Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, June 2021: 1652-1656

Classification Methods for Lung Cancer Using Machine Learning

Anamika[a], Prof. Alok Kumar[b]

[a*,b] Department of ECE, Ajay Kumar Garg Engineering College, Ghaziabad, India

Abstract

Lung cancer is the most dangerous disease in the world. According to WHO (World Health Organization) millions of peoples loses their life due to lung cancer. The spread of lung cancer is increasing day by day. The one more essential reason of increasing lung cancer is that it can't be detected in its early stages. So many researchers gave different techniques to detect lung cancer in its very early stage. But machine learning is the most efficient technique because its computational capability is very high. Through detailed data processing this approach easily forecasts the illness. In this paper we examine different computer classifier learning technologies for the data available in the UCI machine learning repository on available lung cancer. The knowledge is predisposed and modified to a paired configuration and a noteworthy classifier system in weka apparatus is used to group the information set into cancer and non-cancerous. The analysis process indicates that the proposed Multilayer perceptron classification has achieved an unprecedented 98.69 percent precision and is seen as the feasible classification tool for expectations of Lung malignancy knowledge.

Keywords: Cancerous, Lung Cancer, Machine Learning, Multi-Layer Perceptron, Non-Cancerous

1. Introduction

Lung cancer originates inside the lungs and extends to the rest of the body. Lung cancer is classified into two categories that are main. "Small cell lung cancer" and "non-small cell lung cancer" are two types. Small cell lung cancer is a disease which causes malignant cells to grow in the tissues of the lungs. Small cell lung cancer and the different types of non-small cell lung cancer are both classified as small cell carcinoma. Smoking is the greatest cause of small-cell lung cancer. Non-small cell lung cancer is a disease in which cancerous cells grow in lung tissues. Some symptoms specific to the patient, including chest pain, shortness of breath, weight loss, etc. Early detection is very necessary such that preventing mortality may be successful through careful surveillance. The progress rate of lung cancer has been quite satisfactory with proper treatment and screening. Machine learning has already shown to be an effective means of both diagnosing and predicting human diseases. Machine learning determines the diagnosis for the greater level of accuracy. Machine learning is the primary method of medical care today. Machine learning algorithms are being used in healthcare programs around the globe. When computer intelligence is introduced, the precise diagnosis of diseases may be debated. Machine learning (ML) attempts to analyze the data to learn how to diagnose the specific disease problem. This facilitates the detection of the origin of diseases by medical practitioners. Image retrieval Since analyzing the picture using various machine learning methods, the study has proven accurate. It allows for a more concise evaluation of the conditions such that money and effort can be expended and the overall percentage of its profit can therefore be improved. ML monitors infectious outbreaks such that effective action can be taken.". It is imperative that we standardize the use of machine learning techniques so they can be more accurate and meaningful. Therefore, improving further on the algorithm would help physicians, as a medical catalyst of high precision and decent performance in clinical decision-making. Machine learning is divided into three categories: unsupervised, guided, and reinforcement learning Under guided learning, grouping and regression are two separate methods.

2. Related Work

CAD (Computer-Aided Diagnostic) systems were developed by Sheeraz Akram et al. (2015) to support medical professionals recognise and categorise the issue. ANN procedure is used in this framework and the outcome of the precision is 96.68 percent and sensitivity is 96.95 percent. Dhalia et al. (2017) use ACO (Ant colony optimization) to estimate the probability of a pulmonary hamartoma nodule to need follow-up treatment.

Alam et al. (2018) indicates that an algorithm using Watershed transform, GLCM and SVM classifier is most effective when applied to the detection of cancer-affected cells and the subsequent level such as the initial, middle, or final stage. With the help of adaptive thresholding segmentation, SVM, Muthazhagan et al. intended to provide an effective and reliable form of computer-assisted diagnosis of lung cancer. The result was a 98% accurate classifier for prediction of lung cancer tumours.

