Prediction of Lung Cancer Using Machine Learning Classifier

Radhanath Patra(✉)

Electronics Science, Berhampur University, Berhampur, Odisha, India
radhanath.patra@gmail.com

Abstract. Lung cancer generally occurs in both male and female due to uncontrollable growth of cells in the lungs. This causes a serious breathing problem in both inhale and exhale part of chest. Cigarette smoking and passive smoking are the principal contributor for the cause of lung cancer as per world health organization. The mortality rate due to lung cancer is increasing day by day in youths as well as in old persons as compared to other cancers. Even though the availability of high tech Medical facility for careful diagnosis and effective medical treatment, the mortality rate is not yet controlled up to a good extent. Therefore it is highly necessary to take early precautions at the initial stage such that it's symptoms and effect can be found at early stage for better diagnosis. Machine learning now days has a great influence to health care sector because of its high computational capability for early prediction of the diseases with accurate data analysis. In our paper we have analyzed various machine learning classifiers techniques to classify available lung cancer data in UCI machine learning repository in to benign and malignant. The input data is prepossessed and converted in to binary form followed by use of some well known classifier technique in Weka tool to classify the data set in to cancerous and non cancerous. The comparison technique reveals that the proposed RBF classifier has resulted with a great accuracy of 81.25% and considered as the effective classifier technique for Lung cancer data prediction.

Keywords: KNN · ML · RBF · Lung cancer · ANN

1 Introduction

Lung cancer considers as the deadlier disease and a primary concern of high mortality in present world. Lung cancer affects human being at a greater extent and as per prediction it now takes 7th position in mortality rate index causing 1.5% of total mortality rate of the world [2]. Lung cancer originates from lung and spreads up to brain and spreads Lung cancer is categorized in to two major group. One is non-small cell lung cancer and another is small cell lung cancer. Some of the symptoms which are associated with the patients like severe chest pain, dry cough, breathlessness, weight loss etc. Looking in to the cultivation of cancer and its causes doctors give stress more on smoking and second-hand smoking as if the primary causes of lung cancer. Treatment of lung cancer involves surgery, chemotherapy, radiation therapy, Immune therapy etc. In-spite of this

N. Chaubey et al. (Eds.): COMS2 2020, CCIS 1235, pp. 132–142, 2020.
https://doi.org/10.1007/978-981-15-6648-6_11

lung cancer diagnosis process is very weak because doctor will able to know the disease only at the advanced stage [18]. Therefore early prediction before final stage is highly important so that the mortality rate can be easily prevented with effective control. Even after the proper medication and diagnosis survival rate of lung cancer is very promising. Survival rate of lung cancer differs from person to person. It depends on age, sex and race as well as health condition. Machine learning now days plays a crucial role for detection and prediction of medical diseases at early stages of safe human life. Machine Learning makes diagnosis process easier and deterministic. Machine learning now a day's have already dominated medical field. Every county is now adopting machine learning techniques in their health care sector. With the application of machine learning the actual detection of diseases can be explored. Some of the crucial application of machine learning is described as Feature Extraction: In any disease attributes are the real information container of the diseases. Machine learning (ML) helps for easy of data analysis and process the real attributes or information and finds the actual problem creator of diseases. It helps medical expert to find the root cause of diseases. Image processing: Using various process of machine learning the image analysis has been found accurate and valuable. That helps the concerned doctors to have a better diagnosis of the diseases such that money and time can be saved and value proportion can also be increased. Drug manufacturing: Depending up on the increase of various diseases, drug should be multi functional and quantity should be known. So ML has solved the problem and helps the drug industry to use of ML application for manufacturing. Better Prediction of diseases: ML helps to predict the severity of diseases and its outcome. ML controls disease outbreak through early prediction such that appropriate measures can be taken. Still machine learning application needs to be refined such that it can be more standardized and more reliable. Thus the need of more improvement in machine learning algorithm would help the physicians, health catalyst for accurate clinical decision making with high efficiency as well as good accuracy.

Machine learning makes the system to find the solution of problem with own learning strategies. ML classifies in to three categories such as unsupervised learning, supervised learning, Reinforcement learning. Supervised learning identifies two processes under its umbrella, one is classification and another is regression. Classification is process in which input data is processed and categorized in to certain group. The proposed work was carried out in Weka tool. Algorithm like j48, KNN, Naive Bias and RBF are used in Weka tool and a comparative analysis was derived finally.

