Lung Cancer Prediction Using Machine Learning Methodologies

By Karan Bhadja, Dhruv Khandelwal, Devang Chaudhari

SRM University of Science and Technology,

Chennai, Tamil Nadu.

***Abstract*—**Lung cancer, an ailment affecting both men and women, arises from uncontrolled growth of lung cells, leading to critical respiratory complications during inhalation and exhalation. The primary culprits behind this global health menace are cigarette smoking and tobacco smoke. The fatality rate associated with lung cancer, across different age groups, is soaring, surpassing that of other types of cancer. Despite the presence of advanced medical facilities and effective treatments, effective monitoring of mortality rates remains a challenge. As a result, early preventive measures are imperative to identify symptoms and effects at the outset, enabling accurate diagnosis. In the realm of healthcare, the application of machine learning is gaining ground, owing to its robust computational capabilities for timely disease detection and comprehensive data analysis. Our research delved into a range of machine learning techniques to categorize available lung cancer data through a developed malignancy UCI process. The input data underwent preprocessing and conversion into binary format, followed by the application of a well-established Weka classifier to distinguish between cancerous and non-cancerous instances. Our comparative analysis revealed that the proposed RBF classifier exhibited a remarkable accuracy of 81.25 percent, establishing itself as an efficient method for predicting the prognosis of lung cancer**. *Keywords—{Lung cancer, Respiratory problems, Cigarette smoking, Tobacco smoke, Mortality rate, Early diagnosis, Machine learning, Data analysis, Malignancy UCI process, Weka classifiers, RBF classifier, Prognostication}***

I. Introduction

Lung cancer represents a critical global health concern, ranking 7th in the global mortality index with an estimated 1.5% of the world's overall mortality [2-7]. Two primary types of lung cancer exist, namely non-small cell lung cancer and small cell lung cancer, both of which manifest various symptoms including chest pain, dry cough, respiratory complications, and weight loss. Emphasizing the cultivation and etiology of cancer, medical practitioners primarily focus on smoking and second-hand smoke as major contributors to lung cancer. Treatment modalities encompass surgical procedures, chemotherapy, radiation therapy, and immunotherapy, yet early diagnosis remains crucial to curbing mortality rates.

Survival rates for lung cancer patients demonstrate significant variability, contingent upon age, gender, race, and overall health. Machine learning has emerged as a crucial tool in early detection and prognosis of medical conditions, simplifying and streamlining the diagnostic process. Its applications in disease detection, feature extraction, and precise image analysis have transformed the healthcare sector globally. By enabling accurate disease prediction and timely intervention, machine learning facilitates proactive healthcare management.

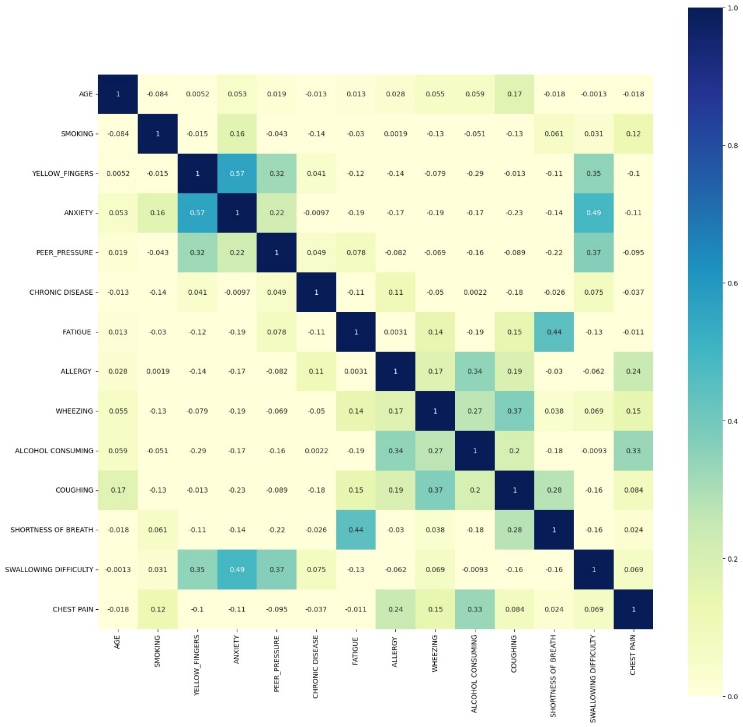
The diverse applications of machine learning encompass the optimization of drug development, more accurate disease prediction, and the regulation of early disease outbreaks for prompt preventive measures. To ensure greater standardization and reliability, the development of more sophisticated machine learning applications is imperative, empowering healthcare professionals to make precise clinical decisions. Employing various machine learning techniques, including unsupervised, supervised, and reinforced learning, the healthcare sector can leverage advanced data-driven approaches to enhance disease management. Notably, the classification process using the Weka tool contributes significantly to this domain [27-29].

