Lung Cancer Detection using Machine Learning Algorithms and Neural Network on a Conducted Survey Dataset Lung Cancer Detection

Ratika
Student
Master of computer applications
Graphic era hill university, Dehradun, India

Nisha Gupta
Student
Master of computer applications
Graphic era hill university, Dehradun, India

Abstract:- Lung cancer is the expansion of malignant cells in the lungs. Due to the rising frequency of cancer, both the death rate for men and women has increased. Lung cancer is a condition in which lung cells proliferate uncontrolled. Although lung cancer cannot be averted, the risk can be decreased. Therefore, early identification of lung cancer is essential for improving patient survival. Lung cancer incidence is directly inversely correlated with the frequency of heavy smokers. Various classification techniques, including Naive Bayes, Random forest, Logistic Regression, Knn, Kernal svm and Artificial neural network were used to investigate the lung cancer prediction. The primary goal of this study is to investigate the effectiveness of classification algorithms and neaural network in the early identification of lung cancer.

Keywords:- *Naive Bayes, Random Forest, Logistic Regression, Knn , Kernal svm, Artificial Neaural Network ,Machine Learning, Lung Cancer.*

I. INTRODUCTION

Lung Cancer is the most treacherous disease for human beings. Lung cancer is responsible for more deaths than combined death count of colon, prostate, ovarian and breast cancer . Lung cancer is a serious health concern for humans and alone in the United States of America with a count of 225,000 people each year . The main factor causing lung cancer is smoking and the duration of smoking is directly proportional to the person getting affected with cancer. To detect lung cancer manually is a very tedious and risky job even for specialists. To gain deeper insights and identification of lung cancer in early stages, different machine leaning methods are used in classification. By applying techniques such as random forest and other classification algorithms, an automated system can be built which can perform with higher accuracy rate and helps in accurate classification.

lung cancer is the leading cause of cancer death in both men and women in the United States. The main objective of this paper is to analyze the lung cancer data available models to lung cancer survivability prediction model and to develop accurate survival prediction models using Machine Learning. Logistic regression,naïve byes, knn ,Random Forest (RF) ,Kernal svm, Artificial neaural network have been applied for constructing a lung cancer survivability prediction model. The classifiers developed in this work predicted the various factors that influence the survival time, would help doctors make more informed decisions about treatment plans and help patients develop more educated decisions about different treatment options.

This study has explained the survival rate analysis of patients with advanced lung cancer who did not receive any type of therapeutic modality and to evaluating performance scores daily activities the results of this study have found slight improvement in survival rates. Random Forest algorithms were found to result in the good prediction performance in terms of accuracy of 88% and Artificial neaural network were found in the best prediction giving accuracy of 89%.

II. LITERATURE REVIEW

In paper [11], Pankaj Nanglia, Sumit Kumar, and others introduced a novel hybrid technique known as the Kernel Attribute Selected Classifier, in which they integrate SVM with Feed-Forward Back Propagation Neural Network, assisting in lowering the computational complexity of the classification. They suggested three block processes for the classification, processed the Block 1 is the dataset. The first block involves feature extraction using the SURF method, the second block involves optimization using a genetic algorithm, and the third block involves classification using FFBPNN.

- Chao Zhang, Xing Sun, Kang Dang, and others use the multicenter data set to conduct a sensitivity analysis in paper [12]. The two categories they selected were Diameter and Pathological outcome.
- In paper [18] K.Mohanambal , Y.Nirosha et al studied structural co-occurrence matrix (SCM) to extract the feature from the images and based on these features categorized them into malignant or benign. The SVM classifier is used to classify the lung nodule according to their malignancy level (1 to 5).
- Radhika P. R. and Rakhi. A. S. Nair's paper [16] primarily focused on the prediction and categorization of medical imaging data. They made use of the data.world dataset and the UCI Machine Learning Repository. Support vector machines had superior accuracy (99.2%), according to a comparative research using several machine learning algorithms. Naive Bayes provides 10%, Decision Tree

provides 80% 87.87% and 66.7% are provided via logistic regression.

- The algorithm for lung cancer detection was examined in paper [17] by Vaishnavi. D1, Arya. K. S2, Devi Abirami. T3, and M. N. Kavitha4. They used the discretely sampled Dual-tree Complex Wavelet Transform (DTCWT) for pre-processing. The second order statistical texture analysis approach known as GLCM provides a table of the co-occurrence of various combinations of Gray levels in an image.

