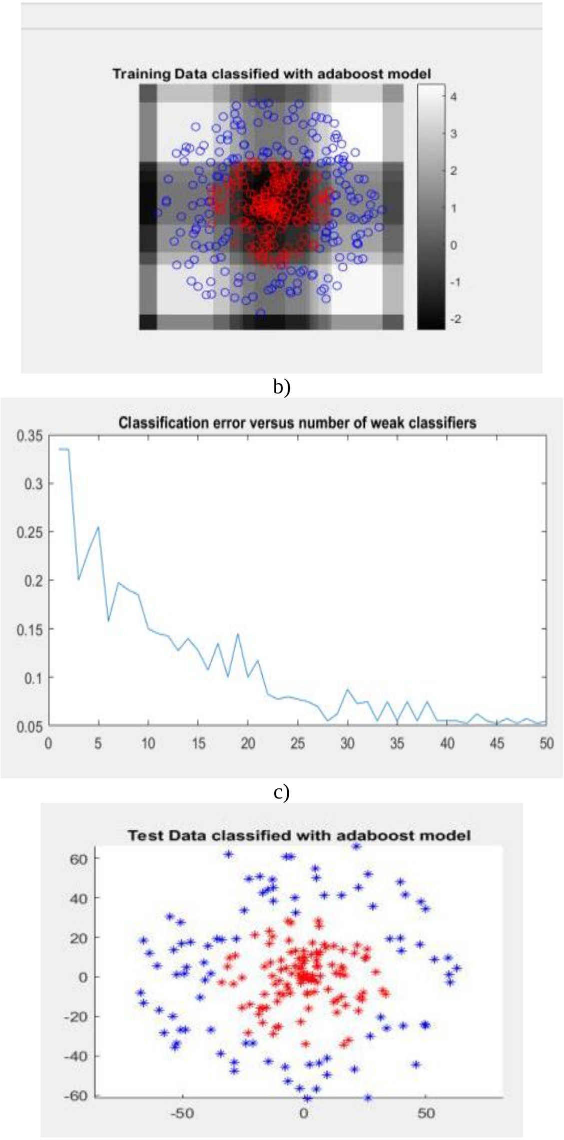


Figure 1
File Edit View Insert Tools Desktop Window Help
Training Data
Training Data classified with adaboost model
Classification error versus number of weak classifiers
Test Data classified with adaboost model
Figure 1
File Edit View Insert Tools Desktop Window Help
Training Data
60
40
20
0
-20
-40
-60
-50
0
50

a)



b)

c)

d)

Fig. 6. Result of Proposed Adaboost method benign and malignant cells. a) training data; b) Training data with classified Adaboost model; c) Classification error versus number of weak classifiers; d) Test data classified with Adaboost model.

A. *Performance Analysis*

In this section, to calculate the performance of proposed Adaboost Algorithm it is compared with different data mining algorithms such as KNN [28], SVM [29], Decision Tree [30], MLPNN (Multilayer Perceptron Neural Network) [31], and CNN (Convolutional Neural Network) [32] as in Fig. 7.

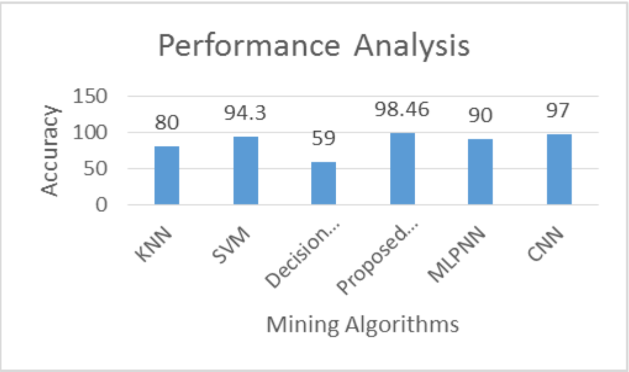


Fig. 7. Performance analysis of proposed Adaboost method.

From Figs. 6 and Fig. 7, Classification Accuracy of Proposed Adaboost method is compared with various mining algorithms to demonstrate a better classification accuracy higher than the rest.

V. Conclusions

In this research, the proposed Adaboost method has been implemented for detecting and classifying CT lung cancer images with benign and malignant classes. Based on the result and performance analysis results, the proposed method is efficient for detecting cancer, and it also classifies the benign (small cell) and malignant (non-small cell) cancer in CT images. The proposed Adaboost method achieves 98.46% classification accuracy, above which it segments to benign cells and malignant cells correctly. The classification accuracy of the proposed Adaboost algorithm shows a better accuracy as well as a low misclassification rate of 1.54%.

This paper has not been concentrated on processing time which means how much time is taken for detecting the lung cancer. Moreover, this proposed algorithm can be further extended for other issues such as brain tumor detection, breast cancer and so on.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Facts and figures of cancer. [Online]. Available: https://www.who.int/cancer/about/facts/es/

[2] K. Saravanan and S. Sasithra, "Review on classification based on artificial neural networks," *International Journal of Ambient Systems and Applications (IJASA)*, vol. 2, no. 4, pp. 11-18.

[3] J. Mutiullah, B. Mehwish, A. Adeel, M. Sabir, and S. Naveed, "Lung cancer detection using digital image processing techniques: A review," *Mehran University Research Journal of Engineering & Technology*, vol. 38, no. 2, pp. 351-360, 2019.

