**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 7.056**

*IJCSMC, Vol. 9, Issue. 10, October 2020, pg.64 – 72*

# Improving the Performance of Lung Cancer Prediction Using Machine Learning Techniques on Big Data

**Dr. M. Kasthuri[1]; M. Riyana Jency[2]**

[1]Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

[2]Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

[1] kasthuri.ca@bhc.edu.in; [2] riyanajency123@gmail.com

**DOI: 10.47760/IJCSMC.2020.v09i10.008**

***Abstract- Medical diligences are a main of developing huge amount of data by itself and form a Big Data Lung Cancer is a disease in which cells in the body grow out of control. When cancer starts in the lungs, it is called lung cancer. From this the prediction of lung cancer is so important to save the patient life. Lung cancer data base stores important as well as raw data into database. For predict the important data from database use different Machine Learning techniques to classify the data. This paper classified about various Machine Learning techniques to predict the lung cancer by using dataset. The accuracy, Precision, Recall and F-Measure are calculated for Support Vector Machine, Navi Bayes, K nearest neighbour, Logistic Regression techniques. The result shows the Support Vector Machine algorithm give the best accuracy of 82.25%. Ensemble method of techniques also used to improve accuracy of lung cancer.***

***Keywords– Lung Cancer, Big data, Machine Learning, Support Vector Machine, K Nearest Neighbour, Navi Bayes, Logistic Regression.***

## I. INTRODUCTION

Lung cancer is one of the major diseases in the health industry that causes several deaths with a series of diseases affecting the circulatory system, non-small cell lung cancer, small cell lung cancer, lung carcinoid tumours, lung failure. In the health industry, lung cancer is one of the main diseases that cause several deaths with a series of diseases involving the circulatory system, non-small cell lung cancer. Lung cancer is carcinogenic cancer that occurs in the lung tissue, usually in air passages line cells. It is both men and women's leading cause of cancer death. Two primary forms exist: small-cell cancer of the lungs and non-small cell cancer of the lungs. Both are different and handled differently. The most mutual system of non-small cell lung cancer.

Lung disease is an unrestrained cell growth in lung tissue, characterised by a malignant lung tumour. Via metastasis, this growth may spread to surrounding tissue or other parts of the body outside the lung. The majority of

*© 2020, IJCSMC All Rights Reserved* 

cancers known as primary lung cancers that begin in the lung are carcinoma. The two major forms are carcinomas in the lung (SCLC) and lung carcinomas in non-small cells (NSCLC). Coughing, weight loss, shortness of breath and chest pains are the maximum common indications, including coughing of the blood. Long-term cigarette smoking is responsible for the vast majority (90%) of lung cancer. In persons who have not ever smoked around 10–15 percent of cases occur. These cases are mostly produced by a grouping of radon gas, asbestos, second-hand smoke and other types of air pollution and genetic factors. Chest x-rays and computed tomography (CT) scans can be used for lung cancer. The diagnosis is confirmed by a biopsy normally done with bronchoscopy or CT.

The main form of prevention is to avoid risk factors, such as smoking and air pollution. The type of cancer, stage (degree of propagation), and overall health of the human being depend on treatment and the long-term outcomes. Most of the cases cannot be healed. Operation, chemotherapy and radiation therapy are standard therapies. NSCLC is mostly treated with operation although SCLC uses chemotherapy and radiation therapy to respond better. Table 1 indicates the factors affecting lung cancer and the death rate.

Table 1. Factors Causing Lung Cancer and Mortality Rate

| Causes | Mortality in% | Figure |
|---|---|---|
| Cigarette smoking | 90% | 70000 (USA) |
| Radon gases | 12% | 21000 |
| Passive smoking | 2.4 % | ------- |

Worldwide, 1.8 million people were diagnosed with lung cancer in 2012 and 1,6 million people died. This makes the cancer-related death of men the most common cause and the second most common cause of death in women after breast cancer. 70 years is the most commonly diagnosed age. In the USA, the survival rate for the next five years is 19.4%, in Japan 41.4%. In the developed world, on average, the outcomes are worse.

