Asian Journal of Research in Infectious Diseases

4(4): 15-22, 2020; Article no.AJRID.60025
ISSN: 2582-3221

Early Lung Cancer Prediction Using Neural Network with Cross-validation

Shawni Dutta[1] and Samir Kumar Bandyopadhyay[2*]

[1]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.
[2]The Bhawanipur Education Society College, Kolkata, India.

Authors' contributions

This work was carried out in collaboration between both authors. Author SD designed the proposed method, coding and statistical work. Author SKB carried out the literature survey and checked the manuscript written by author SD. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRID/2020/v4i430153
Editor(s):
(1) Dr. Giuseppe Murdaca, University of Genoa, Italy.
Reviewers:
(1) Subrato Bharati, Bangladesh University of Engineering and Technology, Bangladesh.
(2) Md. Milon Islam, Khulna University of Engineering & Technology (KUET), Bangladesh.
Complete Peer review History: http://www.sdiarticle4.com/review-history/60025

Original Research Article

Received 01 June 2020
Accepted 06 August 2020
Published 11 August 2020

ABSTRACT

Lung cancer is known as lung carcinoma. It is a disease which is malignant tumor leading to the uncontrolled cell growth in the lung tissue. Lung cancer is caused generally by smoking and the use of tobacco products. It is classified into two broad Small-cell lung Carcinomas and non-Small cell lung carcinomas. Lung cancer treatments include surgery, radiation therapy, chemotherapy, and targeted therapy. Lung Cancer disease is one of the most prominent cause of death in all over world. Early detection of this disease can assist medical care unit as well as physicians to provide counter measures to the patients. The objective of this paper is to approach an automated tool that takes influential causes of lung cancer as input and detect patients with higher probabilities of being affected by this disease. A neural network classifier accompanied by k-fold cross-validation technique is proposed in this paper as a predictive tool. Later, this proposed method is compared with another baseline classifier Gradient Boosting Classifier in order to justify the prediction performance. Experimental results conclude that analyzing interfering causes of lung cancer can effectively accomplish disease classification model with an accuracy of 95%.

**Corresponding author: E-mail: 1954samir@gmail.com;*

Keywords: Lung cancer prediction; neural network; cross-validation; gradient boosting classifier; automated tool.

1. INTRODUCTION

Lung cancer is a serious disease which is the major cause of cancer deaths in people worldwide. Timely detection and screening play leading role in prevention of lung cancer. This paper focuses on predicting patients with lung cancer severity at an early stage so that counter measures can be suggested by the physicians. Prediction at an early stage will assist health care systems to handle this disease carefully. Handling the consequence with care may help medical experts to take informed decision and act accordingly. Data mining and knowledge discovery are applied on past health records to identify hidden patterns and relationship among the data [1].

Disease classification is an important task which has gained special attraction in recent days. Several studies [2-8] have made disease classification methods in various medical fields. All these studies have shown promising classification result in their respective disease detection. Diabetes disease has been diagnosed by implementing 10-fold and 5-fold cross-validation fashion based deep neural network based framework [2]. Haque et. al. [3] focused on liver diagnosis system by implementing Random Forests (RFs) and Artificial Neural Networks (ANNs). Both of these models are estimated using 10-fold cross-validation method [3]. Breast cancer prediction at an early stage has been conducted by implementing Support Vector Machine and K-Nearest Neighbors by training relevant attributes [4]. Another work in [5] developed 10 fold cross validated mathematical model to detect breast cancer using symbolic regression of multi-gene genetic programming (MGGP). Prediction of COVID-19 infected patients' recovery has been predicted using data mining techniques such as decision tree, support vector machine, naive Bayes, logistic regression, random forest, and K-nearest neighbor algorithms [6]. To diagnose COVID-19 automatically from X-ray images convolutional neural network (CNN) and long short-term memory (LSTM) are utilised in [7]. Heart disease detection has been carried out by implementing and comparing several machine learning models such as Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Network (DNN), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (K-NN). After comparison, DNN provided the maximized heart disease detection result [8].

