An Optimized Lung Cancer Prediction System using Machine Learning

Dr. Ashish Tiwari
Computer Science & Engineering
Amity University Lucknow, India
Atiwari3@lko.amity.edu

Vivek Kumar Pandey
Computer Science & Engineering
UCER, Prayagraj, India
vivek5confidential@gmail.com

Saurabh Mishra
Information Technology
UCER, Prayagraj, India
saurabhmisra@gmail.com

ABSTRACT

The lungs are the centre of breath control and ensure that every cell in the body receives oxygen. At the same time, they filter the air to prevent the entry of useless substances and germs into the body. The human body has specially designed defence mechanisms that protect the lungs. However, they are not enough to completely eliminate the risk of various diseases that affect the lungs. Infections, inflammation or even more serious complications, such as the growth of a cancerous tumour, can affect the lungs. Lung cancer generally occurs in both male and female due to uncontrollable growth of cells in the lungs. This causes a serious breathing problem in both inhale and exhale part of chest. Cigarette smoking and passive smoking are the principal contributor for the cause of lung cancer as per world health organization. The mortality rate due to lung cancer is increasing day by day in youths as well as in old persons as compared to other cancers. Even though the availability of high-tech medical facility for careful diagnosis and effective medical treatment, the mortality rate is not yet controlled up to a good extent. Therefore, it is highly necessary to take early precautions at the initial stage such that it's symptoms and effect can be found at early stage for better diagnosis. Machine learning now days has a great influence to health care sector because of its high computational capability for early prediction of the diseases with accurate data analysis. The lungs are the centre of breath control and ensure that every cell in the body receives oxygen. At the same time, they filter the air to prevent the entry of useless substances and germs into the body. The human body has specially designed defence mechanisms that protect the lungs. However, they are not enough to completely eliminate the risk of various diseases that affect the lungs. Infections, inflammation or even more serious complications, such as the growth of a cancerous tumour, can affect the lungs. In this work, we used machine learning (ML) methods to build efficient models for identifying high-risk individuals for incurring lung cancer and, thus, making earlier interventions to avoid long-term complications. The suggestion of this article is the Rotation Forest that achieves high performance and is evaluated by well-known metrics, such as precision, recall, F-Measure, accuracy and area under the curve (AUC). More specifically, the evaluation of the experiments showed that the proposed model prevailed with an AUC of 99.3%, F-Measure, precision, recall and accuracy of 97.1%.

Keywords— Machine Learning, Probability Prediction, Efficient Model, Risk Analysis.

I. INTRODUCTION

When bodily cells proliferate unchecked, a condition named as cancer outcomes. When cancer develops in the lungs, it is referred to as lung cancer. Other bodily parts, such as lymph nodes, organs including the brain, the lungs can also be the site of the start of lung cancer. Lung cancer has the potential to spreading out to further organs. The term "cancer cells" refers to cells which have spread from one organ to another. The two main groups into which they are commonly separated are tiny cell and non-tiny cell lung malignancies, which include adenocarcinoma and squamous cell carcinoma. These numerous types of lung cancer have distinctive patterns of development and therapeutic responses [1]. While small cell lung cancer is more common, non-small cell lung cancer is more common. Lung cancer, which is also the worst disease, is thought to be the main factor in high mortality in the modern world. Compared to other cancers, lung cancer has a greater impact on people, and as expected, it currently occupies position seven in the fatality rate index, contributing 1.6% of world death [2]. The brain is affected by lung cancer after it has spread to the lung. There are two primary classifications of lung cancer. The two forms of lung cancer are tiny cell and non-tiny cell. Acute chest hurt, a dry wheeze, shortness of inhalation, body weight loss, and other symptoms are possible in patients [3]. Doctors who study the causes and progression of cancer emphasise the role of smoking and passive lung cancer is primarily caused by smoking. Lung cancer is treated with abscission, chemo, diffraction, immune remedy, and other procedures. Despite this, doctors can only diagnose lung cancer once it has advanced, making the diagnosis relatively weak [4]. To quickly and effectively lower the mortality rate with effective control, early prediction prior to the last phase is essential. Even with the right treatment and diagnosis, the prediction for lung cancer is quite encouraging[5]. The prognosis for lung cancer varies depending on the patient's age and gender, and race are all factors, as well as health status. The American Cancer Society calculates that a patient's likelihood of surviving lung cancer is 47% if it is identified at a young stage. It is extremely improbable that lung cancer in its early stages will be accidently discovered on an X-ray image [6].