## II. RELATED WORK

Shanti and raj Kumar [2] used wrapper feature selection method as well as stochastic diffusion research algorithm on lung cancer image and concluded that this is one of the best performing algorithms for classification. They proposed to use the modified stochastic diffusion (SDS) algorithm, a novel feature selection algorithm based on the wrapper. In order to define optimum subsets of features the SDS benefits from direct contact by agents. For the classification, the neural network, Naïve Bayes and decision-tab were used. The experiment 's findings show that the proposed approach is capable of achieving better efficiency than previous approaches, such as maximum relevance for minimal redundancy and correlation-based selection.

Monkam et al. [3] conducted a survey with almost 90 percent accuracy on the importance of Convolutional Neural Network in prediction of lung module. Present an exhaustive study and results of these approaches. Second, we present the basic information and the explanations why CNN is ideal for analysing medical pictures. A brief overview of the different medical image datasets and environmental configurations that make lung nodule study with CNNs easier is then provided. In addition, detailed overviews are given on recent progress in the study of pulmonary nodules using CNNs. Finally, current problems and promising directions are addressed in particular for the further development of the application of CNN to the study of medical images and the pulmonary nodule evaluation. The early diagnosis and management of lung cancer is shown to be greatly transformed by CNNs. They trust that this evaluation will deliver all the medical research societies with the necessary knowledge to master the concept of CNN so as to utilize it for improving the overall human healthcare system.

Hussein et al. [4] proposed supervised learning using 3D Convolutional neural network (3D CNN) on lung nodules data set as well as unsupervised learning SVM approach to classify benign and malignant data with a accuracy of 91%.

Ganggayah et al. [5] used various classifiers on breast cancer data having 8066 record with 23 predictor and concluded that random forest classifier gives 82% better accuracy.

Gibbons et al. [6] used supervised learning such as linear regression model, support vector machine, ANN etc. and predicted that SVM results a better accuracy of 96% as compared to other methods.

*© 2020, IJCSMC All Rights Reserved* 

Shakeel et al. [7] used feature selection process and a novel hybrid approach of ANN on lung cancer data available from ELVIRA biomedical data to predict an accuracy of 99.6%.

Dr. M. Kasthuri [8] various forms of noise applied to a image of various pixel block size to enhance the quality of image. The quantitative performance of Peak Signal Noise Ratio (PSNR), Root Mean Square Error (RMSE), Universal Image Quality Index (UIQI), and Enhanced Measure of Enhancement (EME) are used as the evaluation index. The results obtained were evaluated.

Dr. M.Kasthuri [9] various forms of noise applied to various pixel block size of image perform the de-noising testing do to enhance the quality of the image. The performance measure Mean Absolute Error (MAE) and Pearson correlation are used as the evaluate work. The results were investigated with benchmark cameraman image.

## III. DATASET DESCRIPTION

Dataset was available in UCI machine learning repository. Data consists of 32 instances and it has 57 features (1 class attribute and 56 input data), all predictive attributes are nominal range between 0–3 while class attribute level of 3 types [1]. Nominal attribute and class label data are converted in to binary form such that data analysis process becomes easier. Nominal to binary form is the most standardization process for data analysis. Data set comprises of some missing values which degrades the algorithm performances so care full execution before analysis on data is required. Label is described as high, low, medium. In the paper we categorized high to 2, medium to 1 and low to 0.

## IV. CLASSIFICATION TECHNIQUES

Classification comes under supervised learning process in order to predict given input data to a certain class label. The novelty in classification relies on mapping input function to a certain output level. Various learning classifiers are described as Perceptron, Naïve Bayes, Decision Tree, Logistic Regression, K nearest neighbour, Artificial Network, Support Vector Machine. Classification in machine learning is one of prior decision-making techniques used for data analysis. Various classifier techniques are too used to classify data samples. The concept of our paper focuses on novel approach of Machine Learning for analysis of lung cancer data set to achieve a good accuracy. Some of the mostly used classifier techniques are described as.