III. METHODOLOGY

The total economic development of a developing country, such as India, where the majority of the population depends on health, is scared of lung cancer. Therefore, lung cancer detection ought to be more precise and reliable. The open source is used to gather lung cancer parameters. Python is the programming language in use.

Numerous parameters, such as smoking, anxiety, peer pressure, chronic disease, fatigue, allergy, alcohol consuming, etc., are used to predict the lung cancer. The user starts activity in this system by using lung cancer dataset. Data gathered from the user during data collection and pre-processing processes is utilized .The initialization data is then analyzed and splitted into training and testing dataset then the model is fitted into the dataset, which evaluates the dataset and give accuracy to the user.

The system is shown as a block diagram :

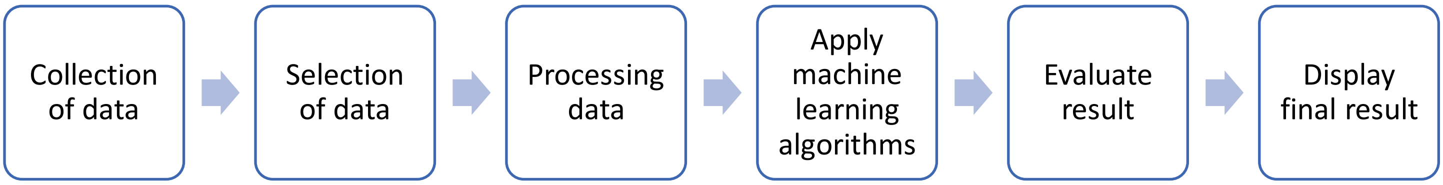


Fig.1 System Block Diagram

➢ *Logistic regression*

The logistic function, often known as the sigmoid function, is used in this method. This S-shaped curve can assign any real value number to a value between 0 and 1, but never exactly within those bounds. Logistic regression so models the default class probability. The logistic function, which enables us to compute the log-odds or the probit, is used to predict the likelihood. As a result, the inputs are combined linearly to create the model, but this linear combination is related to the log-odds of the default class.

➢ *K-Nearest neighbors*

K-Nearest Neighbours is a strategy that classifies new cases based on similarity measures and stores all of the existing examples. The test phase made use of all training data. This accelerates training while slowing down and increasing the expense of the test phase. If there are two classes, the number of neighbours in this method, k, is typically an odd number. Use distance measures such the Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance to calculate the distance between points in order to determine the ones that are the closest in similarity.

➢ *The naive Bayes*

The naïve Bayesian classifier is a probabilistic classifier built on the foundation of the Bayes theorem and has significant assumptions about the independence of the features. As a result, by applying the Bayes theorem, P(X|Y)=P(Y|X)P(X)P(Y), we may determine the likelihood that X will occur given that Y has already occurred. The evidence in this case is Y, and the hypothesis is X. Here, it is assumed that each predictor or trait is independent and that its presence has no effect on the others. The term "naive" is a result. In this instance, we'll assume that the values were drawn from a Gaussian distribution, leading us to think of a Gaussian Naive Bayes.

➢ *Random forest*

This forest is made up of a number of decision trees that were frequently trained using the bagging approach. The fundamental concept of bagging is to reduce variation by averaging numerous noisy but roughly impartial models.

➢ *Kernel Svm*

Using a kernel function, data can be input and then transformed into the format needed for processing. The term "kernel" is employed because the window for manipulating the data in a Support Vector Machine is provided by a set of mathematical operations. In order for a non-linear decision surface to turn into a linear equation in a higher number of dimension spaces, Kernel Function often changes the training set of data. The inner product between two points in a common feature dimension is what it basically returns.

➢ *Artificial neural networks (ANNs)*

A class of machine learning techniques known as artificial neural networks (ANNs) are modelled after the form and operation of biological neural networks seen in the human brain. Artificial neurons (ANNs) are made up of interconnected nodes, also referred to as "units," that are arranged in layers. An input layer, one or more hidden layers, and an output layer are the typical divisions of the layers.An overview of how an artificial neural network functions is given below:

- Input Layer: The neural network's initial data or training features are delivered to the input layer. Each input neuron is associated with a certain characteristic or aspect of the data.

