LUNG CANCER PREDICTION

Sehag A
Computer Science, B.Tech
PES University
PES2UG20CS457
Bangalore, Karnataka
sehag.amar.18@gmail.com

Shreyas Sai Raman
Computer Science, B.Tech
PES University
PES2UG20CS461
Bangalore, Karnataka
shreyasraman53@gmail.com

***Abstract*—Lung cancer is affected to a wide range of people. Prediction of lung cancer at early stages can increase the chances of recovery. People diagnosed with lung cancer tend to suffer from a lot of symptoms that can be caused otherwise. Due to this reason, patients neglect these symptoms and do not go for a diagnosis until the symptoms turn severe.Early diagnosis of lung cancer saves an enormous amount of life, or else serious problems could arise, causing sudden lethal end. Cure rate and prediction of lung cancer is primarily dependent on early identification and diagnosis of the disease. Hence a model which can predict the chances of lung cancer can help people take precautionary measures at early stages and fight the cancer.**

Keywords — Lung cancer, Random Forest classifier, K-Nearest Neighbour(KNN) , Recall, Support Vector Machine (SVM).

I. Introduction

Lung cancer is a type of cancer that progresses in the lungs. It ranks among the number one cause of death from cancer. The nature of lung disease is very complex and hence, the disease must be handled correctly. Lung cancer can be classified into two types, Small Cell Lung Cancer (SCLC) and Non -Small Cell Lung Cancer (NSCLC). NSCLC accounts for 85 percent of all lung cancers. Radiation Therapy and Chemotherapy have always been the most widely used treatments for NSCLC. The process of determining the extent of lung cancer spread is known as staging. TNM staging information describes the amount of cancer as well as where it occurs in the body. This information assists the doctor in selecting the best possible treatments. Research shows, however, that the prediction system performs poorly when only staging information is considered. One of the main reasons why lung cancer survival is poor is late diagnosis.“Fig. 1” [1] shows the survival rates of patients of different age groups. If a person realizes that he/she might be suffering from lung cancer at an early stages, their chances of survival can be increased by a lot. We need a prediction system that can diagnose a person at early stages based on the symptoms observed in that person. This prediction is based on past data collected from similar patients. The algorithms used for this disease are classified based on various methods like K- Nearest Neighbor(KNN) classifier, Random forest classier, Support Vector Machine(SVM) classier.

II. Previous Work

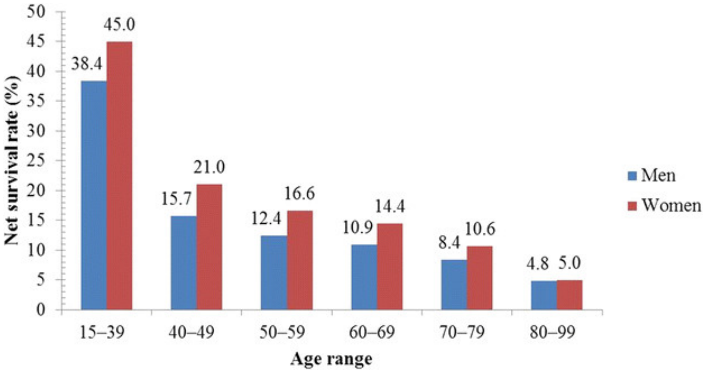

Fig. 1. 5 year net rate of survival of patients with lung cancer in the United Kingdom.

Literature survey is a very important method before we start analysing our data and make predictions.As a part of our literature survey, we reviewed some research papers which have in common with our own approach or the areas of interest under study.In [2],a multi-class lung cancer classifier was developed which classified based on CT scan images.In every stage of classification image enhancement and segmentation have been done separately.The enhancement of images has been achieved by scaling, color space transformation, and contrast enhancement. The proposed algorithm detected 87 cancerous images and 13 non-cancerous images out of 100 cancer-affected images, yielding an accuracy of 87%. With the help of this paper, we are able to detect lung cancer through images and hence take the right treatment for it.

In [3], bagging ensemble method was used to create a model in order to improve the overall prediction accuracy.The bootstrap aggregating technique was found to boost the performance of individual models. The integrated model has an accuracy score of 0.98. This paper helped us choose the algorithms which we can use to make predictions by enhancing accuracy by 3.33%.

In [4], a prediction system was designed using data mining classification techniques. There are huge amounts of healthcare data collected by the healthcare industry, but they aren’t mined to discover hidden information. In many cases, hidden patterns and relationships remain unexplored.Unlike traditional decision support systems, Lung Cancer Disease diagnosis can answer complex ”what if” questions using generic lung cancer symptoms like wheezing, breathlessness, and shoulder, chest, and arm pain, which can predict the risk of patients developing lung cancer. This research helped us to choose our independent variables based on which we can make predictions.If we make predictions based on the early symptoms,we can detect cancer at an early stage and increase the chance of survival by taking appropriate measures to cut down the cancer.