### *A) Support Vector Machine (SVM)*

One of the simple and useful approaches in supervised learning is support vector classification. Support vector Machine is usually preferred for data analysis because of its computational capability with in very less time frame. This classifier works on the decision boundary concept Recognized as hyper plane. The hyper plane is used to classify the input data in to required target group. But in order to fit the decision boundary in a plane maximize distance margin is chosen from data points for classification. User defined support vector classifier can be framed using various kernel function to improve the accuracy. Support vector classifier is well suited for both structured and unstructured data. Support vector classifier is not affected with over fitting problem and makes it more reliable shown in Fig 1.

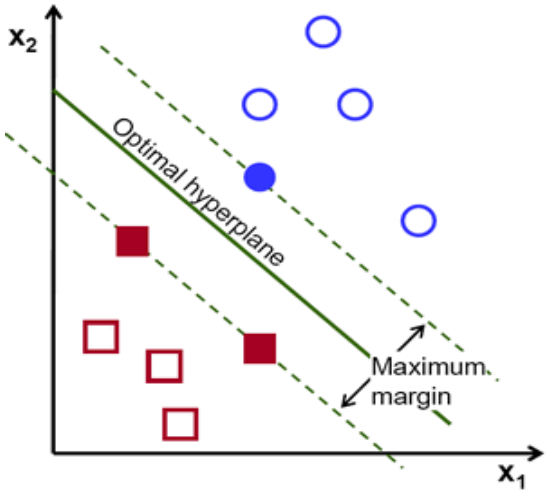


Fig 1. SVM Classifier

© 2020, IJCSMC All Rights Reserved 

***B) Logistic Regression Classifier***

Logistic Regression classifier is brought from statistics. These classifiers are based on the probability of outcome from the input process data. Binary logistic regression is generally preferred in machine learning technique for dealing with binary input variables shown in Fig 2. To categorize the class in to specific category sigmoid function is utilized. Advantages of Logistic Regression classifier.

- Logistic regression classifier is very flexible to implement
- Suitable for binary classification
- Depend on probabilistic model

Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-2}}$$

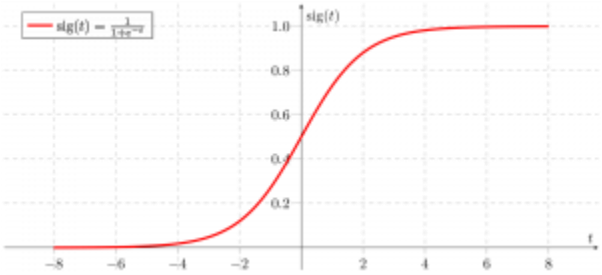


Fig 2 Logistic Regression Classifier

***c) K Nearest Neighbour Classifier***

KNN classifier comes under lazy learning process in which training and testing can be realized on same data or as per the programmer's choice shown in fig 3. In the process, the data of interest is retrieved and analysed depending upon the majority value of class label assigned as per k, where k is an integer. The value of k is based on distance calculation process. The choice of k depends on data. Larger value of k minimizes the noise on classification. Similarly, Parameter selection is also a prominent technique to improve the accuracy in classification. Weighted KNN classifier: A mechanism in which a suitable weight can be assigned to the neighbour's value so that its contribution has great impact to neighbours than distant ones. In the weighted KNN approach the weight has a significant value in evaluating the nearest optimistic value. Generally, the weight is based on reciprocal of distance approach. The weight value of attribute is multiplied with distance to obtain the required value.

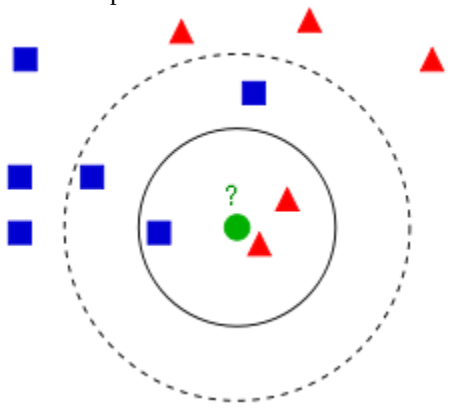


Fig 3 KNN Classifier

Pseudo code for KNN

- Take the input data
- Consider initial value of k
- Divide the train and test data
- For achieving required target iteration for all training data points
- Find the distance between test data and each row of training data. (Euclidean Distance is the best Choice approach)

*© 2020, IJCSMC All Rights Reserved* 

- Arrange the calculated distance in ascending order based on distance values.
- Consider the Top k value from sorted value.
- Find the Majority class label
- Obtain the target class.

