MACHINE LEARNING TECHNIQUES USED TO IDENTIFY LUNG CANCER

[1]Dr.M. Sujatha, [2]S. Prabhakar
Associate Professor, Assistant Professor
Department of Computer Science Engineering
[1]Jyothishmathi Institute of Technological Science, Nustulapur, Karimnagar, India
[2] University College of Engineering and Technology for Women,Kakatiya University Campus, India

Abstract
A novel approach for lung cancer analysed pulmonary nodules to predict the disease of the patients. Lung cancer dataset is implemented by using machine learning to assist physicians in handling non-specific.These systems can decrease variation in classifying nodules to achieve decision-making and decrease the number of good nodules which are unnecessarilyacted upon. In this paper, provides an overview of predicting lung cancer proposed Machine Learning Techniques for Lung cancer (MLTLC) so far and emphasis on strengths and weaknesses. Machine learning algorithms implemented challenges in building and validating lung cancer dataset taken from UCI repository. The proposed method conducted the experiment on Lung cancer to reduce the inconsistencies. The proposed MLTLC method compared with remaining machine learning classifiers.

Index Terms **-** Data preprocessing, Machine learning, Random Forest, Neural Network.

1. INTRODUCTION

The United States National Pulmonary Examination is reduced to 20% of lung cancer [1]. In the beginning of detection of any type of cancer is of paramount importance to pave the way for successful cancer treatment. Unfortunately, most cancers are only detected once they reach an advanced and incurable stage. Individuals who are victims of many types of cancer do not know about it until it's too late.

The important challenges is the process of discovering common knowledge of processdata in machine learning [2]. Data taken directly from the raw data may have dirty data, errors, or more importantly, they are not ready to implement the data. So, pre-processing Lung cancer data is required in data mining to proceed futher. The Lung Cancer dataset contains missing values. The Lung cancer dataset is pre-processed, by imputing missing values based Machine learning techniques. Moreover, the increasing heap of data in the applications of cognitive science, industry and modern business calls for the needs of more complex tools for analysis[3]. Thanks to the pre-processing of the data, it is possible to transform the impossible to possible, adapt the data to meet the input need for every data extraction algorithm. Data pre-processing involves data reduction techniques that aim to reduce data complexity, detect or remove irrelevant, inconsistent, dirty and noisy data[4].

Many people do not even go to the doctor to get themselves checked for several reasons, which caninclude, affordability, fear, travelling cost or even time. Everyone is so absorbed in their work that they disregard the possibility of having cancer, even after the symptoms start to manifest. Hence, this was one of the driving factors for us to make such an application [5]. To be able to measure how much of a societal impact it can have we need to evaluate its impact on both general users and doctors. Therefore, by focusing on just UAE-based users, our proposed application is predicted to have a significant impact on the country as a whole. By focusing on the patient-specific aspect of LCPS, it becomes more apparent why having a tailored prediction, that is specific to the patient's health, is important. It gives the oncologist and patient room to forecast and define a targeted treatment process, helping in better patient care and improving the patient's chances of survival. Thus, having such a system will decrease lung cancer death cases significantly, especially since there are no similar systems in the market.

Of the myriad opportunities for use of Machine Learning in clinical practice, medical imaging workflows are most likely to be impacted in the near term. Machine learning-driven algorithms that automatically process 2- or 3-dimensional image scans to identify clinical signs (e.g., tumors or lesions) or determine likely diagnoses have been published and some are progressing through regulatory steps toward the marke[6]t. Many of these use deep learning, a form of ML based on layered representations of variables, referred to as neural networks.

2.LITERATURE SURVEY

Maryam Aljanabi et al[7] build an efficient predictive model on heart disease. The machine-learningalgorithm is used to pre-processed heart disease data to get better diagnosis results. In this paper, Artificial Neural Network achieved more accuracy compared to Decision Tree. Definitely, the machine learning techniques analyzed of heart disease for patients diagnosis.

Yomna Omar et al[8],treated cancer is a challenge task to detect the disease is in an advanced stage or early stage. They focused on UAE-based users, to predict significance of cancer on the country. By focusing on the patient-specific aspect of LCPS, it becomes more apparent why having a tailored prediction, that is specific to the patient's health, is important.

M.Akhiljabbar et al.[9] has been tested with 6 medical data sets and 1 non medical data set. Out of 7 data sets, 6 data sets were chosen from UCI Repository and heart disease A.P was taken from various corporate hospitals in Andhra Pradesh, and attributes are selected based on opinion from expert doctor's advice.