3. Proposed Work

Using a separate classifier approach, the prepossessed data is translated into an appropriate type for classification. With 10 cross validation methods, the classifier strategy is executed. Cross validation is a strong data analysis tool where it is possible to do 10 folds with the available data and to make a precise judgement on the given data with reasonable prediction. There are 154 Dataset instances and 16 functions are usable (1 is class attribute and 15 is input data). Weka is an open-source framework used for sorting, clustering, regression, and data visualisation. Weka typically supports the input file format for .csv or .arff extensions. With 80 per cent teaching break and 20 per cent research, classification also occurs. But in order to achieve a performance of interest, our analysis method was conducted with 10-fold cross validation in the Weka tool with the chosen classifier technique. There are numerous data processing tabs for Weka Explore, such as prepossess, define, cluster, link, choose and visualise attributes.

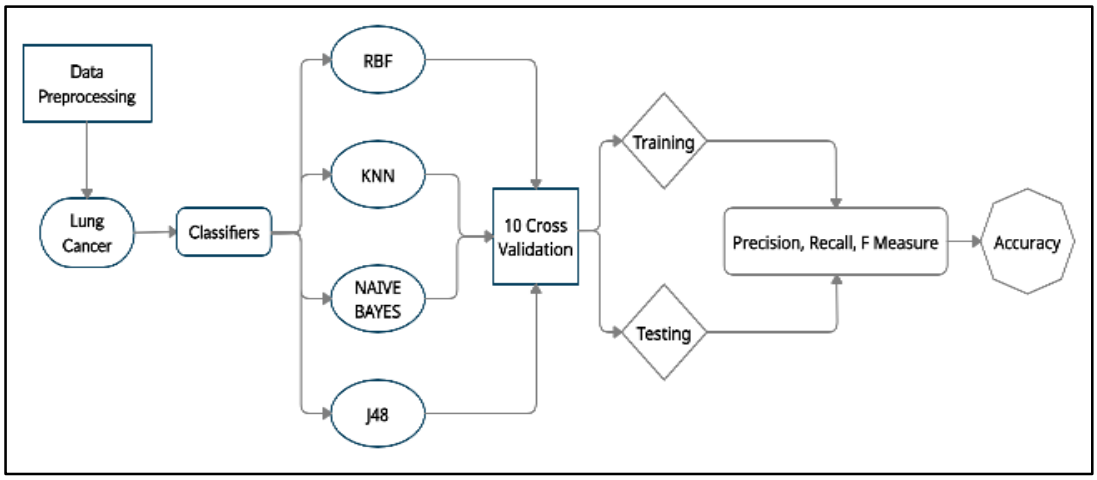

Fig. 1 Process Flow in Weka Tool

3.1 Data Set

There was a dataset in the UCI machine learning repository. Dataset consist 154 instances and it has 16 features (1 is class attribute and 15 is input data). Fig. 1 shows the dataset which is inputted in Weka tool. In the attributes of this dataset; age, gender and symptoms of Lung cancer is shown.

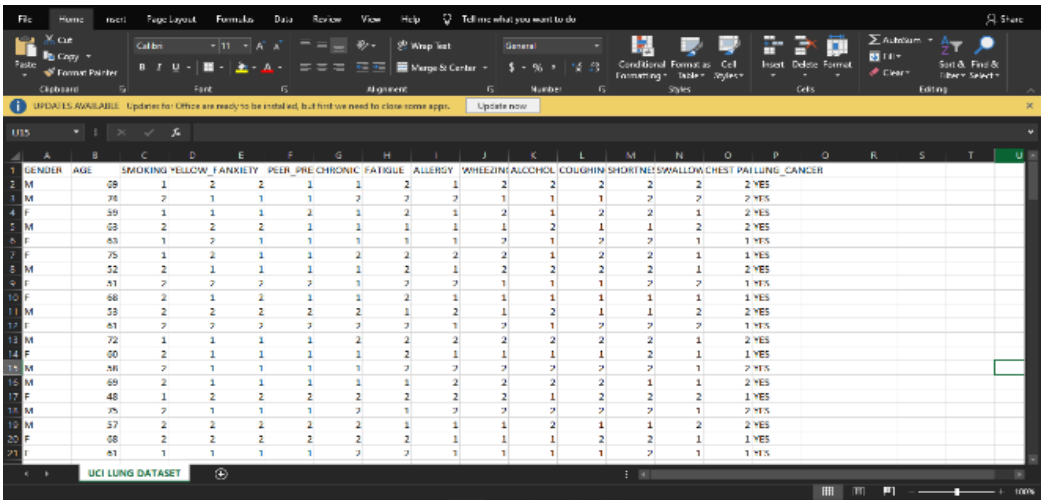

Fig 2. Dataset of Lung Cancer

3.2 Classification Techniques

Classification is a process that processes and categorizes input data into a certain category. The work proposed was carried out using the Weka tool. In the Weka tool, algorithms such as j48, KNN, Naïve Bias and RBF were used and a comparative analysis was eventually derived. In order to predict given input data for a certain class mark, classification falls under a supervised learning method.