***d) Naïve Bayes Classifier***

Naive Bayes classifier is one of the probabilistic classifiers with strong independent assumption between features. Naive Bayes is based on bayes Theorem where Naïve Bayes classifier uses Bayesian network model p using the maximum a posteriori decision rule in Bayesian Setting. The feature which are classified in naive Bayes are always independent to each other. If y is class variable and x is dependent feature vector then.

Bayesian probability says = $Posterior = \frac{Prior * Likelihood}{Evidence}$

## V. PROPOSED MODEL

The knowledge prediction tool called 'Java Classification tool' is developed as software in Eclipse IDE Release2 (4.4.2) with an inbuilt machine learning algorithms and feature selection algorithm. The tool is designed to be generic to accept any datasets that needs to perform classification analysis. This experiment was conducted on a machine operating system windows 10 Home 64 bit Build 18737, the processor was an Intel® Core (TM) i3 10th Generation CPU @ 3,62GHz. The memory of Machine was 6GB RAM. The graphics card was NAVIDIA 6GB.

***a) Result Analysis***

The Input data consists of missing values. So, it is required to prepossess the data such that the missing values have been replaced with the most occurrence value of the corresponding column. Then the processed data is applied in Java Classification tool for analysis. The prepossessed data is converted in to suitable form for classification using different classifier approach. With the classify tab of Java Classification tool different classifier approaches are verified. After careful analysis results of proposed classifiers are compared. Logistic Regression and Naive Bayes algorithm classifies 32 instances in to 25 correctly classified instances and 7 incorrectly classified instances. Likewise, 24 correctly classified instances and 8 incorrectly classified instances are obtained from 32 instances using KNN with 5 nearest neighbours. As per our analysis the SVM classifier is mostly preferred among various classifiers. This is due to its highest classification accuracy which is obtained from its 26 correctly classified instances and 6 incorrectly classified instances from 32 instances. The work process flow of various classification in java tool as shown in fig 4.

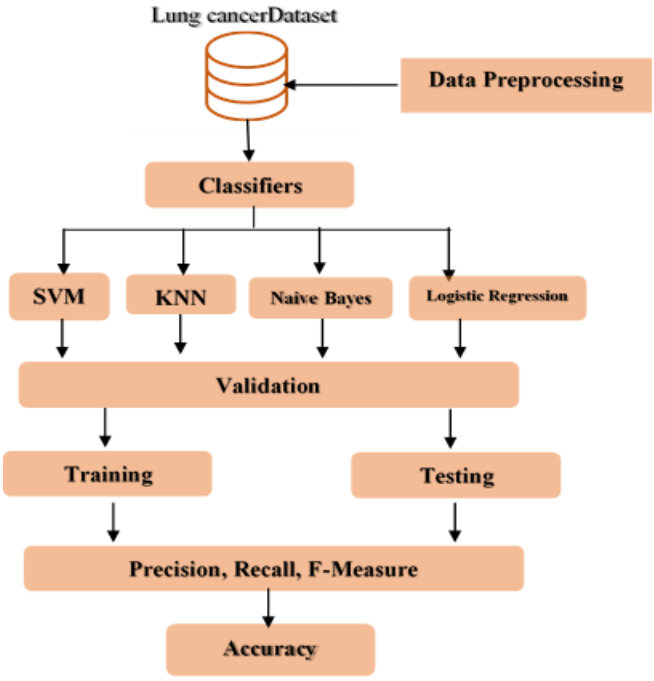


Fig. 4. Work process flow of various classifiers in Java tool

© 2020, IJCSMC All Rights Reserved 

The results of different classifications of lung cancer data used in the Java classification tool are showed in the following table. In general, accuracy, remembering, precision and F-measuring are the main parameters of the classification process in the confusion matrix. Precision of classification is the calculation of the number of accurate predictions from total predictions. These parameters are calculated by a certain result. Which are 'TP' (True Positive) and 'TN' (True Negative) is not correctly predicted for any event values. Likewise, for the inaccurate prediction of no event value, 'FP' is wrongly predicted event value and 'FN' (False negative)

