Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A new tool to predict lung cancer based on risk factors

Ahmad S. Ahmad [*], Ali M. Mayya

Al Andalus University for Medical Science, Faculty of Medical Engineering, Syria

ARTICLE INFO

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

ABSTRACT

Background: Lung cancer is one of the deadliest cancer 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.
Methods: 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.
Results: The tool is tested using local medical cases and the local medical opinion(s) is (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 with the international ones.
Conclusions: 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.

1. Introduction

Lung cancer is the most dangerous and deadliest type of cancer. Smoking is the basic risk factor for lung cancer [1, 2, 3, 4, 5], and it accounts for 85 out of 100 people dying every year [4]. Although people who do not smoke have a lower risk factor, they may still be affected by the smoke of other smokers [3].

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 [4].

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 [4]. In addition, the history of chronic pulmonary diseases is associated with lung cancer [4, 5].

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 [6]; some of these tools predicted lung cancer [7].

1.1. Previous studies

Lung cancer has been a major concern to researchers in both oncology and the field of medical aid that is based on artificial intelligence. Some studies have designed systems for the detection and diagnosis of lung cancer [8, 9], while others have focused on early lung cancer diagnosis [10, 11, 12, 13, 14].

Studies about the diagnosis of lung cancer have been based on techniques such as fuzzy logic[8] and neural networks [15]. Other studies have used hybrid neuro-fuzzy techniques [9, 14]. However, these methods are unable to construct a valid medical diagnostic system with increasing volumes of databases, making them unreliable. There are studies based on advanced machine learning concepts, such as decision trees [12, 13, 16, 17], which have demonstrated higher reliability compared to those old systems.

Hanai and others introduced prognostic models for Non-Small-Cell Lung Cancer (NSCLC) based on neural networks [18]. They built their models on 125 NSCLC patients with 17 potential input risk factors.

Kattan and Bach introduced a study on the variations in lung cancer risk among smokers based on many factors [19]. They measured the influence of those factors on the risk degree of lung cancer. They found that 15% of men who were over 68 years old and smoked two packs of

[*] Corresponding author.
E-mail address: a.ahmad@au.edu.sy (A.S. Ahmad).

https://doi.org/10.1016/j.heliyon.2020.e03402
Received 11 February 2019; Received in revised form 16 July 2019; Accepted 6 February 2020

cigarettes per day for 50 years and still continued to smoke had lung cancer, while only 0.8% of fifty-one-year-old women who smoked a pack of cigarettes per day for 28 years had lung cancer.

Ramachandran and others built an early prevention system for lung cancer based on data mining in which they used 11 different factors [11]. They conducted experiments using a database consisting of 746 samples, but they did not mention any source for their database. In 2014, Thangaraju and others also used data mining techniques to predict the risk factor of lung cancer [12]. They used Bayes Trees and Decision Table for clustering and classification. The experiments were conducted using 303 samples.

Manikandan and others designed a hybrid neuro-fuzzy system for the prediction of lung cancer based on 11 symptoms [14]. They used 163 samples from a database of 271 individuals (221 medical situations and 50 normal persons).

Arulananth and Bharathi defined the symptoms that can be used for lung cancer prediction [20]. They differentiated between diagnostic factors and the symptoms that indicate the presence of cancer. They defined the diagnostic symptoms by age, sex, family history of cancer, smoking, exposure to radiation, exposure to radon, exposure to chemical subjects and air pollution. On the other hand, they defined the symptoms that indicated the presence of cancer by chronic cough, hemoptysis, chest pain, weight loss, fatigue, chronic lung inflammation, wheezing, swallowing difficulties and anorexia.

In 2018, Senthil and Ayshwaya used neural networks and evolutionary algorithms to define the risk degree of lung cancer based on risk factors [15]. They applied these algorithms to the UCI Global Lung Cancer Database, which consisted of only 32 samples, and the symptoms used were not specific.

Recently, in 2018, Markaki and others built a clinical risk prediction model for lung cancer based on smoking symptoms [17]. They depended on the number of years of smoking, number of cigarettes smoked per day, number of years since the start of no-smoking, weight, height, hours spent in contaminated places, frequent cough, sex and age. Some other studies used advanced machine learning algorithms, such as random trees and random forests, which were very useful for the classification of big databases [21]. On the other hand, others have relied on radiotherapy image processing techniques to determine whether lung cancer is present or not [22]. Other researches focused on the prediction of the mortality of people with NSCLC in the U.S. Military Health System [23].

Cassidy concluded that for building a good lung cancer risk-prediction model, it was preferable to seek other factors in addition to smoking and age [7].