As discussed above, disease classification system can be implemented by using machine learning based intelligent models. This paper approaches towards disease classification system construction for lung cancer identification. A recommended system has been proposed in this paper that automatically analyses casual habits of patient in order to determine possibility of being affected by lung cancer. Supervised machine learning approaches are utilized for obtaining intelligent models for this prediction purpose. The system proposed in this paper automatically captures the interfering factors such as patient's age, alcohol consumption, smoking addiction while deciding whether the patient may suffer from lung cancer or not in near future. The proposed system is basically a classifier model that intended to predict lung cancer suffering possibilities. A neural network based framework followed by k-fold cross validation procedure is implemented for obtaining the prediction in advance. Human brain like operations is simulated by implementing neural network based framework in order to accompany complex problem solving approach. After implementing the model, evaluation process takes place. Different values of k such as 5, 10 and 15 folds are applied for cross-validation scheme. The k-value providing the maximized efficiency is drawn for final evaluation. This evaluation results are compared with Gradient Boosting classifier which is serving as baseline classifier in this context.

2. RELATED WORKS

In the world lung cancer is the most common cancer. After breast and prostate it is the third most common cancer. The standard care for people with early stage of lung cancer is thoracic surgery. Smoking is the most direct cause of lung cancer that leads to 90% of lung cancer deaths. There are other causes leading to lung cancer in non-smoking people attributed to genetic factors and air pollutants such as asbestos, radon gas, and passive smoking [9-10].

Machine learning classifiers were used to extract features for Computed Tomography (CT) image dataset for detecting lung disease in CT images of the thorax. Multi-crop convolutional neural

networks approaches are also applied by researchers for lung nodule classification to detect malignancy. Unsupervised deep embedding clustering analysis has been studied extensively in terms of distance functions for detection of lung cancer [11].

Another research [12] utilized machine learning classifier models for classifying images for lung cancer disease. Input features were extracted from images and classifier models accept those parameters while diagnosing lung disease. A system-theoretic method is introduced in [13] that analyses the diagnosis-to-treatment process for lung cancer patients who receive surgical resections. Some researchers conducted studies on patients containing females and males with a tendency of lung cancer. It revealed that the better prognosis was found in females compared to males after adjustment for age, disease stage and smoking history [14]. The purpose of research [15] is to evaluate the performance of texture features on lung CADx and the consistency with expert diagnosis based on visual features from CT images. Using two multiple resolution residually connected network (MRRN) formulations called incremental-MRRN and dense-MRRN, detect and segment the lung tumors CT images [16]. A new hybrid deep learning framework by combining VGG, data augmentation and spatial transformer network (STN) with CNN is proposed in [17]. As shown in [18], lung cancer prognosis can be carried out by implementing and comparing data mining classifier models such as, Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Tree, Random Forest, and Neural Network. After evaluating, naïve bayes has shown highest accuracy of 57.047%.

3. PROPOSED METHODOLOGY

A multi-step procedure is followed to build the proposed model to be applied on lung cancer dataset. Objective of this study is to detect patients with severe lung disease troubles. The entire workflow of this methodology is depicted in Fig. 1. The required steps of this workflow are elaborated in this section.

3.1 Data Collection and Pre-processing

To fulfill the objective of this paper, a dataset related to Lung cancer is collected from kaggle [19]. The dataset can be formulated as a collection of attributes such as patient's age, smoking tendency, alcohol consumption which are quite promising predictor for determining lung cancer possibilities. The attributes present in the dataset along with consisting values are summarized in Table 1. The attributes like name and surname do not contribute to the process of classification task. Hence, they are eliminated. The attributes such as age, smokes, areaQ, and alcohol contain several numeric values. These attribute values are scaled down within the range from 0 to 1. Performing these pre-processing techniques transformed dataset that can be fitted to classifier model.