2 Related Work

Z. Zubi et al. (2014) extracted features from chest x ray images and used concept of back propagation neural network method to improve the accuracy [31]. Rashmee Kohad et al. (2015) used ant colony optimization with ANN and SVM to predict the accuracy of 98% and 93.2% respectively on 250 lung cancer CT images [16]. Kourou et al. (2015) outlined a review of various machine learning approach on several cancer data and concluded that application of integration of feature selection and classifier will provide a promising result in analysis of cancer data [17]. Hosseinzadeh et al. (2013). Proposed SVM model on selection of protein attributes and concluded that the result is having 88% accuracy in compared to other classifier technique for prediction of lung cancer tumors [11].

Naveen and Pradeep (2018) proposed that among SVM, Naive Bayes and C4.5 classifier, C4.5 performs better on North central cancer treatment group (NCCTG) lung cancer data with better accuracy and also predicted that C4.5 is better classifier with the increase of lung cancer training data [25]. Gur Amrit Pal singh and P.K Gupta (2018) proposed new algorithm for feature extraction on image data and applied machine learning classifier to improve the accuracy [29]. Hussein et al. (2019) 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% [12]. Monkam et al. (2019) provided survey on importance of Convolutional neural network for predicting lung module with almost greater than 90% accuracy [21]. Asuntha and Andy Srinivasan (2019) proposed fuzzy particle swarm optimization with deep neural network on lung cancer images to achieve an accuracy of 99.2% [5]. Ganggayah et al. (2019) used various classifiers on breast cancer data having 8066 record with 23 predictor and concluded that random forest classifier gives 82% better accuracy [9]. Gibbons et al. (2019) used supervised learning such as linear regression model, support vector machine, ANN etc. and predicted that SVM results an better accuracy of 96% as compared to other methods [28]. Shakeel et al. (2019) 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% [26]. Bhuvaneswari et al. (2015) used gabor filter for feature extraction and G-knn approach to classify lung cancer images with an accuracy of 90% [7]. Xin Li, Bin Hu, Hui Li Bin (2019) used 3D dense sharp network and IBM SPSS25.0 statistical analysis software on 53 patients to obtain an accuracy of 88% in finding malignant and benign [19]. Shanti and raj Kumar (2020) 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 algorithm for classification [27]. Rezaei Hachesu P et al. (2017) proposed a different approach for analysis of survival rate and the method find a correlation between various attributes and their survival rate and this process is carried out with 470 records having 17 features [10]. Kadir et al. (2018) provided an overview approach of various deep learning strategies used for accuracy prediction of lung cancer CT images [15]. Paing et al. (2019) used computer aided diagnosis process in which in three phases segmentation, detecting and staging process are followed for classification of CT lung cancer images with a greater accuracy [23].

3 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.

4 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 [20, 22]. 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.

4.1 Neural Network

Neural network are the basic block of machine learning approach in which the learning process is carried in between neuron. Artificial neural network (ANN) comprises of input layer, intermediate layer having hidden neurons and output layer. Every input neuron is connected to hidden neuron through appropriate weight and similarly weight is connected between hidden unit to output unit. Neuron presented in hidden neuron and output neuron are processed with some known threshold functional value. Depending on the requirement the activation will be used to process the neuron. The synaptic weight gets multiplied with the corresponding neuron presented in hidden layer and output layer for classification process. The desired target is adjusted through the weight adjustment technique either in feed forward approach or feed back approach to get the required target. Feed forward network approaches are simpler process for classification approaches.

4.2 Radial Basis Function Network

Radial basis function network comes under neural network that uses radial basis function as its threshold function.RBF network has advantage of easy of design and strong tolerance to input noises. Radial basis Function is characterized by feed forward architecture

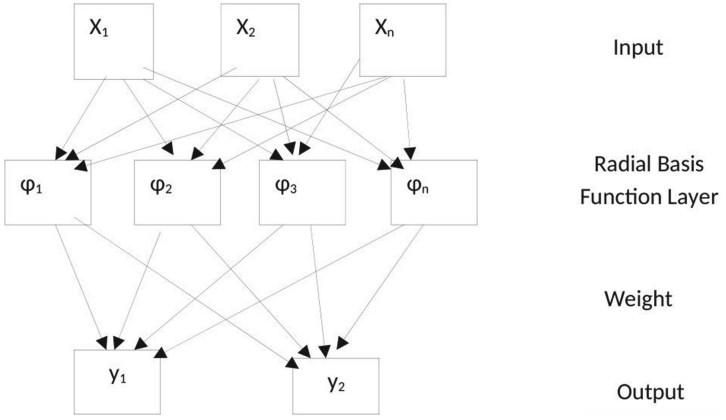

Fig. 1. RBF.

which comprises of an one middle layer between input and output layer. It uses a series of basis function that are centered at each sampling point. Formally for a given input x the network output can be written as (Fig. 1).