III. Proposed solution

We have chosen a dataset in which the symptoms are the independent variables. Our dataset contains a total of 16 attributes. The 16 attributes are gender, age, smoking, yellow fingers, anxiety,peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, chest pain and lung cancer. The attribute ”lung cancer” is the dependent variable. All the attribute values except ”age” attribute are categorical (Yes/No, where Yes=2, No=1). Age is the only continuous attribute.

A. Preprocessing of data

Before we start analyzing our data, we need to make sure that our data is clean i.e, our data does not contain any null values and duplicated values. These values can impact our results highly, thus making our predictions very inaccurate. Hence, duplicated values are removed and null values are either removed or replaced with the mean column value, since our data is categorical, we decided to remove the null values too. We also converted categorical labels into numerical values. This helps us to understand and interpret the trends between the attributes.

B. Exploratory data analysis

EDA helps us to analyze the data using visual techniques. It is used to uncover trends, patterns, or to verify hypothesis with the help of statistical summaries and graphic representations.“Fig. 2” illustrates that most attributes appear to be reasonably represented between their 0(NO) and 1(YES) labels.The dataset is unbalanced and the depiction of ’LUNG_CANCER’ 1(YES) is much higher than that of ’LUNG_CANCER’ 0(NO).

We must examine whether the independent variables are correlated with one another. In general, the uncorrelated variables are best dealt with. If two variables are in correlation, they could describe a similar feature about the data set, and their inclusion could cause an overweight to this generalized effect. It’s not always a bad idea to include correlated features, it’s the data scientist’s decision. After analysing the correlation matrix, we observed that there were no two strongly correlated attributes, which is a good thing, and no action was taken.

C. Model selection

Before we run our data through different models,we need to normalize the data.Since age is the only continuous attribute,we need to standardize it.We standardize the data after we split the data into training data and testing data because we fit and transform the training data so that we learn the parameters of scaling on the training data and in the same

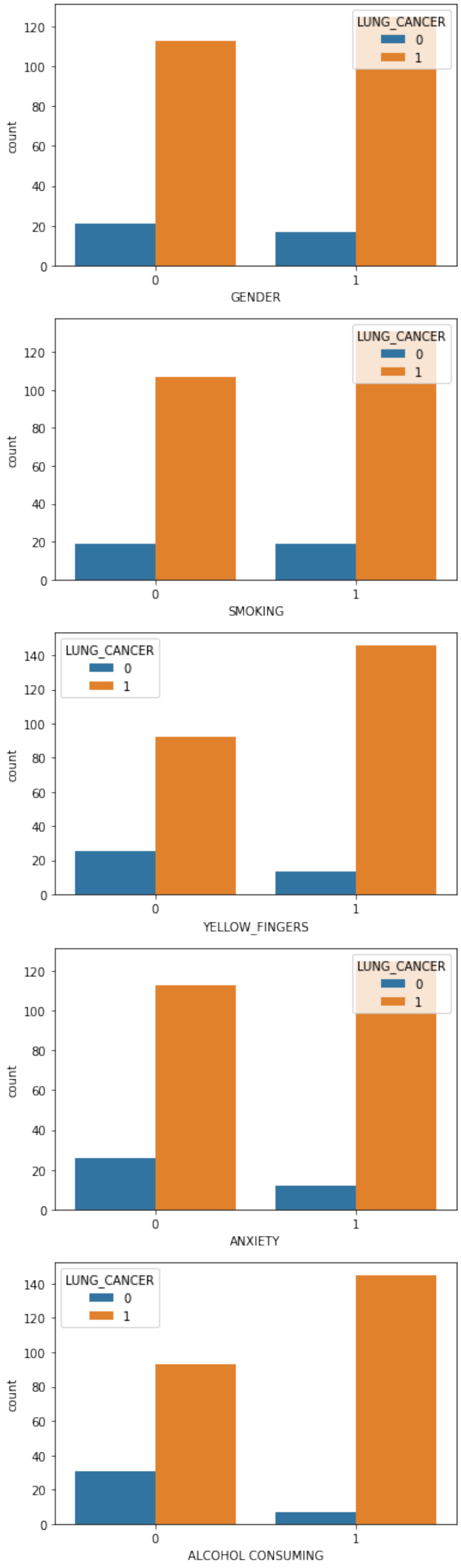

Fig. 2. Categorical attributes VS target variable count.

time we scale the training data but only transform the testing data because we use the scaling parameters learned on the training data to scale the testing data.