3.2 Methodology

Classifications are supervised machine learning techniques that are applied on dataset and mapping inputs to target class [1]. For this purpose neural network architecture is proposed in this paper that accepts several prognostic factors those affect lung cancer and finally predicts possibility of being affected by lung cancer. Hence, the possibility of being affected by lung cancer is the target class of classification procedure. Neural network proposed in this paper is comprised of several neurons. Each of these neurons will accept necessary parameters and apply some activation functions in order to produce outputs. Activation functions are useful to perform diverse computations and produce outputs within a certain range. In other words, activation function is a step that maps input signal into output signal [20].

Table 1. Summary of collected lung cancer dataset

Attribute	Description	Values (in Range)
Name	Patients' name	String values
Surname	Patients' surname	String values
Age	Patients' Age	18-77
Smokes	Smoking Consumption	0-34
AreaQ	Lung Area	1-10
Alkhol	Alcohol Consumption	0-8
Result	Lung Cancer Probability	0-No 1-Yes

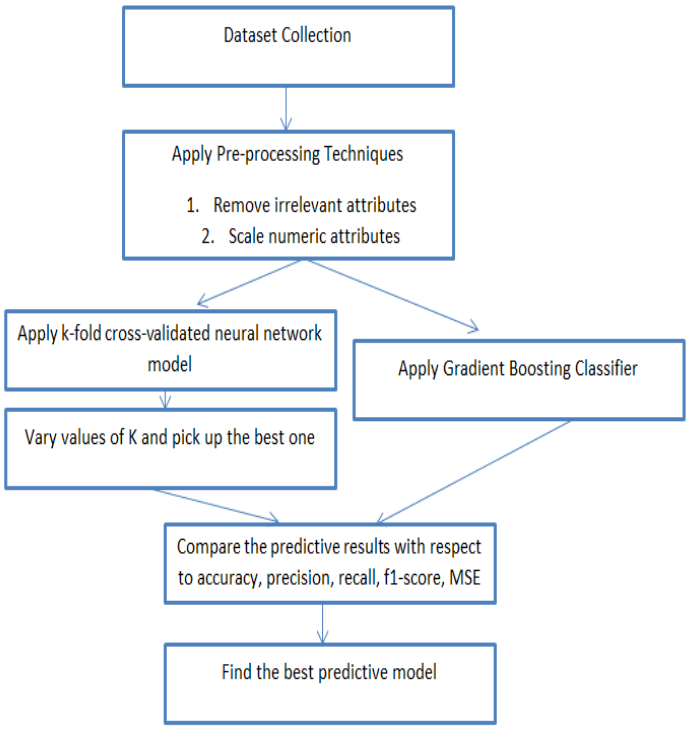

Fig. 1. Proposed system workflow

After configuring this neural model, training process is executed. The training process goes through one cycle known as an epoch where the dataset is partitioned into smaller sections. An iterative process is executed through a couple of batch size that considers subsections of training dataset for completing epoch execution.

3.2.1 Implementation

While designing this model it is necessary to tune hyper-parameters in order to achieve maximized efficiency. This section describes specification of the model along with its hyper-parameters. This model consists of three fully connected layers (Dense) layers having 64,32,1 number of nodes respectively. In this context, sigmoid and relu [20] are two popular activation functions those are applied in each of these specified layer. The first two layers apply relu as activation function and the final layer applies sigmoid activation function.

Finally these aforementioned layers are assembled using adam solver [21] through 30 epochs and with a batch size of 10. Fine-tuning of the hyper-parameters supports the model to obtain best predictive result. The neural network receives a total of 2,433 parameters which are trained to obtain prediction. The summarization of the model is described in Table 2.

This implementation is followed by k-fold cross-validation method for estimating the proficiency of the model. It is a resampling methodology where the dataset is segregated into k groups and in each iteration one group is considered as the test data and the remaining nine folds are considered as training data. Stratified K-fold technique is incorporated in this framework that

Table 2. Summary of neural network model

Layer (type)	Output Shape	Number of Parameters
Dense_31(Dense)	(None, 64)	320
Dense_32(Dense)	(None, 32)	2080
Dense_33(Dense)	(None, 1)	33

validates the cross-validation methodology. The above mentioned model is fitted into the training dataset and it is evaluated against the test dataset. Later evaluation scores for each of these iterations are accumulated and mean score is calculated [22].