To understand how deep learning methods leverage image data to perform recognition tasks, imagine you are entering a dark room and looking for the light switch. From past experience, you have learned to associate light switches with predictable locations within the configuration of a room[10]. Many computer vision–based image processing algorithms, including deep learning, mimic this behavior to identify factors that are associated with the recognition task at hand. Deep learning is especially powerful in its ability to interpret images because of the complexity of the factors it can consider [11].

3. METHODOLOGY

Preprocessing

The pre-processing analysis's various lung structures and nodules to imputes missing values[12]. Feature extraction is segmenting lungs from CT scan recognizes several attributes to support the classifier and to organize candidates in improved manner[13]. The Proposed MLTLC method is done to pre-process the along cancer dataseti.e. the lung cancer is cleansed , then given to proposed method to improve the model is shown in Figure 1.

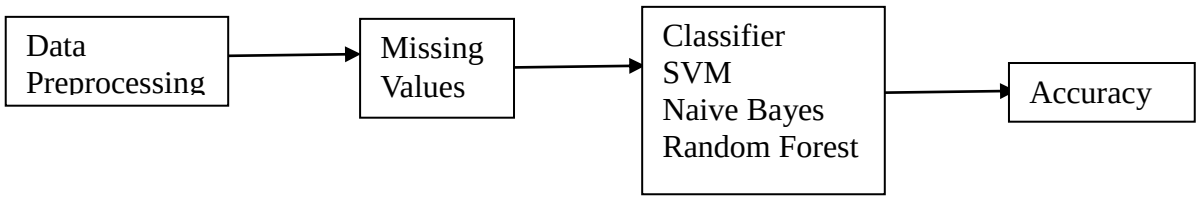

Figure1 TheProposed MLTLC Method

3.1. MACHINE LEARNING TECHNIQUES FOR LUNG CANCER DIAGNOSIS

Clinical analysis of Lung cancer used the proposed MLTLC method for forecast the patients suffering from disease. The recovering from diagnosis of Lung cancer is pre-processed, given to the Proposed MLTLC to achieve the accuracy.

Naive Bayes

Naïve Bayes, which carried out probabilistic inference on idea of conditional independence between predictors [14]. The statistical predictive models, Naïve analyses the subjection among attributes values and classes to obtain a conditional probability.

Random Forest Classifier

Random forest classifier is analysed sub-sampling based on several decision trees to reduced over-fitting well. It is randomization on several decisions trees to improve the development of the entire classification by extracting attributes [15].Random Forest by building a decision tree, a powerful ensemble. Random Forest recover the diagnosis by using a pair of classifier of given training samples. The primary challenge of Random Forest is finding the ensemble of classifiers to produce a good classifier i.e. better than individual classifier. By voting technique, the aggregating method will be improved performance accuracy than individual classifier.

SVM

High dimensional space is given to SVM formtwo-class classification problem, using RBF kernel. SVM classifier transformed space into patterns vectors near to hyper plane decision boundary [16]. It splits two classes of patterns based on given examples based on hyper plane decision boundary. In machine learning, supportvector classifier separate binary classification based on class boundary maximizing the hyper plane the margin in the training data [17].

4. EXPERIMENTAL ANALYSIS

The lung cancer datasets are available in the UCI repository [18]. The experimental analysis on lung cancer is carried out in two phases' i.e. phase 1: training phase, and phase 2: test phase. The Lung cancer dataset was partitioned by 70% of training phase and 30% of testing phase. The proposed MLTLC method conducted the experiments were as follows: (1) Accuracy, (2) False Positive Rate (FPR), (3) False Negative Rate (FNR), (4) True Positive Rate (TPR), and (5) True Negative Rate (TNR) [19]. The real positive rate (TPR) versus false positive rate (FPR) in various threshold settings. The real positive rate is called sensitivity in machine learning. The false positive rate is called probability of false alarm i.e (1 – specificity) is shown in Figure 2 [20].