3.2.1 Artificial Neural Network Classifier

The neural network is the basic principle of the machine learning approach in which the learning mechanism occurs between neurons. "The reference layer, the intermediate layer, and the output layer are part of artificial neural network (ANN). Through an adequate weight, each input neuron is connected to the hidden neuron and weight is evenly divided between the hidden unit and the output unit. The weighting methodology could be used either in the feed forward manner or input method to achieve the desired goal. Feed forward network techniques are easier ways to tackle classification.

3.2.2 Radial Basis Function Classifier

The radial base function method of neural networks is implemented with a radial threshold function. The RBF network has high solution stability and easy implementation.

3.2.3 Support Vector Machine Classifier

With various kernel features, it is possible to improve the Support vector classifier, which gives more accurate results. For both structured and unstructured data, Support Vector classifier classifier is the best choice. The help vector classification algorithm will overcome issues such as overfitting better than other models.

Since it provides quick and easy results, the help vector classifier (SVC) is often used in a shorter time period. This classifier employs judgment boundaries that are often referred to as hyperplanes. The data are classified into the goal categories according to the hyper plane. However, optimizing the interval between sample points and the judgment boundary is used for classification.

3.2.4 Logistic Regression Classifier

Logistic regression is built from and is based on statistics. A classifier is dependent on how likely the results are of classifying the data. Binary logistic regression is a popular method used in machine learning when working with binary input variables. The sigmoid function is used to categorize classes by how similar they are. Logistic Regression Classification Process Benefits.

3.2.5 J48 Classifier

J48 is a cut-off point for a c4.5 representation from Java in the Weka tool. To solve a dilemma, we apply the concept of tree concepts. A class mark is the leaf node in the tree, and the internal node specify the allocated name of the class. The classification is done based on the amount of knowledge recovered, and the type of which the information is useful. Decision tree model.

3.2.6 K Nearest Neighbor Classifier

The classifier is capable of being trained and tested using the same dataset or programmatically according to the preferences of the programmer. In the procedure, according to the majority class mark given as per k, the data of interest is extracted and analysed. The value of k depends on how the distance to be measured is measured. The option of k is dependent on statistical evidence. The higher the cost of k, the better the classifier. The selection of parameters is also a common practice which aids in minimising classification error.

4. Result And Simulation

Generally, the main process parameters for classification are Accuracy, Recall, Precision and F-Measure in the confusion matrix. "The accuracy of the classification is the calculation of the number of accurate predictions made out of the total prediction number. Some particular outcome depends on these parameters. Table 1 shows the result which is concluded from Weka Tool. and Fig. 3 shows the comparison between different classifiers.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad \ldots(i)$$

$$\text{Recall} = TP/(TP+FN) \quad \ldots(ii)$$

$$\text{Precision} = TP/(TP+FP) \quad \ldots(iii)$$

$$\text{F- Measure} = (2* \text{Recall}*\text{Precision})/ (\text{Recall} + \text{Precision}) \quad \ldots(iv)$$

Table 1. Classifier Output in Weka Tool

Classifier	Accuracy	Precision	Recall	F-measure
SVM	96.7	0.958	0.967	0.983
Naïve Bayes	96.07	0.952	0.961	0.956
Multilayer Perceptron	98.69	0.987	0.987	0.987
J48	96.7	0.967	0.967	0.983

