JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A New tool to Predict Lung Cancer Based on Risk Factors

Mrs Anushiya .C –

Assistant Professor – PG & Research Department of Computer Science

Dr H.Sivalingan

Head & Assistant Professor Department of Data Science

Providence College Coonoor

Abstract

Lung cancer is one of the deadliest cancers in the world. Hundreds of researches are presented annually in the field of lung cancer treatment, diagnosis and early prediction. The current research focuses on the early prediction of lung cancer via analysis of the most dangerous risk factors. A novel tool for the early prediction of lung cancer is designed following three stages: the analysis of an international cancer database, the classification study of the results of local medical questionnaires and the international medical opinion obtained from recently published medical reports. The tool is tested using local medical cases and the local medical opinions are used to determine the accuracy of the scores obtained. The Machine Learning approaches are also used to analyze 1000 patient records from an international dataset to compare our results. The designed tool facilitates computing the risk factors for people who are unable to perform costly hospital tests. It does not require entering all risk inputs and produces the risk factor of lung cancer as a percentage in less than a second. The comparative study with medical opinion and the performance evaluation have confirmed the accuracy of the results.

Keywords: Computer science, Cancer prevention, Prediction tool, Lung symptoms, Lung cancer, Risk factors

Introduction

Lung cancer is the most dangerous and deadliest type of cancer. Smoking is the basic risk factor for lung cancer, and it accounts for 85 out of 100 people dying every year. Although people who do not smoke have a lower risk factor, they may still be affected by the smoke of other smoker's .There are many other risk factors, such as second-hand smoking, exposure to radiation and air pollution. Uranium is a metallic chemical element, which breaks down, with time, to form radon gas, which spreads in the air and water causing pollution and great harm to the lungs .Lung cancer risk degree increases when there are cases of lung cancer in relatives, and this may be due to a common environment, genes or both . In addition, the history of chronic pulmonary

diseases is associated with lung cancer .Prognostic models to predict cancer have been developed in many cases, including the incorporation of these tools for patient selection and pretreatment stratification into clinical trials ; some of these tools predicted lung cancer .

General system description

The proposed system includes several stages to achieve the ultimate goal of building a software prediction tool. In the first stage, a global medical database is analyzed to determine the most common symptoms of lung cancer from a standard medical point of view. In the second stage, several medical questionnaires are distributed among a number of doctors and specialists in the fields of internal and thoracic tumors, in order to determine the most effective symptoms of lung cancer from a local medical point of view. In the third stage, medical knowledge from global research projects and reports is extracted, and the most appropriate pathway is defined in order to determine the risk of lung cancer by analyzing the values of available diagnostic factors.

Based on the knowledge derived from stages I and II and the knowledge generated from stage III, a software tool is developed to predict the risk of lung cancer based on the outcome of these three stages. The proposed tool estimates the degree of lung cancer by considering several factors that represent direct cancer risks and environmental factors.

Database analysis

The database consists of 1000 records and 23 attributes that represent the symptoms, risk factors of lung cancer and three categories representing the risk levels of lung cancer: Low, Medium and High. The database is analyzed to see the effect of each characteristic on determining the risk level. The "WEKA" tool is used for the database analysis step.

Analysis of lung cancer risk factors

The analysis of the risk factors is depicted. It shows charts visualizing these factors, which include:

Air pollution can be considered one of the most influential long-term factors of lung cancer. Therefore, high levels of pollution (6–8), on a scale of (1–8), are a major factor in causing the disease. Alcohol consumption, on a scale of (1–8), is one of the risk factors, but it does not affect the lungs directly. The risk of cancer increases in people who drink alcohol.

The inhalation of dust is normal and its effect will disappear when the causative agent disappears. However, high pollution rates of dusty environments increase the risk of cancer.

Genetic risk (i.e. family history), is an important factor. If a person's family has a history of lung cancer, their risk of the disease, on a scale of (1–7), will be in the range of (5–7).

Chronic lung infections are a weak indication of lung cancer, but their recurrence may be an important sign of future cancer. On a scale of (1–7), the risk of cancer starts to materialize for values in the range (4–7), and the risk increases significantly as the value approaches 7.

Analysis of lung cancer symptoms

Lung symptoms are analyzed and plotted as shown in Figure 2. The following points can be noted:

- The risk of cancer increases slightly with increased chest pain. Chest pain cannot be always linked to cancer even if it is hard as it may be due to inflammation or heart problems.

