Nat. Volatiles & Essent. Oils, 2021; 8(6): 1265-1272

NVEO
Natural Volatiles & Essential Oils

Lung Cancer Prediction Using Machine Learning Methodologies

Vemula Suvarchala [1], P Venkata Subbareddy [2], Srinivasa Rao Madala[3]

[1] M.Tech., Scholar, Department of Computer Science and Engineering,

[2] Associate Professor, Department of Computer Science and Engineering,

[3] Professor & HOD, Department of Computer Science and Engineering

PACE Institute of Technology and Sciences

Abstract:

Lung cancer usually occurs in both men and women as a result of unmanageable lung cell growth. This constitutes a severe respiratory problem in both the inhalation and the exhalation of the chest. Cigarette smoking and tobacco smoke are the main contributors to lung cancer in the global health organisation. The death rate due to lung cancer in young and old people is rising day by day in comparison with other cancers. Although high-tech medical facilities for been well and efficient medical treatment are available, the mortality rate is not yet properly monitored. Therefore way earlier safety measures are extremely necessary in the initial stage so that symptoms and effects can be identified early in the process for proper diagnosis. Nowadays, machine learning has a big impact on the health sector as a result of its high computational capacity for early disease diagnosis with reliable data analysis. In our paper, we analysed various techniques for machine learning to classify available lung cancer information into a malignancy UCI process was developed. The input data are pre-empted and transformed into binary form followed by a well-known Weka classifiers technique to classify data from cancer to non-cancerous. The method of comparison demonstrate that the suggested RBF classifier had a high accuracy of 81.25 percent and was regarded as the efficient classifiers method for prognostication of Lung cancer.

1. **Introduction**

Lung cancer is considered the most deadly ailment and the biggest burden of worry in the world today. Lung cancer impacts people more widely, presently ranks 7th in the death rate index with a projected 1.5 percent of the world's highest overall mortality rate[2-7]. Lung cancer is found in the lung and is still spreading in the brain. There are two main types of lung cancer. One is lung cancer of the non-small cell whereas the other is tiny lung cancer of the cell. Some of the health problems for patients include chest pain, dry cough, respiratory disease, weight loss, etc. Looking at the cultivation and causes of cancer, physicians concentrate more on smoking and second-hand cigarettes than on the main causes of lung cancer. Lung cancer treatment involves operations, chemotherapy, radiation therapy, immune therapy, etc. Despite this technique of diagnosing lung cancer, the clinician may only know this at an advanced

stage[18-22]. Early prognoses are thus very important before the last stage, in order to prevent the mortality rate simply. Statistics on pulmonary cancer survival are very encouraging even after correct treatment and diagnosis. Lung cancer survival rates vary from person to person. It depends on age, gender, race and health. In early stages of a safe human existence, machine learning plays an important role in detecting and predicting medical issues. Machine learning simplifies and facilitates the diagnostic process. Machine learning has dominated the medical profession for a day now. Each country in its health industry now utilises machine learning methods. Machine learning may be used to explore actual disease detection. Some of the main uses of machine learning are the extraction of traits: there is a real information container for disease in every attribute of disease. Machine learning enhances data analysis and analyses the actual features or information, identifying the real problem that causes a disease. It allows doctors to detect the underlying cause of disease. Processing of images: Exact and usable picture analysis has been found via several machine learning techniques. This allows doctors to better detect diseases to save money and time and enhance their value. Drug production: the medicine should be multifunctional in proportion to the increase in different diseases and the known quantity. ML has dealt with this problem and allows the pharmaceutical industry to benefit from the ML application. Better disease prediction: ML helps forecast the severity and outcome of disease[27-29]. ML regulates the prediction of early disease outbreaks in order to take appropriate actions. To be more standardised and reliable, another machine learning application needs to be developed. This would allow doctors, the catalyst for health, to make clinical decisions properly and with great accuracy. The system uses its own education methods to address the issue. All kinds of ML are unattended learning, supervised learning and enhanced learning. Supervised learning recognises two processes, one and the other. Classification is the process of processing and aggregating data. The task was carried out in the instrument Weka

2. Related Work

Hosseinzade et al (2013). SVM suggested that protein characteristics should be selected and found that the outcome is 88% accurate compared to the other method of lung cancer prediction[13-16].