This neural network structure accompanied with 5-fold, 10-fold and 15-fold cross validation procedure is applied on lung cancer dataset. The best value of k needs to be chosen after comparing the predictive results. Implementation of this model is evaluated and compared with other benchmark classifiers such as Gradient Boosting Classifier.

3.3 Classifier Performance Evaluation

Once predictions from classifier models are obtained, it is necessary to justify the quality of the predictive results. Justifying the performance of model acquires some evaluating metrics. Use of these metrics will identify the best problem-solving approach. The metrics those are employed by this framework as described as follows:

1. Accuracy is a metric that detects the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model's performance since it does not consider wrong predicted cases. Hence, for addressing the above specified problem, precision and recall is necessary to calculate.
2. Precision identifies the ratio of correct positive results over the number of positive results predicted by the classifier. Recall denotes the number of correct positive results divided by the number of all relevant samples. F1-Score or F-measure is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision and recall.
3. Mean Squared Error (MSE) is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples. A model that exhibits lower value of MSE and higher values of accuracy, F1-Score indicate a better performing model [23].
4. Cohen-Kappa Score is also taken into consideration as an evaluating metric in this paper. This metric is a statistical measure that finds out inter-rate agreement for qualitative items for classification problem [24].

Precisely, the aforementioned metrics can be defined as follows with given True Positive, True Negative, False Positive, False Negative as TP,TN,FP,FN respectively:

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

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

F1- Measure or F1-Score=$\frac{2* \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$

Cohen-Kappa Score= $\frac{\text{p0}-\text{pe}}{1-pe}$

where p_o denotes relative observed agreement among raters and p_e is the probability of agreement by chance.

MSE= $(\sum_{i=1}^{N}(Xi - Xi')^2$ / N) where Xi is the actual value and Xi' is the predicted value.

To address the best problem solving model, it should exhibit lower MSE value and higher values of accuracy, F1-Score, and Cohen-kappa score.

3.4 Baseline Classifier

Gradient boosting classifier is implemented in this paper that serves as baseline while comparing the performance of the proposed method. Gradient boosting algorithm [25] is a boosting technique based classifier that learns by fitting consecutively new models into new models to provide a more accurate estimate of the response variable. It constructs new-base models which decrease the loss function obtained from trained samples. From these calculations the errors are measured and analysed for optimal prediction of results. Loss function calculates the range of detected rate which compares with desired target. Onward stepwise process is most popular method for updating different with various attributes. The accuracy is optimized by reducing loss function and adding base learners at all stages.

The transformed and pre-processed data are partitioned into training and testing set with a ratio of 8:2. Gradient Boost classifier is built based on 500 numbers of estimators on which the boosting is terminated. After implementation, training dataset is fitted into the classifier model and later predictions are obtained for test dataset. Prediction outcomes are evaluated

against accuracy, f1-score, cohen-kappa score and MSE.

4. EXPERIMENTAL RESULTS

In this methodology, k-fold cross-validation method has been used in order to estimate the actual performance model. The value of k should be chosen wisely in order to find out the best predictive result. Multiple values of k are picked up and their performances are validated against accuracy, precision, recall and f1-score. The comparative study shown in Table 3 identifies that the best possible value of k should be chosen as 10 in order to estimate the best classification result.

As shown in Table 3, value of k as 10 provides the best result in lung disease classification process. Hence, neural network along with 10-fold cross-validation method is summarized in Table 4. In this case, gradient boosting method is considered as baseline classifier model for justifying the performance of proposed classifier. Table 4 provides a comparative analysis between the proposed model and Gradient Boosting classifier in terms of specified evaluating metrics such as accuracy, precision, recall, f1-score, cohen-kappa score and MSE. This analysis clearly shows that proposed model is superior while detecting patients having lung disease severity.