Where

$$y = \prod_{i=1}^{N} w_i R_i(x) + w_0 \tag{1}$$

wi: *weight, w*0: *biasterm, R*: *Activation function*

$$R_i(x) = \varphi[\|x_i - c_i\|] \tag{2}$$

φ: *radial function, ci*: *RBF centre*

In RBF architecture the weight that connects to input unit and middle layer represents the centre of the corresponding neuron where as weights connecting to middle layer and output layer are used to train the network.

4.3 Support Vector Classifter

One of the simple and useful approaches in supervised learning is support vector classification. Support vector classifier (SVC) 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.

4.4 Logistic Regression Classifter

Logistic Regression classifier is brought from statistics. These classifiers is 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. 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

4.5 Random Forest Classifter

Combination of classifier trees represents random forest classifier. One of the finest approaches to represent input variables in form of trees that makes a forest like structure. Input Data are represented in trees and each tree specifies a class label. Random

forest depends on its error rate. Error rate signifies in to two directions. First one is the correlation between trees and second one is the strength of the tree. Advantages of random forest.

- Proper method for noisy and Imbalanced data representation.
- Data can be represented without any data reduction.
- Best approach for analysis of large data set.
- Finest approach in machine learning platform for improvement of accuracy.
- It handles the over fitting problem which mostly occurs in different Machine learning algorithm.
- one of the best reliable algorithm

4.6 J48 Classifter

J48 is representation of c4.5 in weka tool developed from java. Decision tree implements tree concepts to find the solution of the problem. class label is represented by leaf node where as attributed are defined with internal node of tree. In decision tree attribute selection process is done by Information gain and gain index. Depending on the concept of information gain and depending on the importance of information gain Decision tree classifier performs the classification. The information gain for a particular attribute X at a node is calculated as

$$\text{Information Gain (N, X)} = \text{Entropy (N)} - \sum_{value\ at\ x} \frac{|N|}{|N_i|} Entropy(N) \tag{3}$$

Where N is the set of instance at that particular node and

$$|N| : cardinality$$

Entropy of N is found as:

$$Entropy\,(N) = \sum_{i=1}^{N} -p_i\, log_2\, p_i \tag{4}$$

4.7 Naïve Bayes Classifter

Naive Bayes classifier is one of the probabilistic classifier with strong indepenent assumption between features. Naive Bayes is based on bayes Theorem where Naive 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.

$$y = argmax_y\, p(y) \prod_{i=1}^{n} p\left(\frac{x_i}{y}\right) \tag{5}$$

P (y) is called class probability and

$$p\left(\frac{x_i}{y}\right) \tag{6}$$

is conditional probability. Bayesian probability says

$$Posterior = \frac{Prior * Likelihood}{Evidence}$$

4.8 Knn Classifter

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. In the process, the data of interest is retrieved and analyzed 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 neighbor's value so that its contribution has great impact to neighbors 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.

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)
- 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.

5 Proposed Model

Data analysis process was carried using both weka tool of version 3.6 and Jupiter platform in python tool [13, 24]. Weka is an open source tool used for classification, clustering, regression and data visualization. Weka generally supports input file either in .csv or .arff extension format. Weka explore has various tabs for data analysis such as prepossess, classify, cluster, association, select attribute and visualize. When data prepossessing is selected it enables to upload the input data in weka tool [3, 30]. Weka tool clearly understands and represents the data for easy of data analysis. Before running any classification

algorithm, Weka tool asks various option like splitting percentage, used training set, supplied test set, Cross validation option etc. Classificationn mostly occurs with splitting of 80% trainning and 20% testing [6]. But In Weka tool our analysis process was carried out with 10 fold cross validation with selected classifier technique to obtain an output of interest [8]. Weka is a user friendly visualization tool with we have tested various classifier technique and its output performance.

6 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 Weka data mining tool for analysis. The prepossessed data is converted in to suitable form for classification using different classifier approach. The classifier approach is executed with 10 cross validation method. The cross validation is a powerful data analysis process where 10 folds can be done with the available data and an accurate decision can be made on the provided data with good prediction. With the classify tab of Weka tool different classifier approaches are verified. After careful analysis results of proposed classifiers are compared. J48 and Naive Bayes algorithm classifies 32 instances in to 25 correctly classified instances and 7 incorrectly classified instances. Like wise 24 correctly classified instances and 8 incorrectly classified instances are obtained from 32 instances using knn with 5 nearest neighbour. As per our analysis the RBF 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. Similarly False