II. Related Work

*Hosseinzade et al. (2013) advocated the use of SVM for selecting specific protein characteristics, yielding an 88% accuracy in lung cancer prediction, outperforming other methods [13-16]. The Northern Centralized Cancer Group (NCCTG) study by C4.5, Naveen, and Pradeep (2018) demonstrated the superiority of C4.5 classification over SVM and Naive Bayes in cases with substantial data, showcasing improved lung cancer classification [25]. Gur Amrit Pal Singh and P.K. Gupta (2018) introduced an innovative method for visual data extraction and employed machine learning techniques to enhance precision [27]. Hussein et al. (2019) proposed the use of 3D CNN and supervised SVM for the categorization of benign and malignant data, achieving an impressive 91% accuracy [12]. Monkam et al. (2019) underscored the significance of Convolutional Neural Network in pulmonary nodule prediction, achieving a precision rate of approximately 90% [21]. Asuntha and Andy Srinivasan (2019) highlighted the optimization of fluorescent particle swarm with a deep neural network, achieving a remarkable 99.2% accuracy in lung cancer image analysis. Ganggayah et al. (2019) conducted a comprehensive analysis using 8066 records and 23 predictors for breast cancer, revealing that random forest classification outperformed other methods with an 82% accuracy [9]. Gibbons et al. (2019) demonstrated the superiority of SVM over linear regression model, vector support machine, and ANN, achieving a 96% accuracy in the supervised learning setting [18]. Shakeel et al. (2019) utilized a novel hybrid selection method in ANN for the prediction of lung cancer data, achieving an accuracy of 99.6% based on the ELVIRA biomedical data [26]. Bhuvaneswari et al. (2015) employed gabor filters for feature extraction and the G-knn method for the classification of lung cancer images, attaining a 90% accuracy rate [7].*

III. Data Set

In the UCI Machine Repository study, a dataset comprising 32 instances with 57 features was accessible. These features include 0-3 predictive attributes and 3 class attributes, with 1 class attribute and 56 inputs. To facilitate data processing, both the nominal features and class label data were converted into binary form, representing a widely adopted standard procedure for data processing. However, the presence of certain missing values within the dataset was found to impact the algorithm's performance, underscoring the importance of thorough data validation prior to analysis. The labels 'high,' 'low,' 'middle,' and 'medium' were encoded as 2, 0.4, and 1, respectively, as outlined in the study.



IV. Technical Classification

Segmentation involves a supervised learning approach aimed at predicting specific class labels based on input data. The efficacy of classification hinges upon the precise mapping of data. Various classification techniques, such as Perceptron, Naïve Bayes, Decision Tree, Logistics Regression, K-Nearest Neighbours, Artificial Neural Networks, and Support Vector Machines, are fundamental components of machine learning, serving as pivotal tools for data analysis and decision-making [20-23].

### The core focus of our study revolves around introducing a novel methodology for the precise analysis of lung cancer data. The paper highlights several commonly employed classification techniques.

### Neural Network Concept Within machine learning, the neural network serves as a foundational element that simulates the behaviour of neurons. The Artificial Neural Network (ANN) comprises an input layer, hidden neurons in the intermediary layer, and an output layer. Each neuron's inputs are linked to the hidden neurons with specific weights and also connected to the output unit through the hidden unit. Processing within the neuron involves the application of a predetermined activation function and threshold value, facilitating the treatment of the neuron as required. The synaptic weight, when multiplied by the corresponding neuron, aids in classification within the hidden layer and output layer. The ultimate objective is achieved through the adjustment of the desired output using a technique of weight modification. Employing network feedback techniques contributes to a more streamlined classification process.

### A. Network of Radial Base Functions

The radial function network is component of a neural network, whose threshold function is its radial base function. The RBF network has a high input noise tolerance and is simply constructed. The radio base function is characterised by a feed architecture consisting of an intermediate layer between the input and the output layer. It uses a number of basic functions centred on each sample point. The network output may be expressly specified for an input x (Fig. 1).