[4] S. R. Shriwas and D. A. Dikondawar, "Lung cancer detection and prediction by using neural network," *International Journal of Electronics & Communication (IIJEC)*, vol. 3, issue 1, pp. 17-21, January 2015.

[5] W. Xindong, V. Kumar *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2008.

[6] A. Agrawal, M. Sanchit, N. Ramanathan, P. Lalith, and C. Alok, "Lung cancer survival prediction using ensemble data mining on seer data," *Scientific Programming*, vol. 20, pp. 29–42, 2012.

[7] Z.-H. Zhou and Y. Jiang, "Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 1, pp. 37–42, 2003.

[8] M. Lundin, J. Lundin, H.B. Burke, S. Toikkanen, L. Pylkkanen, and H. Joensuu, "Artificial neural networks applied to survival prediction in breast cancer," *Oncology*, vol. 57, pp. 281–286, 1999.

[9] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[10] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Poster: A lung cancer mortality risk calculator based on seer data," in *Proc. the IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences*, 2011, pp. 224–233.

[11] K. Punithavathy, M. M. Ramya, and S. Poobal, "Analysis of statistical texture features for automatic lung cancer detection in PET/CT images," in *Proc. Int. Conf. on Robotics, Automation, Control and Embedded Systems (RACE)*, 2015, pp. 1-5.

[12] R. Tekade and K. Rajeswari, "Lung cancer detection and classification using deep learning," in *Proc. Fourth Int. Conf. on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1-5.

[13] A. Asuntha and A. Srinivasan, "Deep learning for lung cancer detection and classification," *Multimedia Tools and Application*s, vol. 79, 2020, pp. 7731-7762.

[14] S. Sasikala, M. Bharathi, and B. R. Sowmiya, "Lung cancer detection and classification using deep CNN," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, issue 2, pp. 2278-3075, 2018.

[15] Z. Liu, C. Yao, H. Yu, and T. Wu, "Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things," *Future Generation Computer Systems*, vol. 97, pp. 1-9, 2019.

[16] M. Toğaçar, B. Ergen, and Z. Comert, "Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 23-39, 2020.

[17] S. Dabeer, M. M. Khan, and S. Islam, "Cancer diagnosis in histopathological image: CNN based approach," *Informatics in Medicine Unlocked*, vol. 16, p. 100231, 2019.

[18] Q. Zhang and X. Kong, "Design of automatic lung nodule detection system based on multi-scene deep learning framework," *IEEE Access*, vol. 8, pp. 90380-90389, 2020.

[19] M. F. Serj, B. Lavi, G. Hoff, and D. P. Valls, "A deep convolutional neural network for lung cancer diagnostic," *Computer Vision and Pattern Recognition*, 2018, pp. 1-10.

[20] A. M. Rossetto and W. Zhou, "Deep learning for categorization of lung cancer CT images," in *Proc. IEEE/ACM Int. Conf. on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2017, pp. 272-273.

[21] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558-1567, 2017.

[22] A. Masood, B. Sheng, P. Li, X. Hou, X. Wei, J. Qin, and D. Feng, "Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images," *Journal of Biomedical Informatics*, vol. 79, pp. 117-128, 2018.

[23] P. Thamilselvan and J. G. R. Sathiaseelan, "Segmentation of lung malignant cancer tissues using PCA and ACART method in MR images for huge," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 6, pp. 1309-1317, 2017.

[24] A. Chon, N. Balachandar, and P. Lu, "Deep convolutional neural networks for lung cancer detection," Ph.D. dissertation, Stanford University, 2017, pp. 1-9.

[25] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 8, pp. 409-417, 2017.

[26] A. Teramoto, T. Sukamoto, Y. Kiriyama, and H. Fujita, "Automated classification of lung cancer types from cytological images using deep convolutional neural networks," *Hindawi BioMed Research International*, pp. 1-6, 2017.

[27] P. Thamilselvan and J. G. R.Sathiaseelan "A novel profuse clustering technique for image denoising," in *Proc. 6th International Conference on Smart Computing and Communications, Procedia Computer Science*, vol. 125, 2018, pp. 132–142.

[28] R. J. Ramteke and Y. K. Monali, "Automatic medical image classification and abnormality detection using k nearest neighbor," *International Journal of Advanced Computer Research*, vol. 2, no. 4, pp. 190-196, 2012.

[29] B. R. Froz, A. O. C. Filho, A. C. Silva, A. Paiva, R. A. Nunes, and M. Gattass, "Lung nodule classification using artificial crawlers, directional texture and support vector machine," *Expert Systems with Applications*, vol. 69, no. 1, pp. 176-188, 2017.

[30] M. Y. Lee and C. S. Yang, "Entropy based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images," *Computer Methods and Programs in Biomedicine*, vol. 100, pp. 269-282, 2010.

[31] C. Bhuvaneswari, P. Aruna, and D. Loganathan, "A new fusion model for classification of the lung disease using genetic algorithm," *Egyptian Informatics Journal*, vol. 15, pp. 69-77, 2014.

[32] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique," *Medical Physics*, vol. 43, pp. 2821-2827, 2016.

P. Thamilselvan is currently working as an assistant professor in Department of Computer Science, Bishop Heber College (Autonomous) affiliated to Bharathidasan University Tiruchirappalli, Tamilnadu, India. He has been working as an assistant professor since 2018. He has published more than 18 research articles in national and international journals which are published in IEEE, Elsevier, and Scopus. His area of research interests includes artificial neural networks, digital image processing, data mining, and image mining.