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

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$FMeasure = \frac{2 * recall * Pecisopn}{recall + precision}$$

Table II. Classifiers output in Java Classification tool

| Algorithm | Precision | Recall | F-Measure | ROC Area | Correctly classified | Incorrectly classified |
|---|---|---|---|---|---|---|
| KNN | 0.74 | 0.74 | 0.71 | 0.70 | 76% | 24% |
| Naive Bayes | 0.784 | 0.79 | 0.78 | 0.78 | 79.125% | 20.87% |
| SVM | 0.824 | 0.824 | 0.824 | 0.759 | 82.25% | 17.75% |
| Logistic Regression | 0.778 | 0.791 | 0.776 | 0.718 | 79.12% | 20.87% |

Fig.5. specifies the accuracy analysis of the Machine Learning techniques with Four Machine Learning algorithms as the motivation of the work is to identify the best features that increase the accurate prediction of Lung cancer Dataset. Among the Four experimental algorithms, SVM improves the accuracy of other in Lung cancer dataset prediction with 82%.

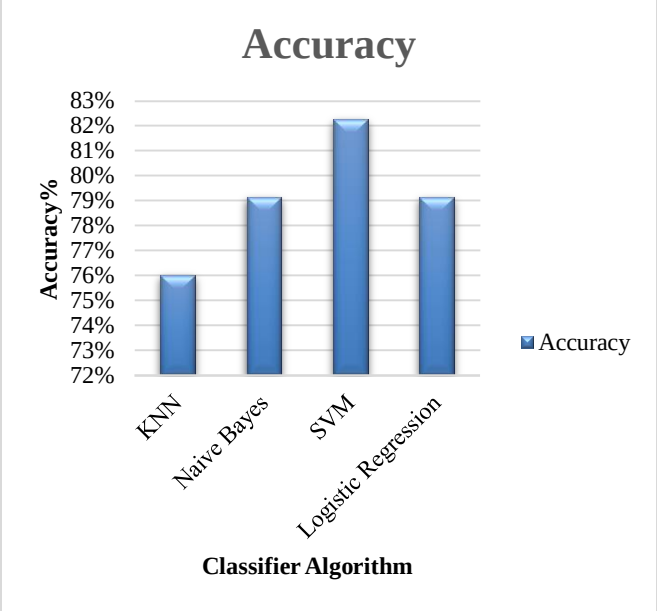


Fig 5. Accuracy Analysis of Machine Learning algorithms

Fig.6. depicts the precision value analysis of experimental algorithms for Lung cancer dataset. SVM algorithm has the highest precision value than other algorithms.

