Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

Early Lung Cancer Prediction Using Neural Network with Cross-validation

Shawni Dutta[1] and Prof. Samir Kumar Bandyopadhyay[2]
[1]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.
[2] Academic Advisor, The Bhawanipur Education Society College, Kolkata, India.

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 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 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.

Keywords: Lung Cancer Prediction, Neural Network, Cross-validation, Gradient Boosting Classifier, Automated tool.

1. Introduction-

Past health record of a patient can be utilised in early prediction of any disease. 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. A recommended system is proposed in this paper that automatically analyses previous health records of patient in order to determine possibility of being affected by lung cancer. Supervised machine learning approaches are utilized 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 10-fold cross validation procedure is implemented for obtaining the prediction in advance. After implementing the model, evaluation process takes place. The evaluation results are compared with Gradient Boosting classifier which is serving as baseline classifier in this context.

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

2. Related Work-

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 [1-2]. 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 [3-6].

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 [7]. It may be the evidence of sex being a predictor in lung cancer prognosis. Similarly, it also showed results suggesting poor prognosis for older patients compared to younger patients [8]. So age of patients as not an important prognostic factor in lung cancer survival and treatment.

Machine learning classifiers were used to extract features for 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 [9].

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 required steps are explained as follows-

3.1 Data Collection and Pre-processing-

To fulfill the objective of this paper, a dataset related to Lung cancer is collected from kaggle [Add Ref]. 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. Pre-processing techniques such as missing values handling, irrelevant attributes (like patient's name) elimination are applied to the collected dataset. Scaling of attributes within specified range will provide a transformed dataset that can be fitted to classifier model.

3.2 Methodology-

Classifications are the techniques that are applied on dataset and mapping inputs to target class. For this purpose neural network architecture is proposed in this paper that accepts several factors those affect lung cancer and finally predicts possibility of being affected by lung cancer. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

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.

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.

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 Dense layers having 64,32,1 number of nodes respectively. In this context, sigmoid and relu 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 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 Figure1.

Layer (type)	Output Shape	Param #
dense_31 (Dense)	(None, 64)	320
dense_32 (Dense)	(None, 32)	2080
dense_33 (Dense)	(None, 1)	33

Figure1. Summary of Neural Network model

This implementation is followed by 10-fold cross-validation method for estimating the proficiency of the model. It is a resampling methodology where the dataset is segregated into 10 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 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.

This neural network structure accompanied with 10-fold cross validation procedure is applied on lung cancer dataset. Implementation of this model is evaluated and compared with other benchmark classifiers such as Gradient Boosting Classifier.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

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.
4. Cohen-Kappa Score [21] 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.

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-

$$\text{Accuracy} = TP+TN/(TP+FP+TN+TP)$$

$$\text{Recall} = TP/(TP+FN) \qquad \text{Precision} = TP/(TP+FP)$$

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

$$\text{Cohen-Kappa Score} = (p_0-p_e)/(1-p_e)$$ where p_o denotes relative observed agreement among raters and p_e is the probability of agreement by chance.

$$\text{MSE} = \left(\sum_{i=1}^{N}(Xi - Xi')^2 / \mathbf{N}\right)$$ 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

3.4 Baseline Classifier-

Gradient boosting classifier is implemented in this paper that serves as baseline while comparing the performance of the proposed method. This classifier is based on boosting technique

Gradient boosting algorithm[18] is another 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-

The prediction performance of proposed model that is, neural network along with 10-fold cross-validation method is indicated in Table1. A comparative analysis with Gradient Boosting classifier in terms of specified evaluating metrics are also provided. This analysis clearly shows that proposed model is superior while detecting patients having lung disease severity.

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

Table1. Performance of Proposed model with respect to baseline classifier

5. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2020 doi:10.20944/preprints202006.0333.v1

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%, f1-score of 0.94, cohen-kappa score of 0.9 and MSE of 0.05. Incorporating more influential factors to this model may help in providing more accurate predictions.

References-

[1]. N. Camarlinghi, ``Automatic detection of lung nodules in computed tomography images: Training and validation of algorithms using public research databases,'' Eur. Phys. J. Plus, vol. 128, no. 9, p. 110,Sep. 2013.

[2]. R. L. Siegel, K. D. Miller, and A. Jemal, ``Cancer statistics, 2016,'' CA, Cancer J. Clin., vol. 66, no. 1, pp. 730, Jan. 2016.

[3]. P. P. RebouçasFilho, E. D. S. Rebouças, L. B. Marinho, R. M. Sarmento,J. M. R. Tavares, and V. H. C. de Albuquerque, ``Analysis of human tissue densities: A new approach to extract features from medical images, ''Pattern Recognition. Letter, vol. 94, pp. 211218, Jul. 2017.

[4]. W. Shen et al., ``Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification,'' Pattern Recognition, vol. 61, pp. 663673, Jan. 2017.

[5]. FangfangHan, 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.

[6]. Hyo Kyung Lee, Student Member, IEEE, FengJu, Member, IEEE, Raymond U. Osarogiagbon, Nicholas Faris, Xinhua Yu, Fedoria Rugless, Shan Jiang, and JingshanLi, Fellow,” A System-Theoretic Method for Modeling, Analysis, and Improvement of Lung Cancer Diagnosis-to-Surgery Process”,2017.

[7]. JueJiang , Yu-chi Hu, Chia-Ju Liu, Darragh Halpenny, Matthew D. Hellmann, Joseph O. Deasy, Gig Mageras and Harini Veeraraghavan” Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images”,2018.

[8]. Schiller J, Parles K, Cipau A. 100 Questions & answers about lung cancer. Second edition. United Sate of America: Jones & Bartlett Learning; 2009.

[9]. Cancer Research UK. Lung cancer and smoking statistics -Key Facts; 2011. Available:http://info.cancerresearchuk.org/cancerstats/keyfacts/lung-cancer/.