The Northern Centralized Cancer Group (NCCTG) C4.5, Naveen and Pradeep (2018), suggested better results in better data between the SVM, Naive Bayes, and C4.5 classifications on lung cancer and predicted that C4.5 would be better classified as a result of the increase in lung cancer training data[25]. Gur Amrit Pal singh & P.K Gupta (2018) presented the novel extraction method for visual data and used machine learning classifications to enhance precision[27].

Hussein et al. (2019) suggested to categorise 91 percent of correct benign and malignant data as the 3D CNN, supervised learning of the lung nodule and unattended SVM approaches[12]. Monkam et al. (2019) presented a review of the significance of the neural network of Convolution with a precision of nearly 90% for the prediction of a pulmonary module[21]. Asuntha and Andy Srinivasan (2019) suggested optimisation of fluorescent particle swarm by a deep neural network in pictures of lung cancer to reach 99.2% accuracy.

Ganggayah et al. (2019) tested different classifications using 8066 records of 23 predictors in their data on breast cancer and found that the random forest classification was 82percent more accurate[9]. The

supervised learning of Gibbons et al. (2019) employed a linear regression model, vector support machine, ANN, etc., forecasting SVM is 96 percent more accurate than other methods[18]. Shakeel et al. (2019) have utilised ANN's novel hybrid selection method to predict 99.6 per cent of accuracy in lung cancer data from the ELVIRA biomedical data[26]. Bhuvaneswari et al. (2015) utilised gabor filters for the purpose of characteristic extraction and G-knn method to classify images of lung cancer with 90% accuracy[7].

3. Data Set

In the UCI Study Machine Repository, Dataset was available. The data consists of 32 instances with 57 characteristics, all 0-3 predictive and 3 class attributes (1 class attribute and 56 inputs). Nominal features and class label data are converted into binary for easier data processing. The most standardised way of data processing is nominal binary form. The data collection contains certain missing values which degrade the performance of the algorithm to guarantee full execution before data analysis is performed. The etiquette is high, low, middle and medium. High to 2, medium to 1 and low to 0.4 in the article.

4. Technical Classification

Segmentation is carried out via a supervised learning experience to anticipate particular input data on a class label. The uniqueness of the classification depends on a certain degree of mapping. Various classifications are called Perceptron, Naïve Bayes, Decision Tree, Logistics Regression, K next door, Artificial Network, and Vector Machine Support. The classification of machine learning is one of the first techniques of data analysis making decisions. Different techniques are also used to categorise samples of data[20-23]. The concept of our paper centres on a new way of analysing the data on lung cancer to guarantee high precision. Some of the classification techniques most commonly employed are described.

Network of Neurons

The neural network is the fundamental element in machine learning that teaches neurons. ANN includes the input layer, the intermediate layer hidden neuron, and the output layer. Each neuron input is connected to the hidden neuron with an adequate weight and also to the output unit through the hidden unit. The neuron is processed using a specified functional threshold value in the cloaked neuron and neuron output. Activation is used to treat the neuron as necessary. The synaptic weight is multiplied by the corresponding neuron shown for classification in the oversized layer and output layer. The intended objective is adjusted to achieve the desired result via a technique of weight modification. Network feedback techniques provide a simpler classification process.

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).