Figure 1: CT Scan image for lung Cancer.

Figure 1 shows, During the study piece, several machine learning (ML) models were employed for the topic at hand in order to compare how well they performed against one another. More particular, we examined the Support Vector Machine (SVM), a widely used kernel-based classifier [6]. Additionally, a linear classifier was trained using stochastic gradient descent (SGD) [5] under an SVM convex loss function. [7] were taken advantage of from the ensemble random forest (RF). Finally, a straightforward artificial neural network and a distance-based classifier called K-nearest neighbours (K-NN) [7] were assessed.

II. Risk Factors of Cancer

Multiple Risks elements have been recognized through means of studies that could increase your chance of spreading lung cancer. Lung cancer chance is primarily increased by smoking. For 80% to 90% of lung cancer fatalities in the US, smoking cigarettes is to blame. Smoking tobacco, including cigarettes, cigars, and pipes, raises the chance of lung cancer developing. There are about 7,000 compounds in tobacco smoke, Consequently, it is very poisonous. Lots of them are lethal. One way or another, minimum 70 of them have been joined to either human or animal cancer [9]. Smokers have a 15–30-fold higher danger of non-smokers to acquire lung cancer or die from it. Even light or infrequent cigarette usage raises the chance of lung cancer. Smoking more frequently and for longer periods of time raises the chance. Smokers who left smoking have a lower chance of lung cancer compare to they would have otherwise, but they still have a higher risk than non-smokers [10]. Smoking cessation can lower the danger of lung cancer at any age. In practically each and every bodily part, smoulder increases the chance of cancer. Smoking shoots up the risk of grow a number of cancers, including those of the voice box (larynx), trachea, stomach, colon, rectal, liver, pancreas, mouth, throat, oesophageal, stomach, colon, and bronchial. Lung cancer risk is also increased by second hand smoke, which includes tobacco, cigar, and pipe smoke. Anyone who inhales second hand smoke is doing the same thing as someone who smokes [11]. One in four non-smokers and 14 million children in the United States during 2013 and 2014 were exposed to second hand smoke. In the US, smoking and radon are the two leading causes of lung cancer. Water, soil, and rocks can all be the source of the radon-filled natural gas. It has no flavour or smell and is translucent. Radon may become trapped and start to build up in the air when it enters homes or other buildings through cracks or holes [12]. Those People occupy or are employed by these residences, businesses are exposed to high amounts of radon. Lung cancer can develop after a long duration due to radon exposure. The Environmental Protection Agency (EPA) in the United States estimates that Radon is a factor in the annual death toll from lung cancer of 21,000 persons. Lung cancer is more likely to develop if you are exposed to radon in smokers compared to non-smokers [13]. However, the EPA claims probably greater than 10% of deaths from lung cancer associated with radon occur in smokers who have never smoked cigarettes. Nearly one in every fifteen homes in the US have excessive radon levels. Find out how to radon test your home and how help reduce radon levels if they are excessive. Cancer in the lung is most prevalent category of cancer, consider for one in six fatalities annually and accounting for 1.76 million deaths as of 2016. A patient's life can be saved or extended with the right early cancer diagnosis and therapy, which raises the survival rate.

Table 1. The order of features in the balanced data

Random Forest		Gain Ratio	
Age	0.3463	Senstivity	0.3952
Senstivity	0.2808	Liquor	0.3698
Liquor	0.2664	SwallowDifficulty	0.3255
Inhaling	0.2568	Inhaling	0.3082
Whoop	0.2443	PeerPressure	0.293
SwallowDifficulty	0.2328	Coughing	0.2475
PeerPressure	0.2244	Age	0.1565
ChronicDisease	0.1663	ChronicDisease	0.1176
ChestPain	0.0959	ChestPain	0.0435
Unease	0.0775	YellowFingers	0.0292
Smoulder	0.0752	Unease	0.028
YellowFingers	0.0726	Smoulder	0.023
ShortnessofBreath	0.0433	ShortnessofBreath	0.0135
Sex	−0.0055	Sex	0.0026
Exhaustion	−0.0333	Exhaustion	0.0009