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

$$TNR = \frac{TN}{TN + FP}$$

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

$$FPR = \frac{FP}{FP + TN}$$

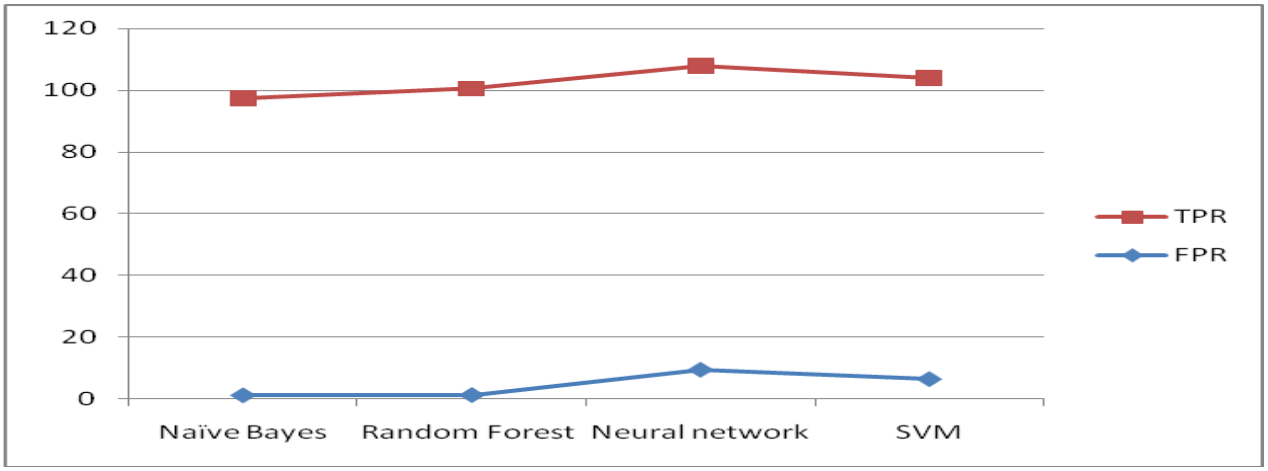

Figure 2 The proposed MLTLC method for lung cancer TPR Vs. FPR

Real negative rate measures the percentage of negatives i.e.The patient who are suffering from disease. The percentage of correctly identified the patients suffering from lung cancer is shown in Figure 3.

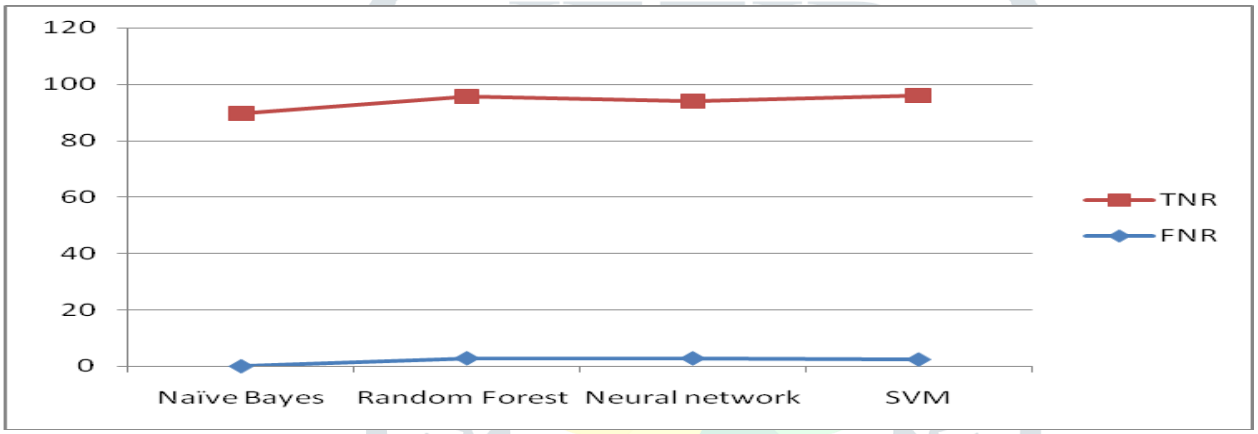

Figure 3 The proposed MLTLC method for lung cancer TNR Vs. FNR

The accuracy of the Proposed MLTLC method is implemented on Lung cancer. The classification of proposed method has more accuracy for Neural Networks compared to Naive Bayes classifier, Random Forest ensemble, and SVM classifier is shown in Figure 4.