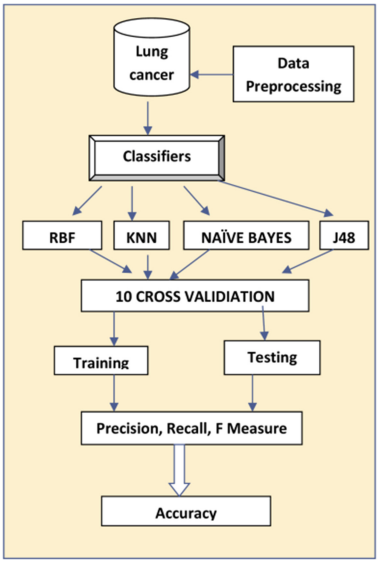

Fig. 2. Process flow of various classifiers in Weka tool.

Positive and False Negative both have a value of 3 each. The output result of various classifiers used in Weka tool on lung cancer data is represented in below table. Generally in confusion matrix Accuracy, Recall, Precision and F-Measure are the key process parameter for classification [4, 14]. Classification accuracy is the measure of number of correct prediction made out from total number of prediction. These parameters depend on some specific outcome. Those are 'TP (True Positive) which is the correctly predicted event values and 'TN (True Negative) is correctly predicted no event values. Similarly 'FP (False Positive) is incorrectly predicted event values and 'FN (False Negative) for incorrectly predicted no event values. Relationship are derived as below (Figs. 2, 3) and (Table 1).

$$Accuracy = \frac{{}^{J}TP + {}^{J}TN}{{}^{J}TP + {}^{J}TN + {}^{J}FP + {}^{J}FN} \tag{7}$$

$$Recall = \frac{{}^{J}TP}{{}^{J}TP + {}^{J}FN} \tag{8}$$

$$Precision = \frac{{}^{J}TP}{{}^{J}TP + {}^{J}FP} \tag{9}$$

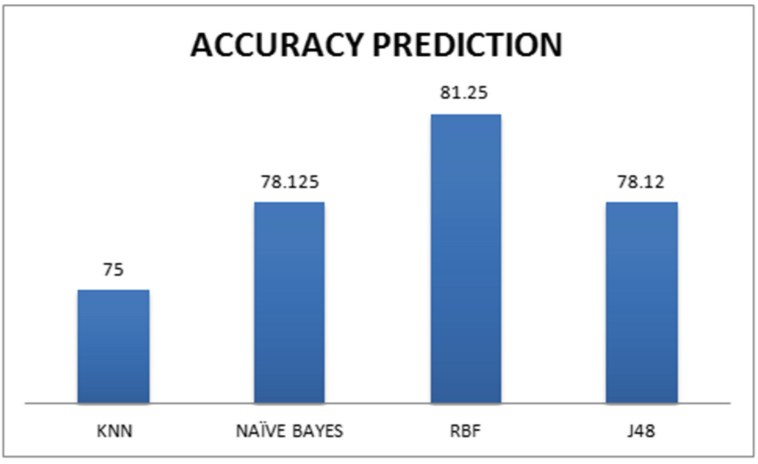

Fig. 3. Accuracy graph.

Table 1. Classifiers output in Weka tool.

Classifier	Precision	Recall	F-Measure	ROC area	Correctly. classified	Incorrectly classified
KNN(5)	.73	.75	.70	.69	75%	25%
Na¨ıve Bayes	.775	.78	.77	.77	78.125%	21.87%
RBF	.813	.813	.813	.749	81.25%	18.75%
J48	.768	.781	.766	.708	78.12%	21.87%

$$F\_Measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{10}$$

7 Conclusion

In this paper we have shown that with RBF classifier the accuracy is found to be 81.25% 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 RBF, accuracy will be further improved.