5. DISCUSSION

Neural networks are quite promising in simulating brain like operation in order to accomplish complex problem solving paradigm. Several human habits and other characteristics (like age) are fed as input to the neural network models. The neurons (or nodes) present in the network can map the input features into the target class. In this case the target class is the probability of occurring lung cancer disease. The implementation of this feed-forward neural network receives certain fine-tuned hyper-parameters in order to maximize the classification accuracy. However, k-fold cross-validation approach is also incorporated in this research work while estimating the performance of classification. The value of k has been varied for the values of 5, 10, 15 and it is observed that the value of k=10 provides the best possible efficiency. This classification method does not receive any lung cancer related images rather it focuses on smoking tendency, age, and alcohol consumption etc. in order to recognize patients with lung cancer severity. According to [26], it is well-known fact that alcohol consumption is the leading cause and smoking is the second leading cause of lung cancer occurrence. Hence, it is justified to analyze these parameters while investigating the lung cancer probabilities. Instead of being focused to CT image related classification procedure as carried out by [11-18], past or current habits of patients can be utilized in the domain of lung disease classification tool. It is the leading cause of years of life lost because it is associated with the highest economic burden relative to other tumour types. Researches have been going on for achieving improved outcomes of lung cancer.

Table 3. Classification performance for different values of k in cross-validation

Value of K in Cross-validation	Accuracy	Precision	Recall	F1-score
K=5	91.52%	0.93	0.89	0.91
K=10	95.0%	0.95	0.94	0.94
K=15	88.33	0.81	0.87	0.83

Table 4. Performance of proposed model with respect to baseline classifier

Performance Measure Metrics	Accuracy	Precision	Recall	F1-Score	Cohen-Kappa Score	MSE
Neural Network with Cross-validation	95.0%	0.95	0.94	0.94	0.9	0.05
Gradient Boosting Classifier	91.67%	0.92	0.93	0.92	0.82	0.08

6. CONCLUSIONS

Machine learning based lung cancer prediction model has been approached to support clinicians in managing patients' trouble. Neural network along with 10-fold cross validation procedure is proposed in this paper that predicts lung cancer in advance. The predictive model accepts past medical records and the model is accompanied by designing with fine-tuning parameters. Experimental results have shown promising prediction results with an accuracy of 95%, precision of 0.95, recall of 0.94, f1-score of 0.94, cohen-kappa score of 0.9 and MSE of 0.05. Analysing habits like smoking, alcohol consumption, age, it is possible to detect lung cancer trouble in patients. Incorporating more influential factors to this model may help in providing more accurate predictions. However, this research area can even be extended by analysing CT images for obtaining lung cancer identification tool. Along with receiving and analysing extracted features from CT images, emphasis on human habits can help in diagnosing lung disease. This paper established that relative to the huge health, social, and economic burden associated with lung cancer, the level of world research output lags significantly behind that of research on other malignancies. Quality of research works for detecting lung cancer is much lower than to basic science and medical research.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Michalski RS, Carbonell JG, Mitchell TM. Machine learning an artifical intelligence approach. Tioga Press, Palo Alto; 1983.
2. Safial Islam Ayon, Md. Milon Islam. Diabetes Prediction: A deep learning approach. International Journal of Information Engineering and Electronic Business (IJIEEB). 2019;11(2):21-27. DOI: 10.5815/ijieeb.2019.02.03
3. Md. Rezwanul Haque, Md. Milon Islam, Hasib Iqbal, Md. Sumon Reza, and Md. Kamrul Hasan. Performance evaluation of random forests and artificial neural networks for the classification of liver disorder. 2018 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IEEE, Rajshahi, Bangladesh. 2018;1-5:8-9.
4. Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, Md. Kamrul Hasan. Prediction of breast cancer using support vector machine and K-Nearest neighbors. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, Dhaka, Bangladesh. 2017;226-229:21-23.
5. Md. Kamrul Hasan, Md. Milon Islam, A. Hashem MM. Mathematical model development to detect breast cancer using multigene genetic programming. 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), IEEE, Dhaka, Bangladesh. 2016;574-57:13-14.
6. Muhammad Lawan Jibril, Md. Milon Islam, Usman Sani Sharif, Safial Islam Ayon, Predictive data mining models for novel coronavirus (COVID-19) infected patients recovery. SN Computer Science, Springer. 2020;1(4):206.
7. Islam Z, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) Using X-ray Images, (2020), med Rxiv. 2020;1–20. Available:https://doi.org/10.1101/2020.06.18.20134718.
8. Safial Islam Ayon, Md. Milon Islam and Md. Rahat Hossain. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, Taylor & Francis. 2020;1-20. Available:https://doi.org/10.1080/03772063.2020.1713916
9. Cancer Research UK. Lung cancer and smoking statistics -Key Facts; 2011. Available:http://info.cancerresearchuk.org/cancerstats/keyfacts/lung-cancer/.
10. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016.' CA, Cancer J. Clin. 2016;66(1):730.
11. Shen et al. W. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," Pattern Recognition. 2017;61:663673.