- Coughing blood is the most common sign of lung cancer, especially when combined with other risks, such as, fatigue and pain. An increased blood flow rate significantly increases the risk of lung cancer. There are cases where fatigue is an indicator of lung cancer, especially when frequent and concomitant with other symptoms. Higher degrees of fatigue with repetition may lead to higher risks of cancer.
- Regarding weight loss, there are cases in which the range is between 2 and 4, on a scale of (1–8), and the risk still materializes. Therefore, the coefficient of weight loss exists for most types of cancer; however, this factor alone is not enough.
- Shortness of breath is another lung cancer symptom, but there are many cases where lack of breathing is normal and is not linked to cancer. Parts of this symptom are on the scale of (6–9) and this indicates the likelihood of cancer.
- Increased difficulty of swallowing and the frequent occurrence of this condition (range (5–8)) increase the risk of lung cancer.
- Other factors, such as frequent cold, dry cough and snoring are symptoms of lung cancer. In general, high levels of frequent cold do not indicate a high risk and, similarly, low levels do not indicate a low risk. The risk is noticed significantly at high scopes of frequent cold factor.
- Dry coughs are associated with similar symptoms such as difficulty in swallowing, wheezing and shortness of breath. These are significant indicators of lung cancer. The diagrams of these symptoms are compared and found to be similar; they all increase the risk of cancer.

Results of database analysis

During the analysis of the considered database, the symptoms and factors are divided into three categories based on the degree of their effects on the probability of lung cancer. The factors and symptoms that have high-risk effects are smoking, air pollution, dust allergy, genetic risks and coughing blood. The medium-risk factors and symptoms are alcohol consumption, chronic inflammations, balanced diet, obesity, fatigue, weight loss, shortness of breath, frequent cold and dry cough. Finally, low-risk factors and symptoms are occupational hazards, chest pain, wheezing, swallowing difficulties and snoring.

International medical opinion and studies

Tobacco has more than 7000 chemicals which are known to cause cancer [4]. Smoking, of any type, increases the risk of lung cancer. The good news is that quitting smoking decreases risk of cancer . Smoking is considered the most dangerous risk factor [1, 2, 3, 4, 26, 27, 28, 29]. People who do not smoke can still get lung cancer if they are second-hand smokers [1, 2, 3, 4]. However, although smoking is a major risk factor for lung cancer, 40% of Asian lung cancer patients are non-smokers . Internationally, the third most common risk factor is exposure to radon gas [1, 4, 26, 27, 28]. Some medical studies have linked exposure to radon gas to lung cancer, while others have not . Some medical studies have listed contact with asbestos or other cancer-causing agents as a lung cancer risk factor [1, 2, 4, 26]. The personal history of lung diseases and the family history of lung cancer are considered the second most common risk factors for lung cancer [1, 2, 3, 4, 26, 27, 28].

Classification study of database used

In order to compare the proposed LPCT output with the results of the international database, the classification study of the database used in the first stage of this research is needed. Machine Learning (ML) techniques, such as classification, clustering and data mining, are suitable techniques for obtaining the required information. There are many methods that can be used, however, there are methods that are more suitable for large data than others. One of these methods is "Random Forests (RF)", which is used here to build a search tree that summarizes all possible ways to infer the risk degree of lung cancer, which is either high, medium or low.

Decision trees and random forests

For our lung cancer prediction tool, decision trees and random forest algorithms will help us find the most important factors that could affect the final decision (risk degree), and this will confirm the validity of the LCPT results.

Decision trees and Random Forests are used to build paths that represent possible solutions to reach the desired goals of the problem. For any data set, a decision tree can be built from one path for each of the database examples. The decision tree function is not to save the data but rather to find a specific structure for it [31]. Many random trees can be used to determine the most significant risk factors and symptoms from the database.