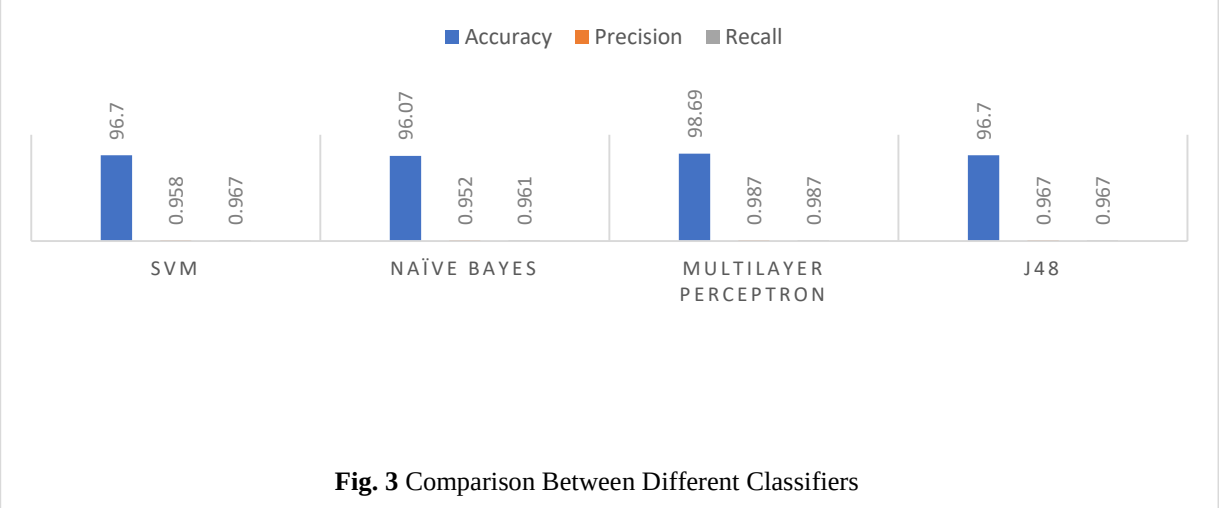

Fig. 3 Comparison Between Different Classifiers

5. Conclusion & Future Scope

We have shown in this paper that the accuracy is considered to be 98.69 percent with the MLP classifier. Compared with the current methodology, the experimental findings have been compared and the accuracy of classification is found to be superior with very fewer features.

The researchers will describe the different factors influencing the efficiency of the proposed biomedical signal classifier methods in the future. Consequently, the findings of this analysis are promising. These findings would also encourage other researchers to investigate further in this direction.

6. Acknowledgements

It is a matter of honor for us to express our gratitude towards all those who directly or indirectly assisted us and made our dissertation research possible .Authors would like to thank Ajay Kumar Garg Engineering College, Ghaziabad ,to provide the framework resources to accomplish your task.

References

[1] Radhanath patra “Prediction of Lung Cancer using Machine Learning Classifier”. Communication in Computer and Information Science book Series (CCIS,volume 1235).

[2] S. Akram, Muhammad Y j “ Artificial Neural Network based classification of Lungs Nodule Using Hybrid Feature from Computerized Tomographic Images”.Appl. Math.Inf.SCI.9,No.1,183-195(2015).

[3] Priyanka Basak and Ashoke Nath “Detection of different stages of Lung Cancer in CT scan images using Image processing technique”.June 2017.International Journal of Innovative Research IN computer and Communication Engineering5(5);9708-9719.

[4] M. Prabukumar L. Agilandeeswari & K. Ganesan “An Intelligent Lung cancer diagnosis system using cuckoo search optimization and support machine classifier”. Journal of Ambient Intelligence, Humanized Computing volume 10,pages 267-293(2019).

[5] J.Dhalia and H.khanna"Computer aided diagnosis of pulmonary hamartoma from CT scan images using ant coloy optimization based feature selection". Alexandria Engineering Journal Volume 57,Issue 3,September 2018,pages 1557-1567.

[6] Janee Alam , S.Alam and H. Alamgir"Multi-Stage Lung Cancer Detection and prediction using Multiclass SVM classifier".International Conference on Computer , Communication, Chemical,Material and Electronic Engineering(IC4ME2)18098246 (2018).

[7] B. Muthazhagan T.Ravi & D. Rajnigirinath "Enhanced computer-assisted lung cancer detection method using content- based image retrieval and data mining technique".Journal of Ambient Intelligence and Humanized Computing (2020).