A neural network, a form of artificial intelligence, is used to train the input data specimen and then test them. At the start of the procedure, the weights of the neural network are generated randomly from the input data. The same dataset that was utilised for training the neural networks serves as the basis for their evaluation. To determine the frequency of errors or error rates that occur during classification process, data is weighted. Errors are then corrected by reweighting the dataset. We then calculated the importance score of each feature that was involved in the features analysis for the target class. Two feature ranking techniques—gain ratio and random forest—were taken into consideration for this purpose. In order to evaluate a feature's capacity to best distinguish between instances in the two classes, Random Forest computes the Gini impurity [9]. Table 1 displays the ranking scores in downward-sloping. We can observe that five out of fourteen features were placed in the same sequence as significance by both approaches based on the calculated scores, while some of the other features were arranged in proximal or reverse order. Values that are close to 0 and/or negative indicate characteristics that are of low or no importance. All of the qualities will be taken into account while training and validating the models because they are necessory predictors of lung cancer development and medical professionals' guidance of it.

III. Machine Learning Principles Design

The number of occurrences from all of the data that were correctly predicted is measured and used to evaluate the presentation of the classification job. We also looked at recall, which measures a model's sensitivity to distinguish between patients who genuinely had lung cancer and were rightly classified as productive in comparison to all deserving contributors. Table 1 displays the apparent of the traits in every class. Men and women are almost uniformly likely to be given a lung cancer diagnosis based on their gender. Additionally, based on this table, we can consequently, each of the characteristics we examined is turned on in lung cancer patients by 27% to 36%, despite the fact that a significant number of patients reported these symptoms even before receiving a lung cancer diagnosis. Even though the illness hadn't formed, keeping an eye on risk factors, warning signs, and subsequent clinical checks may assist to shut out or decrease the disease's unfavourable outcome.

IV. Results and Discussion

The Weka [6] environment was used to evaluating the ML models' performance since it given a number of athenaeum for data preparation, classification, clustering, forecast, and visualisation. A computer system with

the following characteristics was used to conduct the experiments as well: an x64 CPU, Windows 11 Home, a 64-bit operating system, and an 11th generation Intel(R) Core(TM) i7-1165G7 processor operating at 2.81 GHz with 15.9 GB of RAM. We utilised SMOTE and 10-fold cross authenticate to evaluate the models' performance on the balanced dataset of 541 cases. The best parameter choices for the suggested ML models are finally listed in Table 2.

Table 2. The order of features in the balanced data.

Models	Varibles
SVM	eps=0.002 gamma=0.0 kerneltype:linear loss=0.2
KNN	K=3.1 SearchAlgorithm:LinearNNSearch withEuclidean
RF	maxDepth=0 numIterations=100 numFeatures=0

Numerous machine learning models, including SVM, KNN, and RF, are assessed in the framework of this study work to be able to identify utilising the model, greatest predictive result in relation to accuracy, precision, recall, F-Measure, and AUC. Our presentation evaluation of the models following SMOTE with 11-fold cross-evidence is provided in Table 2. Percentages greater than 93.4% (RT) are shown by all of our suggested models. With an AUC of 99.4%, it has 97.2% accuracy, precision, recall, and F-Measure. The fact that RF, with 99.2%, and AdaBoostM1, with 98.6%, which uses RF as its basis classifier, both obtain high percentages of AUC should also be noticed. The proposed machine learning models' AUC ROC curve is finally plotted in Figure 2 for reference.