References

1. https://archive.ics.uci.edu/ml/dataset/Lung+cancer. Accessed 12 Feb 2020
2. WHO Deaths by cause, sex and mortality stratum, World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 25 Jan 2020
3. Ada, R.K.: Early detection and prediction of lung cancer survival using neural network classifier (2013)
4. Alcantud, J.C.R., Varela, G., Santos-Buitrago, B., Santos-Garcia, G., Jimenez, M.F.: Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making. PLoS ONE **14**(6), e0218283 (2019)
5. Asuntha, A., Srinivasan, A.: Deep learning for lung cancer detection and classification. Multimedia Tools Appl. **79**, 1–32 (2020)
6. Bhatia, S., Sinha, Y., Goel, L.: Lung cancer detection: a deep learning approach. In: Bansal, J.C., Das, K.N., Nagar, A., Deep, K., Ojha, A.K. (eds.) Soft Computing for Problem Solving. AISC, vol. 817, pp. 699–705. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1595-4_55
7. Bhuvaneswari, P., Therese, A.B.: Detection of cancer in lung with k- nn classification using genetic algorithm. Procedia Mater. Sci. **10**, 433–440 (2015)
8. Chaubey, N.K., Jayanthi, P.: Disease diagnosis and treatment using deep learning algorithms for the healthcare system. In: Applications of Deep Learning and Big IoT on Personalized Healthcare Services, pp. 99–114. IGI Global (2020)
9. Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., Dhillon, S.K.: Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med. Inform. Decision Making **19**(1), 48 (2019)
10. Hachesu, P.R., Moftian, N., Dehghani, M., Soltani, T.S.: Analyzing a lung cancer patient dataset with the focus on predicting survival rate one year after thoracic surgery. Asian Pacific J. Cancer Prevention: APJCP **18**(6), 1531 (2017)
11. Hosseinzadeh, F., KayvanJoo, A.H., Ebrahimi, M., Goliaei, B.: Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus **2**(1), 238 (2013)
12. Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U.: Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. IEEE Trans. Med. Imag. **38**(8), 1777–1787 (2019)
13. Jacob, D.S., Viswan, R., Manju, V., PadmaSuresh, L., Raj, S.: A survey on breast cancer prediction using data mining techniques. In: 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 256–258. IEEE (2018)
14. Jakimovski, G., Davcev, D.: Using double convolution neural network for lung cancer stage detection. Appl. Sci. **9**(3), 427 (2019)

15. Kadir, T., Gleeson, F.: Lung cancer prediction using machine learning and advanced imaging techniques. Transl. Lung Cancer Res. **7**(3), 304 (2018)
16. Kohad, R., Ahire, V.: Application of machine learning techniques for the diagnosis of lung cancer with ant colony optimization. Int. J. Comput. Appl. **113**(18), 34–41 (2015)
17. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struc. Biotechnol. J. **13**, 8–17 (2015)
18. Krishnaiah, V., Narsimha, G., Chandra, D.N.S.: Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol. **4**(1), 39–45 (2013)
19. Li, X., Hu, B., Li, H., You, B.: Application of artificial intelligence in the diagnosis of multiple primary lung cancer. Thoracic Cancer **10**(11), 2168–2174 (2019)
20. Lynch, C.M., et al.: Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int. J. Med. Inform. **108**, 1–8 (2017)
21. Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., Qian, W.: Detection and classification of pulmonary nodules using convolutional neural networks: a survey. IEEE Access **7**, 78075–78091 (2019)
22. Murty, N.R., Babu, M.P.: A critical study of classification algorithms for lungcancer disease detection and diagnosis. Int. J. Comput. Intell. Res. **13**(5), 1041–1048 (2017)
23. Paing, M.P., Hamamoto, K., Tungjitkusolmun, S., Pintavirooj, C.: Automatic detection and staging of lung tumors using locational features and double-staged classifications. Appl. Sci. **9**(11), 2329 (2019)
24. Patel, D., Shah, Y., Thakkar, N., Shah, K., Shah, M.: Implementation of artificial intelligence techniques for cancer detection. Augmented Human Res. **5**(1), 6 (2020)
25. Pradeep, K., Naveen, N.: Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4. 5 and naive bayes algorithms for healthcare analytics. Procedia computer science **132**, 412–420 (2018)
26. Shakeel, P.M., Tolba, A., Al-Makhadmeh, Z., Jaber, M.M.: Automatic detection of lung cancer from biomedical data set using discrete adaboost optimized ensemble learning generalized neural networks. Neural Comput. Appl. **32**(3), 777–790 (2020)
27. Shanthi, S., Rajkumar, N.: Lung cancer prediction using stochastic diffusion search (sds) based feature selection and machine learning methods. Neural Process. Lett. **1**, 1–14 (2020)
28. Sidey-Gibbons, J.A., Sidey-Gibbons, C.J.: Machine learning in medicine: a practical introduction. BMC Med. Res. Methodol. **19**(1), 64 (2019)
29. Singh, G.A.P., Gupta, P.K.: Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Comput. Appl. **31**(10), 6863–6877 (2018). https://doi.org/10.1007/s00521-018-3518-x
30. Varadharajan, R., Priyan, M., Panchatcharam, P., Vivekanandan, S., Gunasekaran, M.: A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers. J. Ambient Intell. Human. Comput. **1**, 1–12 (2018)
31. Zubi, Z.S., Saad, R.A.: Improves treatment programs of lung cancer using data mining techniques. Journal of Software Engineering and Applications **2014**, 69–77 (2014)