*© 2020, IJCSMC All Rights Reserved* 

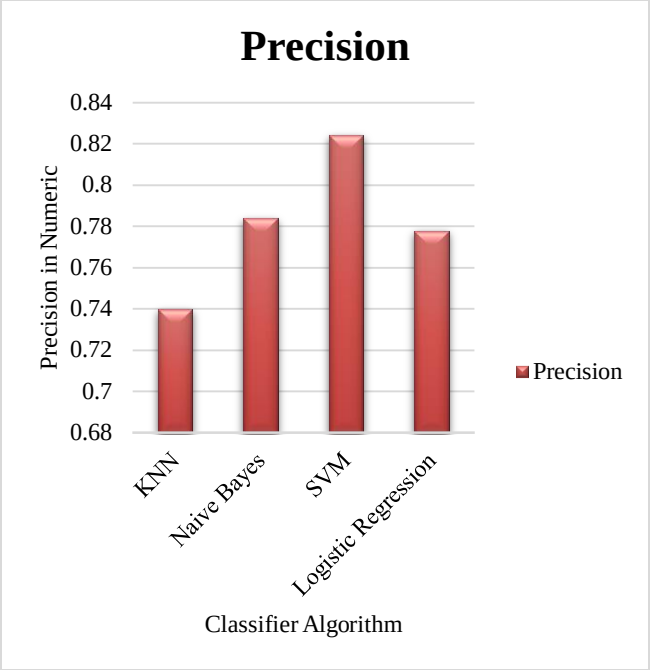


Fig.6. Precision Measure.

Recall is a measure that results how many relevant Lung cancer are selected. Recall measure analysis shown in Fig.7. Proves that the algorithm SVM outperforms than the other algorithms

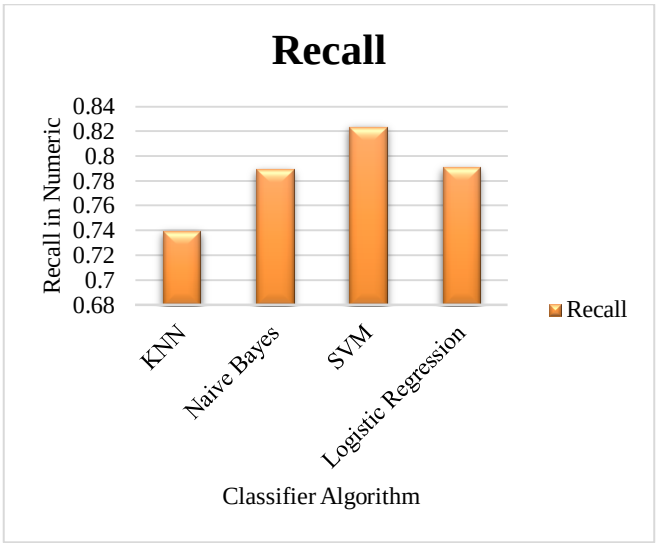


Fig 7. Recall Analysis.

F-Measure analysis of proposed algorithms is shown in Fig.8. Like other results, the F-Measure value of SVM is high. The results denote that the algorithm outperforms the experimental algorithms in terms of quality and quantity.

*© 2020, IJCSMC All Rights Reserved* 

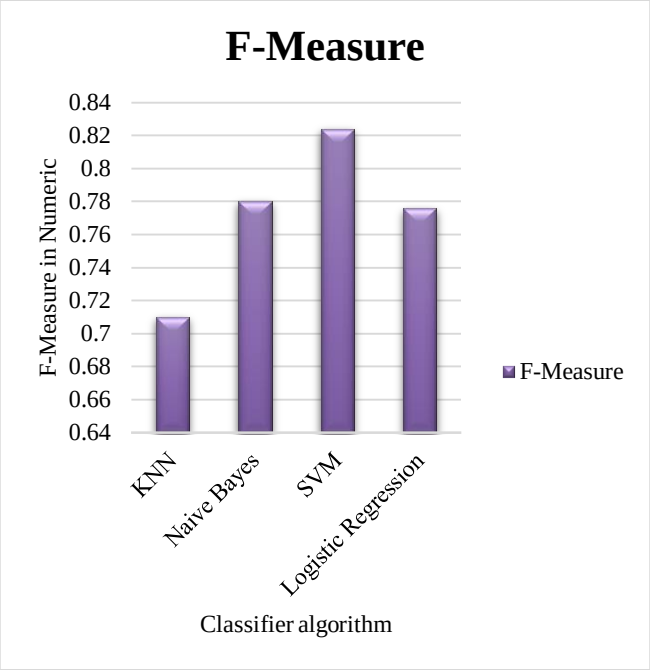


Fig 8. F-Measure Analysis

## VI. CONCLUSION

The detection and classification of lung cancer can greatly increase the early stage detection of lung cancer and improves patient survival. The designing and analysing four Machine Learning algorithms for the lung cancer classification of without computing texture features and the morphology. In this paper we have shown that with SVM classifier the accuracy is found to be 82.25% on lung cancer data. So, in the analysis it can be predicted that with suitable feature selection method and integrated approach with other supervised learning process and modified functional approach in SVM, accuracy will be further improved.