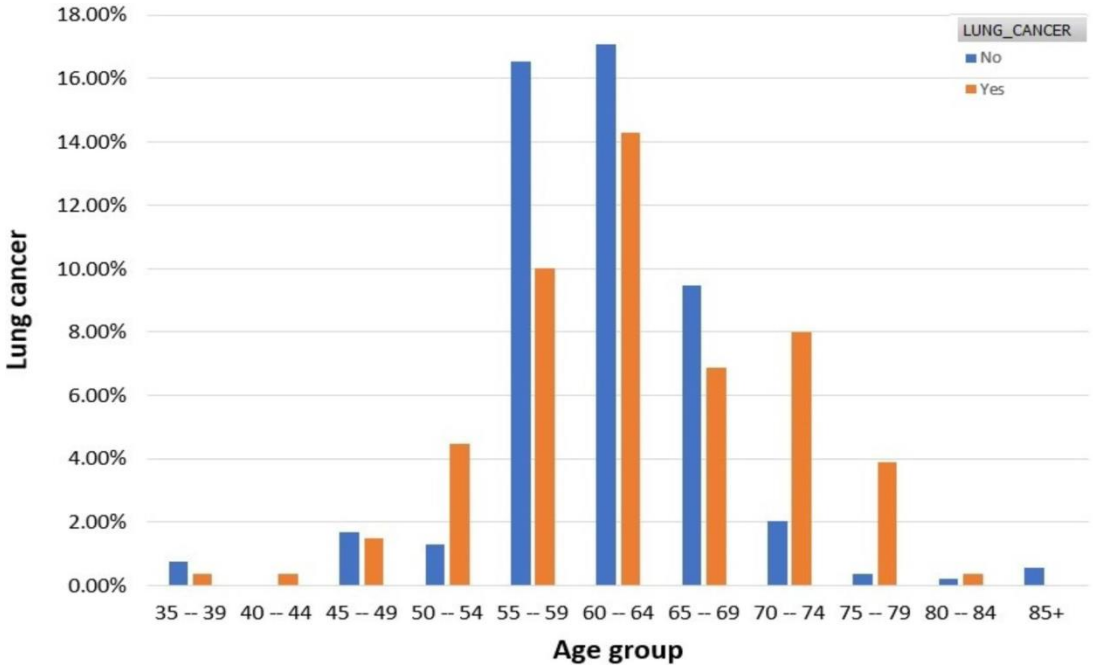

Figure 2. Distribution of participants among the age groups in the balanced data.

CONCLUSION AND FUTURE USE

The primary respiratory organs are the lungs. Due to the lungs' ability to feed their blood with oxygen, that is necessary for human existence, humans never cease breathing until they pass away the most typical cancer-

causing factor-related death in people of both genders is lung cancer. The advanced phase of the cancer determines the patient's life expectancy. The life expectancy increases with the timing of the diagnosis. In this work, we make use of supervised learning to create models for determining whether a person has lung cancer manifestation based on a variety of features-symptoms. To evaluate the F-Measure, AUC, F-Measure, and recall of various machine learning models, such as SVM, KNN, and RF. Based on the results of the experiment and utilising SMOTE with 10-fold cross-validation, the RF outperformed the other models, achieving an accuracy, precision, recall, F-Measure, and AUC of 97.2% and 99.4%, respectively.

REFERENCES

[1] Schiller, H.B.; Montoro, D.T.; Simon, L.M.; Rawlins, E.L.; Meyer, K.B.; Strunz, M.; Vieira Braga, F.A.; Timens, W.; Koppelman,G.H.; Budinger, G.S.; et al. The human lung cell atlas: A high-resolution reference map of the human lung in health and disease.Am. J. Respir. Cell Mol. Biol. 2019, 61, 31–41.

[2] Hervier, B.; Russick, J.; Cremer, I.; Vieillard, V. NK cells in the human lungs. Front. Immunol. 2019, 10, 1263.

[3] Barroso, A.T.; Martín, E.M.; Romero, L.M.R.; Ruiz, F.O. Factors affecting lung function: A review of the literature. Arch. De Bronconeumol. 2018, 54, 327–332.

[4] Mirza, S.; Clay, R.D.; Koslow, M.A.; Scanlon, P.D. COPD guidelines: A review of the 2018 GOLD report. In Mayo Clinic Proceedings; Elsevier: Amsterdam, The Netherlands, 2018; Volume 93, pp. 1488–1502.

[5] Dotan, Y.; So, J.Y.; Kim, V. Chronic bronchitis: Where are we now? Chronic Obstr. Pulm. Dis. J. COPD Found. 2019, 6, 178.