Let's choose the models that work best for binary classification. The models that we will be using are: K-nearest neighbors classifier, Random Forest classifier, Support Vector Machines classifier. We need to evaluate these 3 models and decide the most appropriate model for this dataset. There are several metrics based on which we can evaluate the performance of a ML model. Precision, recall and accuracy are the usual metrics. Precision tells us, of the samples that we predicted to be positive, how many were actually positive. Recall tells us, of the samples that could have been positive, how many did we correctly predict to be positive. Accuracy gives us an idea of how effective the model is overall, but doesn't give us a metric to account for false negatives.

a) K-nearest neighbours classifier: The k-nearest neighbours algorithm, also known as KNN, is a non-parametric, supervised learning classifier that uses proximity to classify or predict the grouping of an individual data point. Initially, we calculated the recall scores for every neighbour(k) count. Maximum recall is 0.89 for k=1. Therefore, we build a KNN model with k=1. The precision, recall for 'LUNG_CANCER' 1 and accuracy are as follows:74%,89.8%,79%.

b) Random forest classifier: Random forests are an ensemble learning method that constructs a large number of decision trees during the training phase for classification, regression, and other tasks. When it comes to classification tasks, the random forest output is the class selected by the large percentage of trees. Random decision forests mitigate the tendency of decision trees to overfit to their training set. Random forests generally perform more accurately than decision trees. Random forest classifier showed a precision of 92%, recall of 95% and an accuracy of 93%. "Fig. 3" shows the confusion matrix. We observe that almost all the instances are classified accurately.

c) Support vector machine classifier: Support vector machines are supervised learning models that use learning algorithms to analyse data for classification and regression. SVMs are among the most reliable prediction methods because they are based on statistical learning frameworks. The goal of the SVM algorithm is to find a hyperplane in an N-dimensional space that precisely categorises the data points. The dimensions of the hyperplane is determined by the total number of features. After training the model, we tested the model against the test data.The results were quite disappointing as SVM classified all the instances as positive.The recall of the model is 100% but its precision and accuracy is 50%.

D. Results

Precision and accuracy are the most commonly used metrics for evaluation of models. But for our data, we will be making use of recall. The number of correct positive predictions made out of all possible positive predictions is measured by recall. Precision only measures the correct positive predictions out of all positive predictions but recall indicates missed positive predictions.

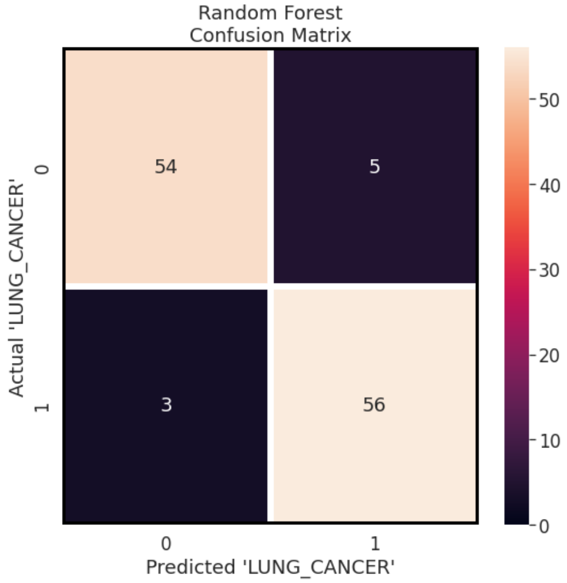

Fig. 3. Confusion matrix for Random forest classifier.

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \quad (1)$$

$$Recall = TP/(TP + FN) \quad (2)$$

Recall takes precedence in reducing the number of false negatives, which are positive cases misclassified as negatives by the model. As a necessary consequence, it is essential in mission-critical applications where a false negative could result in loss of life. It is critical to maximise recall in such applications.

In the most extreme case, perfect recall can be obtained by classifying all cases as positive. This would ensure that no false negatives are produced, but it could result in a large number of false positives, or negative cases misclassified as positives by the model. Typically, recall is combined with other metrics such as false positive rate and precision to quantify the trade-off between false negatives and false positives.

We visualised the comparison between the 3 chosen models based on accuracy,recall and average recall as shown in "Fig. 4".In the case of the SVM, it has simply classified everything as positive, so it has the best Recall when only that class is considered. This makes evaluating the overall Accuracy of the models a relevant secondary step. So if we also consider Accuracy, the best models are the Random Forest Classifier and k-nearest neighbours classifier.