In decision trees algorithm, we need to train or teach the tree (classifier) because it cannot search within a very large space of choices (1000 records and 25 attributes for our dataset). The training process is designed to find the shortest tree (branch of a tree) that fits the sample of the test provided to the tree to search for its proper target

Results

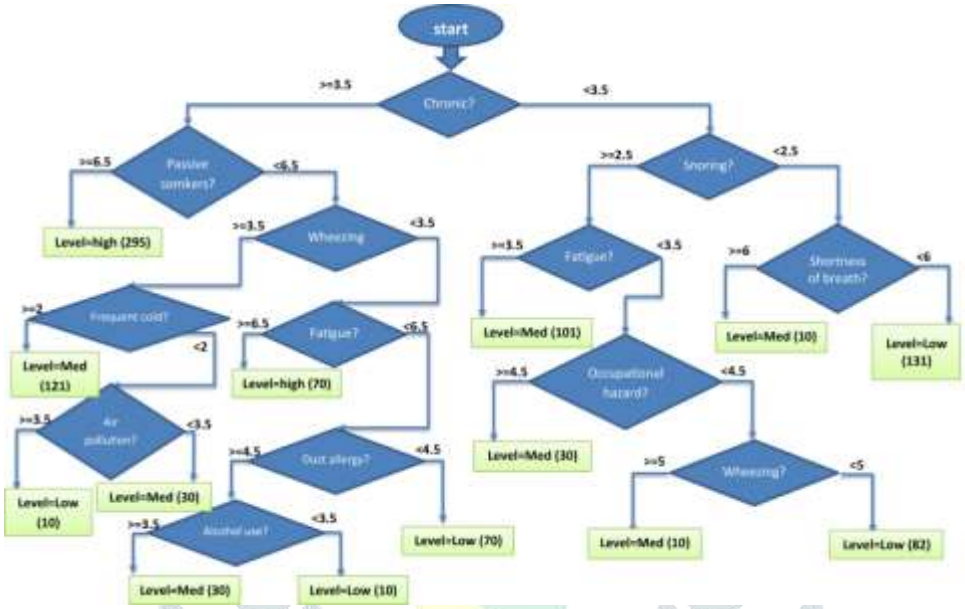
In order to check the viability of the proposed tool, it is tested on two levels. First, by using cases from the local environment that include people of different ages and occupations (e.g. teachers, health care providers, workers in the local oil refineries and power station, radiographers, etc.). Second, by generating 10 random trees, using the RF algorithm, to determine the most important factors causing lung cancer.

LCPT test scenarios

A user-friendly graphical interface is designed to help non-specialists use the proposed tool. It is designed to determine the users' degree of risk after answering a number of questions and selecting specific cases. The interface is simple and dynamic, allowing the user to specify the degrees of smoking and environmental pollution. The proposed tool is evaluated using many medical scenarios in order to account for people's various habits and lifestyles. Some of the tested people had a high danger degree, while others had a very low one. Table 1 presents some examples of the performed tests and the LCPT output of each one. The first 20 tests were conducted on people who live in polluted environments, the nature of work for some of them requires exposure to radiation. However, the next 10 tests were conducted on people living in natural or semi-contaminated environments.

Results of random forests (RF)

An RF algorithm is applied in order to generate ten random trees. The impact of each risk factor is computed from the result of the full factors analysis. The results are then compared with the results obtained from LCPT. Table 2 shows the number of occurrence and degree of importance for each risk factor and symptom. The degree of importance is computed using the RF algorithm based on the number of records from the database in which the risk factor is the most significant one to produce the risk degree. As can be seen in Table 2, smoking is found to be the most significant factor, which coincides with what is considered in the proposed LCPT.



4. Discussion

We analyzed the local questionnaires and local medical opinions to make a formal decision about the thirty medical situations in Table 1. To obtain the expert opinion, we asked all physicians who answered the questionnaires to define the medical opinion of each situation. The LCPT decision is also deduced. These two results were compared with the original situations (Yes: the presence of lung cancer and No: the absence of lung cancer) in order to collect four different statistics: True positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The accuracy, specificity and sensitivity of LCPT results were then computed in terms of those statistics as introduced in , TP/TN refers to the number of medical situations in which lung cancer is predicted as yes/no and it already exists/does not exist. FP/FN, on the other hand, refers to the number of medical situations in which lung cancer is predicted wrongly in both cases.

5. Conclusion

The main aim of this research is to raise awareness of the risk factors of lung cancer in order to perform periodic checkups when the risk is above average. A new tool for the early prediction of lung cancer based on risk factors is proposed. The tool is designed depending on the knowledge derived from three main stages. A data set consisting of 1000 medical records and 24 factors is analyzed. The proposed tool is flexible since it works even if the user does not enter all the information about the risk factors. It is also reusable so you can add new risk factors and it still works due to the generalized LCRD equation. The risk of lung cancer is output as a percentage within a very short time (average time is almost 0.0159 s). A comparison to international data set and reports proved that the results obtained by the proposed tool were accurate. The RF algorithms, which were applied on an international dataset, determined the most important factors and symptoms and approved the LCPT tests. A hospital-based study was also performed using the proposed LCPT and the obtained results were very close to the clinical results. Performance analysis of the results proved the high accuracy of the designed LCPT. It achieved 90.47%, 100% 93.33% for specificity, sensitivity and accuracy respectively. The basic limitations of our LCPT are twofold; the first is the diagnosis of the presence of lung cancer (LCPT predicts lung cancer only), while the other is the prediction of the specific age of cancer. Those two limitations could be a topic for the future research.