# References

[1] https://archive.ics.uci.edu/ml/dataset/Lung+cancer. Accessed 07 Sep 2020.

[2] Shanthi, S., and N. Rajkumar. "Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods." Neural Processing Letters Volume 8, Issue 10, pp. 1-14 Springer, 2020.

[3] Monkam, Patrice, Shouliang Qi, He Ma, Weiming Gao, Yudong Yao, and Wei Qian. "Detection and classification of pulmonary nodules using convolutional neural networks: A survey." IEEE Access 7, Vloume 7, Issue 10, pp. 78075 – 78091, ISSN: 2169-3536, 2019.

[4] Hussein, Sarfaraz, Pujan Kandel, Candice W. Bolan, Michael B. Wallace, and Ulas Bagci. "Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches." IEEE transactions on medical imaging, Volume: 38, Issue: 8, pp. 77-1787, ISSN: 0278-0062, 2019.

[5] Ganggayah, Mogana Darshini, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Sarinder Kaur Dhillon. "Predicting factors for survival of breast cancer patients using machine learning techniques." BMC medical informatics and decision-making Volume 19, Issue 1, pp. 19:48, 2019.

[6] Sidey-Gibbons, Jenni AM, and Chris J. Sidey-Gibbons. "Machine learning in medicine: a practical introduction." BMC medical research methodology Volume19, Issue. 1, pp. 19:64, ISSN: 2874-0681,2019.

[7] Shakeel, P. Mohamed, Amr Tolba, Zafer Al-Makhadmeh, and Mustafa Musa Jaber. "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks." Neural Computing and Applications Volume 32, Issue 3 pp. 777-790, Springer, 2020.

*© 2020, IJCSMC All Rights Reserved* 

[8] M.Kasthuri, "Mean Filtering to De-Noising Image Using Various Block Size", Internat ional Journal of Advanced Scient ific Research and Management, Volume 4 Issue 6, pp. 43 – 47, ISSN 2455-6378, June 2019.
[9] M.Kasthuri, "Image of De-Noising Using Linear Mean Filter For Various Block Size ", International Journal of Scientific & Technology Research, Volume 8, Issue 10, pp. 2606 – 2608, ISSN 2277-8616, October 2019.
[10] Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." Multimedia Tools and Applications, Volume 79, Issue 4, pp. 1-32, ISSN: 7731–7762,2020.
[11] Chaubey, Nirbhay Kumar, and Prisilla Jayanthi. "Disease Diagnosis and Treatment Using Deep Learning Algorithms for the Healthcare System." In Applications of Deep Learning and Big IoT on Personalized Healthcare Services, Volume 10, Issue 9, pp. 99-114. ISSN: 4018-7998, 2020.
[12] Pujan Kandel, Candice W. Bolan, Sarfaraz, Michael B. Wallace, and Ulas Bagci. "Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches." IEEE transactions on medical imaging Volume: 38, Issue: 8, pp. 77-1787, ISSN: 0278-0062, 2019.
[13] Ganggayah, Mogana Darshini, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Sarinder Kaur Dhillon. "Predicting factors for survival of breast cancer patients using machine learning techniques." BMC medical informatics and decision making, Volume 19, Issue 1, pp.19-48, ISSN: 1186-0801, 2019.
[14] Jakimovski, Goran, and Danco Davcev. "Using double convolution neural network for lung cancer stage detection." Applied Sciences, Volume 9, Issue 3, pp. 142, 2019.
[15] Li, Xin, Bin Hu, Hui Li, and Bin You. "Application of artificial intelligence in the diagnosis of multiple primary lung cancer." Thoracic Cancer, Volume 10, Issue 11, pp. 2168-2174, ISSN: 2658-3589, 2019.

*© 2020, IJCSMC All Rights Reserved* 