Additionally, if we consider Average Recall, or recall for both classes of 'LUNG_CANCER' - 'YES' and 'NO', the Random Forest model is the clear winner. Random forest classifier was able to predict 110 correct cases out of 118. Area under the curve(AUC) is the best metric to evaluate a binary classification model.AUC is the measure of the ability

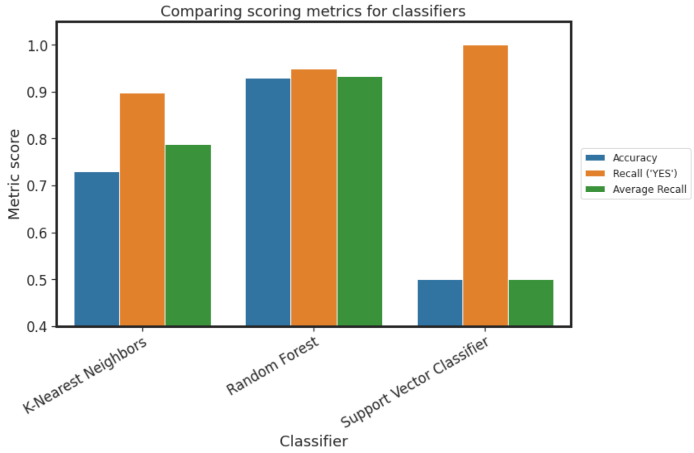

Fig. 4. Metric comparison between the models.

of a classifier to distinguish between classes. Greater the AUC, more accurate the performance of the model at different threshold points between positive and negative classes. This simply means that when AUC is equal to 1, the classifier can distinguish between all Positive and Negative class points perfectly. When AUC is zero, the classifier predicts all Negatives as Positives and vice versa. When AUC is 0.5, the classifier is not able to differentiate between the Positive and Negative classes. Our Random forest classifier has an AUC value of 0.93 which is a very good value.

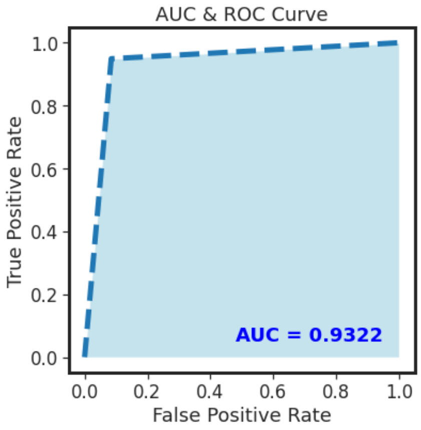

Fig. 5. AUC curve of Random forest classifier.

IV. Discussions

We went through peer review to assess the quality of a manuscript before it is published. It was suggested to use appropriate correlation analysis method since our dataset consisted of only binary categorical values. Pearson's correlation works best on numerical data,point-biserial correlation is used between one dichotomous variable and one continuous variable. When both variables are dichotomous, the best method would be phi-coefficient.

Another issue that was addressed was that our dataset consisted of one continuous variable. Hence, it had to be standardized accordingly for the model to predict accurately.

V. Conclusion

In this paper,a random forest classifier which can predict lung cancer based on the symptoms is presented. We can predict the presence of lung cancer at early stages based on the symptoms expressed by a person. This helps to take immediate action against the cancer and increase the chances of survival. Additionally, the results obtained showed that the suggested methodology is promising in terms of accuracy and recall in predicting lung cancer efficiently and accurately. Even though the metrics are high, there are still few misclassified cases. It cannot be tolerated in health-critical system as it could cost life. Unfortunately, our model misclassified 3 out of 118 cases as not suffering from lung cancer but are actually suffering from lung cancer, this is not acceptable. We hope to make more accurate model which can classify all the cases correctly. Our model fails to differentiate severe symptoms from generic symptoms as all the independent variables are given equal weights. Our model fails if the person is not sure about the symptoms. If a person is clearly sure about the symptoms they are suffering from, our model is highly accurate in predicting the outcome. The proposed methodology for predicting lung cancer is general and is easily adaptable to other diseases' prediction models.

Contributions

Shreyas Sai Raman : Literature survey and EDA
Sehag A : Model selection and evaluation

References

[1] Epidemiology of lung cancer and approaches for its prediction: A systematic review and analysis - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Five-year-net-survival-rate-of-lung-cancer-patients-by-age-in-the-United-Kingdom-UK_fig1_305722874 [accessed 10 Nov, 2022]

[2] J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465593.

[3] Reddy, DendiGayathri Kumar, Emmidi Reddy, DesireddyLohithSaiCharan P, Monika. (2019). Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method. 353-357. 10.1109/ICAIT47043.2019.8987295.

[4] Krishnaiah, V. V. Jayarama et al. "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques." (2013).