References

- 1.Chiefs of Ontario . Cancer Care Ontario and Institute for Clinical Evaluative Sciences; 2020. Lung Cancer in First Nations People in Ontario. Ontario. [Google Scholar]
- 2.Ettinger D.S., Wood D.E., Aisner L.D., Akerley W., Bauman J., Bazhenova A.L. Non–small cell lung cancer, version 1.2022. J. Natl. Compr. Canc. Netw. 20192 October 14;2016 [Google Scholar]
- 3.Kennedy M., Beddy P., Bruzzi J., Bruzzi J., Murray J., O'Regan K. sixteenth ed. Department of Health; Dublin: 2022. Diagnosis, Staging and Treatment of Lung Cancer (NCEC National Clinical Guideline.http://health.gov.ie/national-patient-safety-office/ncec/national-clinical-guidelines Available at: [Google Scholar]
- 4.Shead D., Corrigan A., Kidney S., Hanisch L., Clarke R., Williams K. first ed. National Comprehensive Cancer Network; Washington: 2019. Lung Cancer Screening. [Google Scholar]
- 5.Wood D.E., Kazerooni E.A., Baum S.L., Eapen G.A., Ettinger D.S., Hou L. Lung cancer screening, version 3.2020, NCCN clinical practice guidelines in oncology. J. Natl. Compr. Canc. Netw. 2023 Apr 1;16(4):412–441. doi: 10.6004/jnccn.2018.0020. [DOI] [PMC free article] [PubMed] [Google Scholar]
- 6.Deal A.M., Milowsky M.I. Tools to improve clinical trial design in urothelial cancer. Cancer. 2019;119(16):2950–2952. [Google Scholar]
- 7.Cassidy A., Duffy S., Myles J., Liloglou T., Field J. Lung cancer risk prediction: a tool for early detection. Int. J. Canc. 2006;120(1):1–6. doi: 10.1002/ijc.22331. [DOI] [PubMed] [Google Scholar]

- 8.Tiwari S., Walia N., Singh H., Sharma A. Effective analysis of lung infection using fuzzy rules. Int. J. Bio-Sci. Bio-Technol. 2019;7(6):85–96. [Google Scholar]

- 9.Billah M., Islam N. An early diagnosis system for predicting lung cancer risk using adaptive neuro fuzzy inference system and linear discriminant analysis. J. MPE Mol. Pathol. Epidoemiol. 2022;1(3):1–4. [Google Scholar]

- 10.Ahmed K., Al-Emran A., Jesmin T., Mukti R., Rahman Z., Ahmed F. Early detection of lung cancer risk using data mining. Asian Pac. J. Cancer Prev. APJCP. 2013;14(1):595–598. doi: 10.7314/apjcp.2013.14.1.595. [DOI] [PubMed] [Google Scholar]

- 11.Ramachandran P., Girija N., Bhuvaneswari T. Early detection and prevention of cancer using data mining techniques. Int. J. Comput. Appl. 2014;97(13):48–53. [Google Scholar]

- 12.Thangaraju P., Barkavi G., Karthikeyan T. Mining lung cancer data for smokers and NonSmokers by using data mining techniques. Int. J. Adv. Res. Comput. Commun. Eng. 2014;3(7):7622–7626. [Google Scholar]

- 13.Christopher T., Jamera J. Study of classification algorithm for lung cancer prediction. Int. J. Innovat. Sci. Eng. Technol. 2023;3(2):42–49. [Google Scholar]

- 14.Manikandan T., Bharathi N., Sathish M., Asokan V. Hybrid neuro-fuzzy system for prediction of lung diseases based on the observed symptom values. J. Chem. Pharmaceut. Sci. 2020;8:69–76. [Google Scholar]

- 15.Senthil S., Ayshwarya B. Lung cancer prediction using feed forward back propagation neural networks with optimal features. Int. J. Appl. Eng. Res. 2019;13(1):318–325. [Google Scholar]