- Hidden Layers: There may be one or more hidden layers between the input and output layers. Multiple synthetic neurons or units are present in each hidden layer, processing data and transmitting it to subsequent layers.
- Weights and Bias: Each neuronal link inside the network has a corresponding weight. These weights are modified during the training phase to enhance the performance of the network. Each neuron also has a bias, which can be thought of as an activation threshold.
- Activation Function: A neuron's output is determined by its inputs and internal state by the activation function. Sigmoid, ReLU (Rectified Linear Unit), and tanh (hyperbolic tangent) are often used activation functions. They give the network non-linearities, which help it learn intricate patterns.
- Loss Function: A loss function evaluates the discrepancy between the neural network's output and the predicted output. Whether regression or classification is being used to solve the problem will determine the loss function that is used.
- Backpropagation: The primary algorithm used to train the neural network is backpropagation. In order to minimise the loss, it calculates the gradient of the loss function with respect to the network weights and modifies the weights in the opposite direction of the gradient. Usually, optimisation methods like stochastic gradient descent (SGD) or its variations are used for this process.
- Training: The neural network is trained by supplying training examples to the network periodically, modifying the weights via backpropagation, and optimising the loss function. The aim is to reduce the loss and enhance the forecast accuracy of the network.
- Prediction: After the neural network has been trained, predictions can be made using brand-new, unexplored data. Forward propagation is used to feed the input data through the network, and the output layer delivers the anticipated outcome.

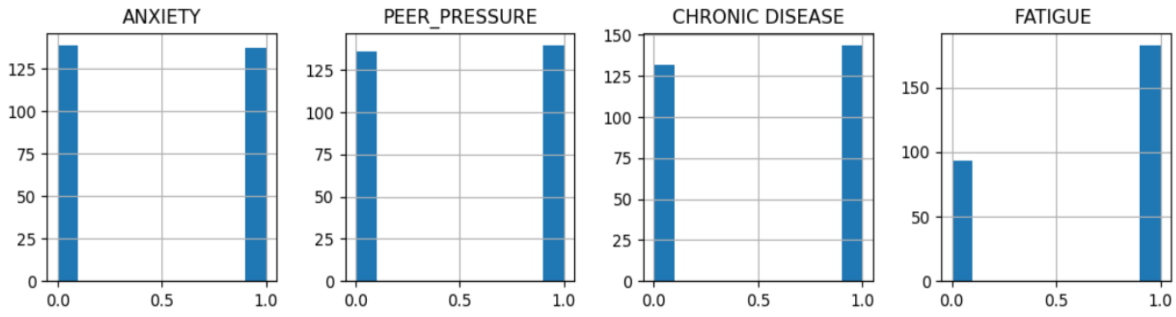


Fig.2 parameters affecting lung cancer diagram

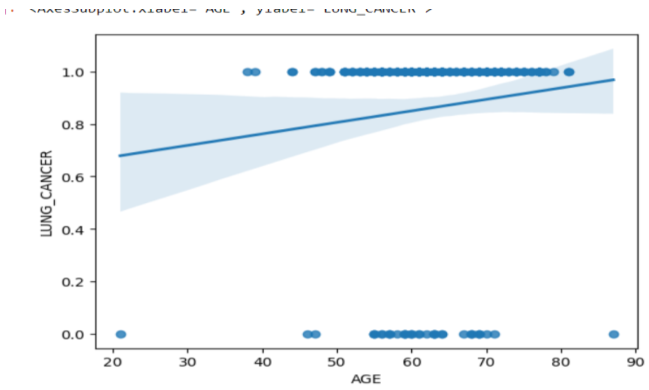


Fig3. lung cancer-age diagram

This graph shows persons having age 50 above is having lung cancer nowadays which is great in number.

IV. RESULT AND DISCUSSION

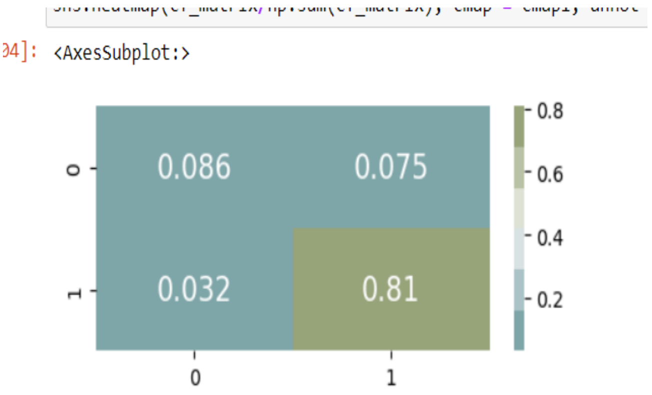
The dataset was trained and the random forest model achieved a training accuracy of 88% and Artificial neural network gives accuracy of 89% which is highest then any other model. The table shows the accuracy achieved by all other models:

Table 1 the accuracy achieved by all other models:

MODEL	LOGISTIC REGRESSION	KNN	RANDOM FOREST	NAÏVE BYES	KERNEL SVM	ANN
ACCURACY	87%	86%	88%	83%	84%	89%

ISSN No:-2456-2165

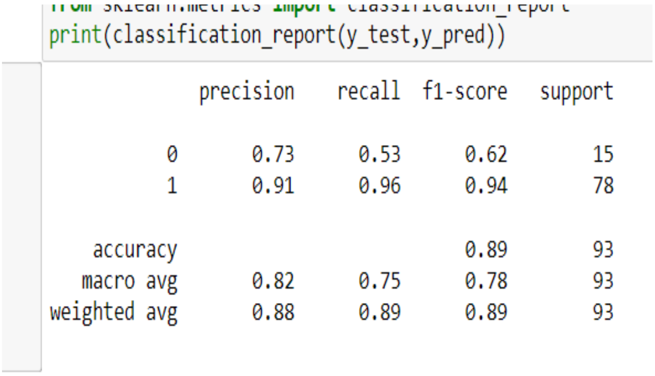
The confusion matrix is given as: depicts the true label vs. the predicted label



```
sns.heatmap(cf_matrix/np.sum(cf_matrix), cmap = cmap1, annot
34]: <AxesSubplot:>
```

Fig 4. confusion matrix

The figure shows the precision, recall, f1-score and support for the different categories.



```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))

              precision    recall  f1-score   support

           0       0.73      0.53      0.62        15
           1       0.91      0.96      0.94        78

    accuracy                           0.89        93
   macro avg       0.82      0.75      0.78        93
weighted avg       0.88      0.89      0.89        93
```

Fig 5 precision, recall, f1-score

V. CONCLUSION AND FUTURE ENHANCEMENTS

To conclude this research the lung cancer features were classified with high accuracy and with limited computation power. The preprocessing of the data was done efficiently which helped the model for less time consumption. In the end of the research comparative study was done to asses the quality of results. The random forest model obtaining the accuracy of 88% gives a quality result and ANN gives accuracy of 89%. As well as it was observed that Naïve byes has the lowest achieving accuracy of 83%. This makes ANN an efficient neural network in terms of accuracy .

In the future work the lung cancer detection can be done on imaging format which can build a 4D image structure such as 4D MRI. This can used for accurate segmentation of the dataset and which can help to detetect the lung cancer more accurately.

Second application of this research could be used for full scaled system for assistance to the radiologists and doctors for better decision making. In future work, more numbers of datasets and parameters should be taken into consideration which can benefit the classifiers.

REFERENCES

[1]. SRS Chakravarthy and H. Rajaguru. "Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm." Asian Pacific Journal of Cancer Prevention, 20, 7, 2019, 2159-2166, doi: 10.31557/APJCP.2019.20.7.2159.

[2]. AA. Borkowski, MM. Bui, LB. Thomas, CP. Wilson, LA. DeLand, SM. Mastorides. "Lung and Colon Cancer Histopathological Image Dataset." (LC25000). ArXiv: 1912.12142v1 [eess.IV], 2019.

[3]. W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach," 2018 11th Biomedical Engineering International Conference (BMEiCON), Chiang Mai, 2018, pp. 1-5, doi: 10.1109/BMEiCON.2018.8609997.

[4]. K. Yu, C. Zhang, G. Berry, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nat Commun 7, 12474 (2016), doi: 10.1038/ncomms12474

[5]. G. A. Silvestri, et al. "Noninvasive staging of non-small cell lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition)." Chest vol. 132, 3 Suppl (2007): 178S-201S. doi:10.1378/chest.07-1360.

[6]. https://www.cdc.gov/cancer/lung/basic_info/symptoms.htm

[7]. https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection

[8]. https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620

[9]. https://www.datacamp.com/blog/classification-machine-learning

[10]. M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.8783041.