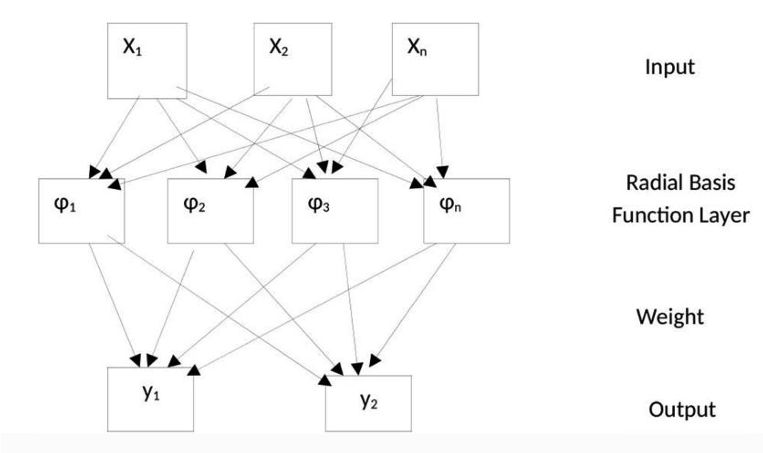


Fig. 1. RBF

### The weight of the input and the medium layer within the RBF design is the central part of the corresponding neuron, which uses weights to train the network connected to the mid layer and the output layer.

### B. Vector Classifier Support

Support for vector classification is one of the key and effective methods for supervised learning. The support vector classification (SVC) is frequently selected because of its computer capability for data processing in short timeframes. This categorization is based on the concept of the decision limit Known as a hyper-flat. The hyperplane is used to categorise the data entry into the target group needed. However, the greatest distance of the aircraft from data points to the categorization determination border is selected. The user-defined vector classifier may be framed to improve accuracy using different kernel functions. Vector classification assistance is appropriate for organised and unstructured data. Vector classification system support is not affected and is more reliable.

### C. Classifier of Logistic Regression

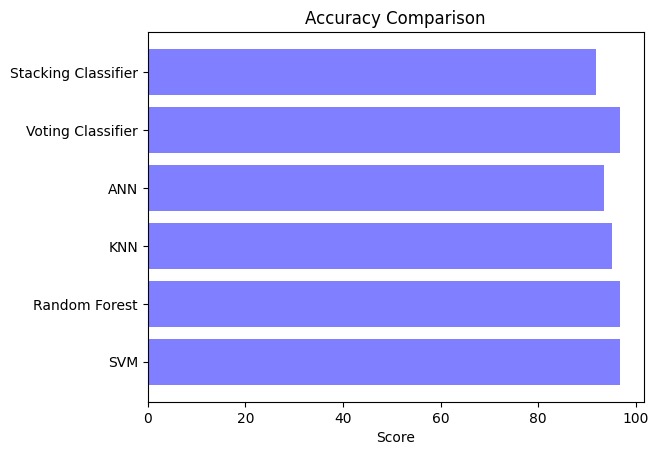
The classification of logistic regression is obtained using statistics. These classifiers are based on the likelihood of outcomes of data input. Binary logistic regression is typically used to handle binary input variables in the technology of machine learning. The class is divided into a specific category of sigmoid function.

### D. Random Classifier of Forest

The combination of the categorization of trees is a random classification of forests. One of the best ways to express input variables as trees that create a forest-like structure. Input Data is shown in the trees with a class name defining each tree. Random forests are dependent on their rate of error. Error rate implies many directions. The first is the relationship between trees; the other is the tree's strength. Knn Knn classifier is a slow learning technique where training and testing may be conducted on the same data or by choice of the programmer. Interest data are collected and processed according to the majority of the label values shown by k, where k is an integer. The value of k depends on the technique of determining the distance. The selection of k relies on the data. Larger k value lowers the rating of noise. Similarly, the selection of parameters is also an important means of improving classification precision. Weighted Knn Classification: a technique for giving the neighbor's value to a sufficient weight in order to have a substantial impact on the neighbour compared to the distance. The weight of the weighted knn method is important for the assessment of the closest optimistic value. Weight is typically based on an approximation to reciprocal distance. The weight of the property is multiplied by the time the necessary value is obtained.