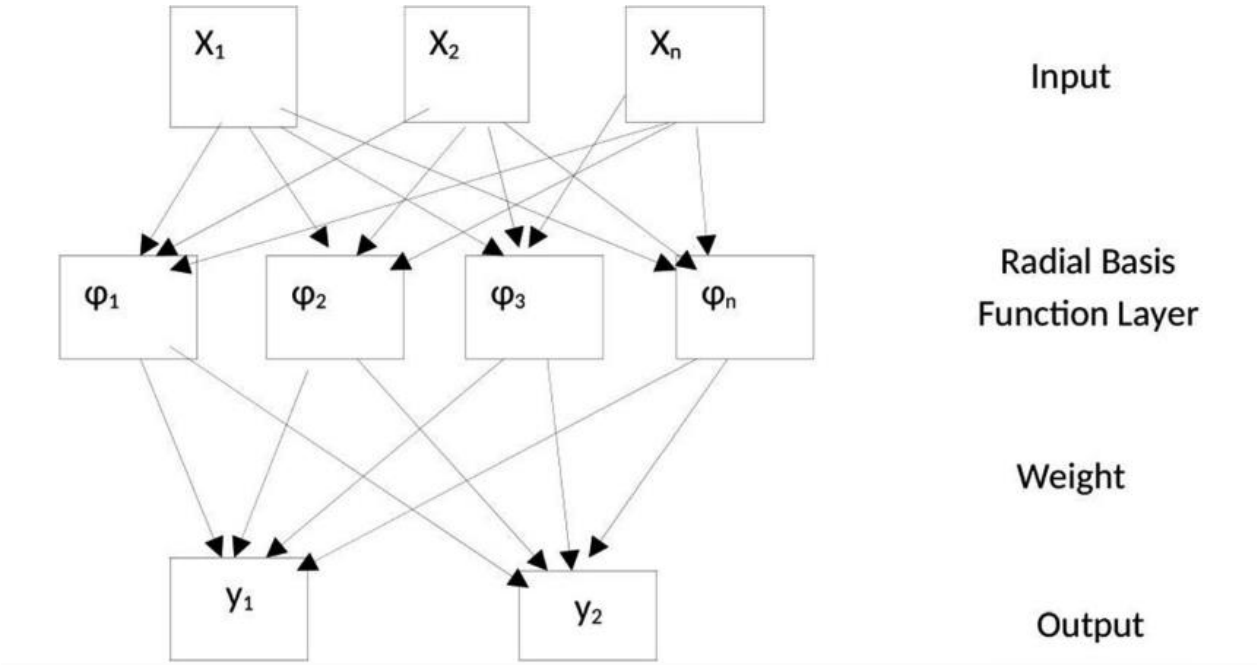

Fig. 1. RBF.

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 midlayer and the output layer.

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.

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.

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.

5. 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-analyze 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.

6. 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).

7. 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. https://archive.ics.uci.edu/ml/dataset/Lung+cancer. Accessed 12 Feb 2020
2. WHO Deaths by cause, sex and mortality stratum, World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 25 Jan 2020
3. Ada, R.K.: Early detection and prediction of lung cancer survival using neural network classifier (2013)Google Scholar
4. Alcantud, J.C.R., Varela, G., Santos-Buitrago, B., Santos-Garcia, G., Jimenez, M.F.: Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making. PLoS ONE **14**(6), e0218283 (2019)Cross Ref Google Scholar
5. Asuntha, A., Srinivasan, A.: Deep learning for lung cancer detection and classification. Multimedia Tools Appl. **79**, 1–32 (2020)CrossRef Google Scholar
6. Bhatia, S., Sinha, Y., Goel, L.: Lung cancer detection: a deep learning approach. In: Bansal, J.C., Das, K.N., Nagar, A., Deep, K., Ojha, A.K. (eds.) Soft Computing for Problem Solving. AISC, vol. 817, pp. 699–705. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1595-4_55CrossRef Google Scholar
7. Bhuvaneswari, P., Therese, A.B.: Detection of cancer in lung with k- nn classification using genetic algorithm. Procedia Mater. Sci. **10**, 433–440 (2015)CrossRef Google Scholar
8. Chaubey, N.K., Jayanthi, P.: Disease diagnosis and treatment using deep learning algorithms for the healthcare system. In: Applications of Deep Learning and Big IoT on Personalized Healthcare Services, pp. 99–114. IGI Global (2020)Google Scholar
9. Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., Dhillon, S.K.: Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med. Inform. Decision Making **19**(1), 48 (2019)CrossRefGoogle Scholar
10. Hachesu, P.R., Moftian, N., Dehghani, M., Soltani, T.S.: Analyzing a lung cancer patient dataset with the focus on predicting survival rate one year after thoracic surgery. Asian Pacific J. Cancer Prevention: APJCP **18**(6), 1531 (2017)Google Scholar
11. Hosseinzadeh, F., KayvanJoo, A.H., Ebrahimi, M., Goliaei, B.: Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus **2**(1), 238 (2013)CrossRefGoogle Scholar

12. Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U.: Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. IEEE Trans. Med. Imag. **38**(8), 1777–1787 (2019)CrossRefGoogle Scholar
13. Jacob, D.S., Viswan, R., Manju, V., PadmaSuresh, L., Raj, S.: A survey on breast cancer prediction using data mining techniques. In: 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 256–258. IEEE (2018)Google Scholar
14. Jakimovski, G., Davcev, D.: Using double convolution neural network for lung cancer stage detection. Appl. Sci. **9**(3), 427 (2019)CrossRefGoogle Scholar
15. Kadir, T., Gleeson, F.: Lung cancer prediction using machine learning and advanced imaging techniques. Transl. Lung Cancer Res. **7**(3), 304 (2018)CrossRefGoogle Scholar
16. Kohad, R., Ahire, V.: Application of machine learning techniques for the diagnosis of lung cancer with ant colony optimization. Int. J. Comput. Appl. **113**(18), 34–41 (2015)Google Scholar
17. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struc. Biotechnol. J. **13**, 8–17 (2015)Google Scholar
18. Krishnaiah, V., Narsimha, G., Chandra, D.N.S.: Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol. **4**(1), 39–45 (2013)Google Scholar
19. Li, X., Hu, B., Li, H., You, B.: Application of artificial intelligence in the diagnosis of multiple primary lung cancer. Thoracic Cancer **10**(11), 2168–2174 (2019)CrossRefGoogle Scholar
20. Lynch, C.M., et al.: Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int. J. Med. Inform. **108**, 1–8 (2017)CrossRefGoogle Scholar
21. Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., Qian, W.: Detection and classification of pulmonary nodules using convolutional neural networks: a survey. IEEE Access **7**, 78075–78091 (2019)CrossRefGoogle Scholar
22. Murty, N.R., Babu, M.P.: A critical study of classification algorithms for lungcancer disease detection and diagnosis. Int. J. Comput. Intell. Res. **13**(5), 1041–1048 (2017)Google Scholar
23. Paing, M.P., Hamamoto, K., Tungjitkusolmun, S., Pintavirooj, C.: Automatic detection and staging of lung tumors using locational features and double-staged classifications. Appl. Sci. **9**(11), 2329 (2019)CrossRefGoogle Scholar
24. Patel, D., Shah, Y., Thakkar, N., Shah, K., Shah, M.: Implementation of artificial intelligence techniques for cancer detection. Augmented Human Res. **5**(1), 6 (2020)CrossRefGoogle Scholar
25. Pradeep, K., Naveen, N.: Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4. 5 and naive bayes algorithms for healthcare analytics. Procedia computer science **132**, 412–420 (2018)Google Scholar
26. Shakeel, P.M., Tolba, A., Al-Makhadmeh, Z., Jaber, M.M.: Automatic detection of lung cancer from biomedical data set using discrete adaboost optimized ensemble learning generalized neural networks. Neural Comput. Appl. **32**(3), 777–790 (2020)CrossRefGoogle Scholar
27. Madala, S. R., Rajavarman, V. N., & Vivek, T. V. S. (2018). Analysis of Different Pattern Evaluation Procedures for Big Data Visualization in Data Analysis. In Data Engineering and Intelligent Computing (pp. 453-461). Springer, Singapore.
28. Madala, S. R., & Rajavarman, V. N. (2018). Efficient Outline Computation for Multi View Data Visualization on Big Data. International Journal of Pure and Applied Mathematics, 119(7), 745-755.

29. Vivek, T. V. S., Rajavarman, V. N., & Madala, S. R. (2020). Advanced graphical-based security approach to handle hard AI problems based on visual security. International Journal of Intelligent Enterprise, 7(1-3), 250-266.