[6] Stern, J.; Pier, J.; Litonjua, A.A. Asthma epidemiology and risk factors. In Seminars in Immunopathology; Springer: Berlin/Heidelberg,Germany, 2020; Volume 42, pp. 5–15.

[7] Bell, S.C.; Mall, M.A.; Gutierrez, H.; Macek, M.; Madge, S.; Davies, J.C.; Burgel, P.R.; Tullis, E.; Castaños, C.; Castellani, C.; et al.The future of cystic fibrosis care: A global perspective. Lancet Respir. Med. 2020, 8, 65–124.

[8] Tiwari, A., & Garg, R. (2022). Adaptive Ontology-Based IoT Resource Provisioning in Computing Systems. International Journal on Semantic Web and Information Systems (IJSWIS), 18(1), 1-18.

[9] Buyya, R., Yeo, C. S., & Venugopal, S. (2008, September). Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In 2008 10th IEEE international conference on high performance computing and communications (pp. 5-13). Ieee.

[10] Tiwari, A., & Garg, R. (2022). A Optimized Taxonomy on Spot Sale Services Using Mathematical Methodology. International Journal of Security and Privacy in Pervasive Computing (IJSPPC), 14(1), 1-21.

[11] Bowen, J. A. (2011). Legal issues in cloud computing. Cloud Computing: Principles and Paradigms, 593-613.

[12] Tiwari, A., & Garg, R. (2022). Reservation System for Cloud Computing Resources (RSCC): Immediate Reservation of the Computing Mechanism. International Journal of Cloud Applications and Computing (IJCAC), 12(1), 1-22.

[13] Kumar Sharma, A., Tiwari, A., Bohra, B., & Khan, S. (2018). A Vision towards Optimization of Ontological Datacenters Computing World. International Journal of Information Systems & Management Science, 1(2).

[14] Tiwari, A., & Sharma, R. M. (2018). Rendering Form Ontology Methodology for IoT Services in Cloud Computing. International Journal of Advanced Studies of Scientific Research, 3(11).

[15] Rangaiah, Y. V., Sharma, A. K., Bhargavi, T., Chopra, M., Mahapatra, C., & Tiwari, A. (2022, December). A Taxonomy towards Blockchain based Multimedia content Security. In 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT) (pp. 1-4). IEEE.

[16] Rohinidevi, V. V., Srivastava, P. K., Dubey, N., Tiwari, S., & Tiwari, A. (2022, December). A Taxonomy towards fog computing Resource Allocation. In 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT) (pp. 1-5). IEEE.

[17] Singh, N. K., Jain, A., Arya, S., Gonzales, W. E. G., Flores, J. E. A., & Tiwari, A. (2022, December). Attack Detection Taxonomy System in cloud services. In 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT) (pp. 1-5). IEEE.

[18] Chouhan, A., Tiwari, A., Diwaker, C., & Sharma, A. (2022, February). Efficient Opportunities and Boundaries towards Internet of Things (IoT) Cost Adaptive Model. In 2022 IEEE Delhi Section Conference (DELCON) (pp. 1-5). IEEE.

[19] Singh, Shubhuam, Pawan Singh, and Sudeep Tanwar. "Energy aware resource allocation via MS-SLnO in cloud data center." Multimedia Tools and Applications (2023): 1-23.

[20] Singh, Pawan. "Energy Management in Cloud Through Green Cloud Technologies." Journal of Management and Service Science (JMSS) 2, no. 2 (2022): 1-11.

[21] Rawat, A., & Singh, P. (2021). A Comprehensive Analysis of Cloud Computing Services. Journal of Informatics Electrical and Electronics Engineering (JIEEE), 2(3), 1-9.

[22] Khan, H., & Singh, P. (2021). Issues and Challenges of Internet of Things: A Survey. Journal of Informatics Electrical and Electronics Engineering (JIEEE), 2(3), 1-8.

[23] Singh, P., Hailu, N., & Chandran, V. (2014). Databases for Cloud Computing: Comparative Study and Review. European Journal of Academic Essays, 1(6), 12-17.