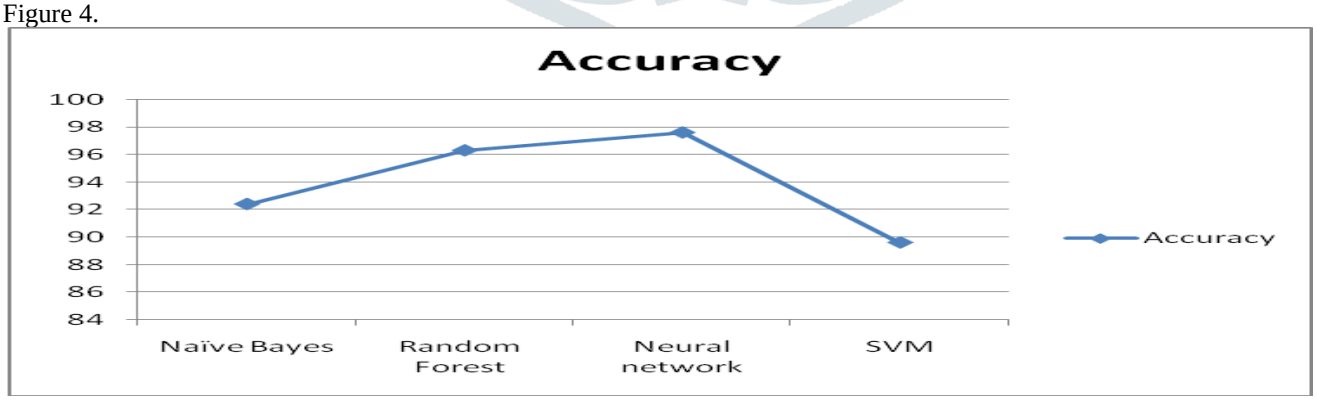

Figure 4 Accuracy of the Proposed method

5. CONCLUSION

An overview of the main methods used for nodule classification and prediction of lung cancer data Computed tomography (CT). Through our experience, with the provision of adequate training data, the latest technology is being achieved using SVM and Radom Forest learning to achieve taxonomic performance. When evaluating the performance of the system, it is important to be aware of restrictions or other training groups and validation of the data used, ie, smokers, non-smokers, or patients with current or previous history of malignancies. The classification of proposed method has more accuracy for Neural Networks compared to Naive Bayes classifier, Random Forest ensemble, and SVM classifier.

REFERENCES

1 National Lung Screening Trial Research Team ,Aberle DR, Adams AM, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N Engl J Med, PP 395-409, 2011.

2 American Cancer Society Cancer facts and figures 2017 Tech. rep. American Cancer Society, 2017. url: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf

3 Lung CT Screening Reporting & Data System. Available online: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads

4 Apoorva Mahale1, ChetanRawool, Dinesh Tolani1, DeepeshBathija, Prof. KajalJewani, "A Survey On Lung Cancer Detection Using Image Data Analysis And Machine Learning", International Journal Of Innovative Research In Computer And Communication Engineering Vol. 5(1), 2017.

5 WafaaAlakwaa, Mohammad Nassef, AmrBadr, Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN),nternational Journal of Advanced Computer Science and Applications, Vol. 8(8),PP. 409-418, 2017.

6 N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26.

7 Maryam Aljanabi, Mahmoud H. Qutqut, Mohammad Hijjawi, Machine Learning Classification Techniques for Heart Disease Prediction: A Review, International Journal of Engineering and Technology, vol.5 , PP 1-7, 2018.

8 Yomna Omar, Abdullah Tasleem, Michel Pasquier and AssimSagahyroon, Lung Cancer Prognosis System using Data Mining Techniques, HEALTHINF- 11th International Conference on Health Informaticsm, PP 361-368, 2018.

9 M.Akhiljabbar, B.L Deekshatulu, Priti Chandra ,Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm ,International Conference on Computational Intelligence: Modeling Techniques and Applications, Vol.10, PP. 85-94, 2013.

10 ELCAP Public Lung Image Database. Available online: http://www.via.cornell.edu/databases/lungdb.html.

11 Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Clarke, L.P. Data From LIDC-IDRI. The Cancer Imaging Archive. Available online: http://doi.org/10.7937/K9/TCIA.2015.

12 Classification of lung nodules. In Proceedings of 13the EMBC,.; pp. 5461–5464, 2013.

13 Liu, X.L.; Hou, F.; Hao, A. Multi-view multi-scale CNNs for lung nodule type classification from CT images.PatternRecognit. 2018, 77, 262–275.

14 Witten, I., Frank, E., Data Mining: Practical Machine Learning Tool and Techniques, Amsterdam: Morgan Kaufman, 2016.

15 Liaw, Andy, Wiener, Matthew: Classification and regression by random forest. R news 2(3), 18–22 (2002)

16 Pathak, AkhandaNand, And Ramesh Kumar Sunkaria. "Multiclass Brain Tumor Classification Using Svm." International Journal Of Computer Applications (2014).

17 Mahale, Apoorva, Et Al. "Svm Classifier Based Cad System For Lung Cancer Detection." Ijecs, Vol.6(5), 2017.

18 https://archive.ics.uci.edu/ml/datasets/lung+cancer

19 https://www.ihealthcareanalyst.com/pre-screening-diagnostic-technology-adoption-computer-aided-detection-market/

20 https://en.wikipedia.org/wiki/Receiver_operating_characteristic

JETIR