V. Model proposed

Data analysis has been carried out using both version 3.6 of the Weka tool and the tool platform Jupiter Python [13, 24]. Weka is a tool for classifying, clustering, regressing and open-source data. Weka usually accepts the.csv or.arff input file. Weka Explorer provides a broad range of data analysis tabs, including pre-possession, categorisation, cluster, mix, selection and display properties. The Weka tool allows you to enter input data [3] when pre-possession data is chosen. Weka tool interprets and represents easy-to-analyse data. Weka Tool calls for several choices before the classification method is performed, including the percentage of division, training set, test set, cross validation option, etc. Classification is usually carried out via 80% training and 20% testing [6]. However, our research at Weka Tool was carried out utilising the classification method chosen for an appealing result with 10 times cross validation [8]. Weka is an easy-to-use display tool with several categorization methods and test results.



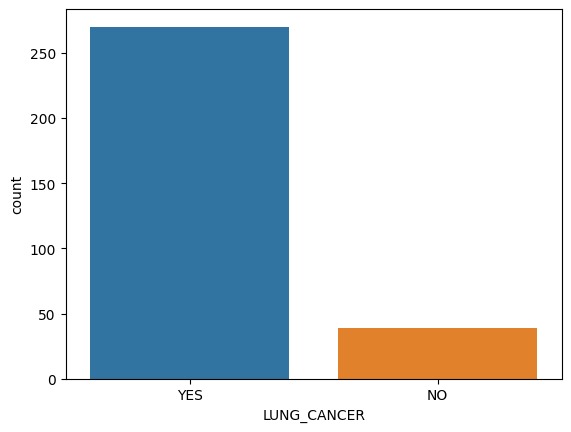
VI. Experimental analysis

The data provided is missing. Therefore, the data must be ready to replace the missing values with the most common column value. The processed data is then utilised for analysis in the Weka data mining tool. The data obtained is categorised using various categorization techniques appropriately. The approach classifier is subject to 10 cross validation techniques. The cross-validation process is a powerful data analysis technique, where 10 folds can be done using the provided data and the data can be predicted correctly. The Weka tool classification tab verifies several categorization methods. The outcomes of the suggested classifiers are compared after thorough examination. In 25 correctly classified examples, J48 and Naive Bayes algorithms categorise 32 situations and in 7 erroneously categorised instances. As with 24 properly classified and 8 incorrectly classified instances, there are 32 cases with 5 nearest neighbour knn. Our study has shown that the RBF grade is primarily selected from many grades. This is owing to its maximum accuracy in 26 cases properly categorised and 6

occurrences from 32 instances incorrectly categorised. Likewise, the value of both False Positive and False Negative is 3. The results of the several classifiers used in the Weka tool for lung cancer data are given in the table below. Usually in the abuse matrix Precision, recall, precision and F-measure are important characteristics for classification processes [4, 14]. Classification precision is the assessment of the predicted total predictions accurately. Some findings are based on these factors. These are 'TP' (true positive), the correctly anticipated event values, and 'TN' (true negatives).

VII. Conclusion

In this article we showed that the RBF classification is 81.25 percent accurate using data on lung cancer. The research may thus demonstrate that the accuracy of the functional selection method and the integrated approach with other supervised learning processes and the modified functional approach in the RBF increase further.



References

1. A. UCI, "Lung cancer dataset," [Online]. Available: https://archive.ics.uci.edu/ml/dataset/Lung+cancer. Accessed: Feb. 12, 2020.
2. World Health Organization, "WHO Deaths by cause, sex and mortality stratum," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: Jan. 25, 2020.
3. R.K. Ada, "Early detection and prediction of lung cancer survival using neural network classifier," 2013.
4. J.C.R. Alcantud et al., "Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making," PLoS ONE, vol. 14, no. 6, e0218283, 2019.
5. A. Asuntha and A. Srinivasan, "Deep learning for lung cancer detection and classification," Multimedia Tools Appl., vol. 79, pp. 1–32, 2020.
6. S. Bhatia, Y. Sinha, and L. Goel, "Lung cancer detection: a deep learning approach," in Soft Computing for Problem Solving, J.C. Bansal et al., Eds. Singapore: Springer, 2019, pp. 699–705.
7. P. Bhuvaneswari and A.B. Therese, "Detection of cancer in lung with k- nn classification using genetic algorithm," Procedia Mater. Sci., vol. 10, pp. 433–440, 2015.