12. Rebouças Filho PP, Rebouças EDS, Marinho LB, Sarmento RM, Tavares JMR, de Albuquerque VHC. Analysis of human tissue densities: A new approach to extract features from medical images. Pattern Recognition. Letter. 2017;94:211218.
13. Lee et al. HK. A System-theoretic method for modeling, analysis, and improvement of lung cancer diagnosis-to-surgery process in IEEE. Transactions on Automation Science and Engineering. 2018;15(2):531-544.
DOI: 10.1109/TASE.2016.2643627.
14. Dela Cruz, Charles S et al. Lung cancer: epidemiology, etiology, and prevention. Clinics in chest medicine. 2011;32(4):605-44.
DOI:10.1016/j.ccm.2011.09.001
15. Fangfang Han, Guopeng Zhang, Huafeng Wang, Bowen Song, Hongbing Lu, Dazhe Zhao, Hong Zhao and Zhengrong Liang. A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI Database. 2013 IEEE International Conference on Medical Imaging Physics and Engineering, Shenyang. 2013;14-18.
DOI: 10.1109/ICMIPE.2013.6864494.
16. Jiang J. et al. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT Images, in IEEE. Transactions on Medical Imaging. 2019;38(1):134-144.
DOI: 10.1109/TMI.2018.2857800.
17. Subrato Bharati, Prajoy Podder, Rubaiyat Hossain Mondal M. Hybrid deep learning for detecting lung diseases from X-ray images. Informatics in Medicine Unlocked. 2020;20:100391.
Available:https://doi.org/10.1016/j.imu.2020.100391.
18. Bharati S, Podder P, Mondal R, Mahmood A, Raihan-Al-Masud M. Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer. Advances in Intelligent Systems and Computing. 2020;941. Springer, Cham. Available:https://doi.org/10.1007/978-3-030-16660-1_44
19. Yusuf Dede. Lung Cancer Data Set, Version 1. Retrieved on May 28,2020 from Available:https://www.kaggle.com/yusufdede/lung-cancer-dataset
20. Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation functions: comparison of trends in practice and research for deep learning. arXiv, abs/1811.03378. 2018;1–20.
21. Kingma DP, Ba JL, Adam: A method for stochastic optimization. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 2015;1–15.
22. Kirschen RH, O'Higgins EA, Lee RT. A study of cross-validation and bootstrap for accuracy estimation and model selection. Am. J. Orthod. Dentofac. Orthop. 2000;118(4);456–461.
DOI: 10.1067/mod.2000.109032.
23. H. M, S. M.N. A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manag. Process. 2015;5(2):01–11.
DOI: 10.5121/ijdkp.2015.5201.
24. Vieira SM, Kaymak U, Sousa JMC. Cohen's kappa coefficient as a performance measure for feature selection. 2010 IEEE World Congr. Comput. Intell. WCCI 2010; 2010.
DOI: 10.1109/FUZZY.2010.5584447.
25. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front. Neurorobot. 2013;7. DEC, 2013,
DOI: 10.3389/fnbot.2013.00021.
26. Jose Ramon Troche, Susan T. Mayne, Neal D. Freedman, Fatma M. Sheb, Christian C. Abnet. The association between alcohol consumption and lung carcinoma by histological subtype. American journal of epidemiology. 2016;183(2):110-21.
DOI:10.1093/aje/kwv170

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/60025