Some previous models did not consider all risk factors and symptoms, and others used a very small database. In this study, a lung cancer prediction tool based on risk factors and their specifications is built. In addition, the symptoms and their effect on lung cancer are studied. Both local and international studies and reports are considered in order to build a powerful international prediction tool. Machine learning techniques are also applied to analyze an international lung cancer database of 1000 records and 23 attributes.

2. Materials and methods

2.1. 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 [24] 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

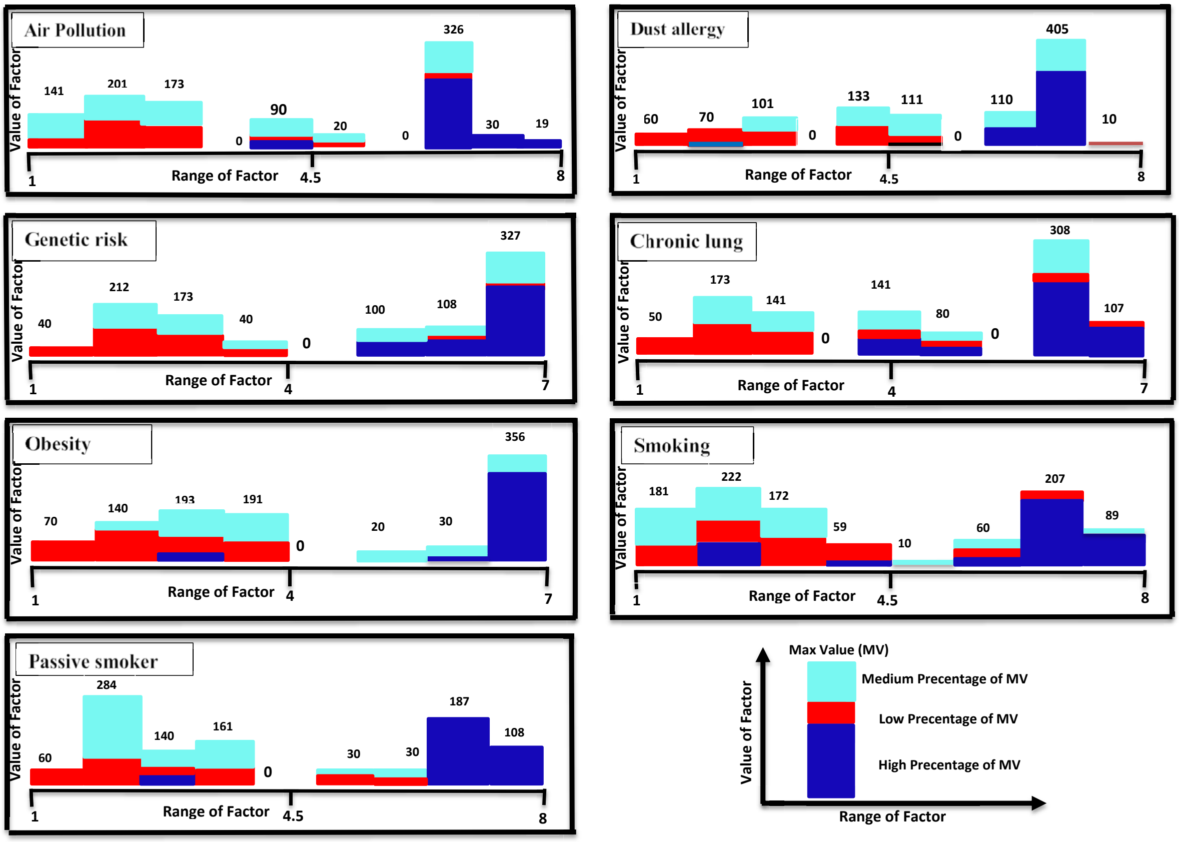

Figure 1. Analysis of Risk Factors of the Studied Database (colors represent risk levels: blue for high, cyan for medium and red for low), numbers above histograms mean cases from 1000 patients of the database.

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.

2.2. 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 [25] is used for the database analysis step.

2.2.1. Analysis of lung cancer risk factors

The analysis of the risk factors is depicted in Figure 1. 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.
- Obesity, as a food-related factor, is a significant risk factor for cancer. Cancerous cells do not have a better environment to flourish in than a body full of fat. It appears that obese people have a higher risk of cancer compared to normal people.
- Smoking & passive smoking are the most significant risk factors. The risk of lung cancer increases significantly with increased smoking or with increased exposure to other people's smoke. There are cases where the rate of smoking is low or a person is not a smoker, yet the risk of cancer is moderate because of other factors.

2.2.2. 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.

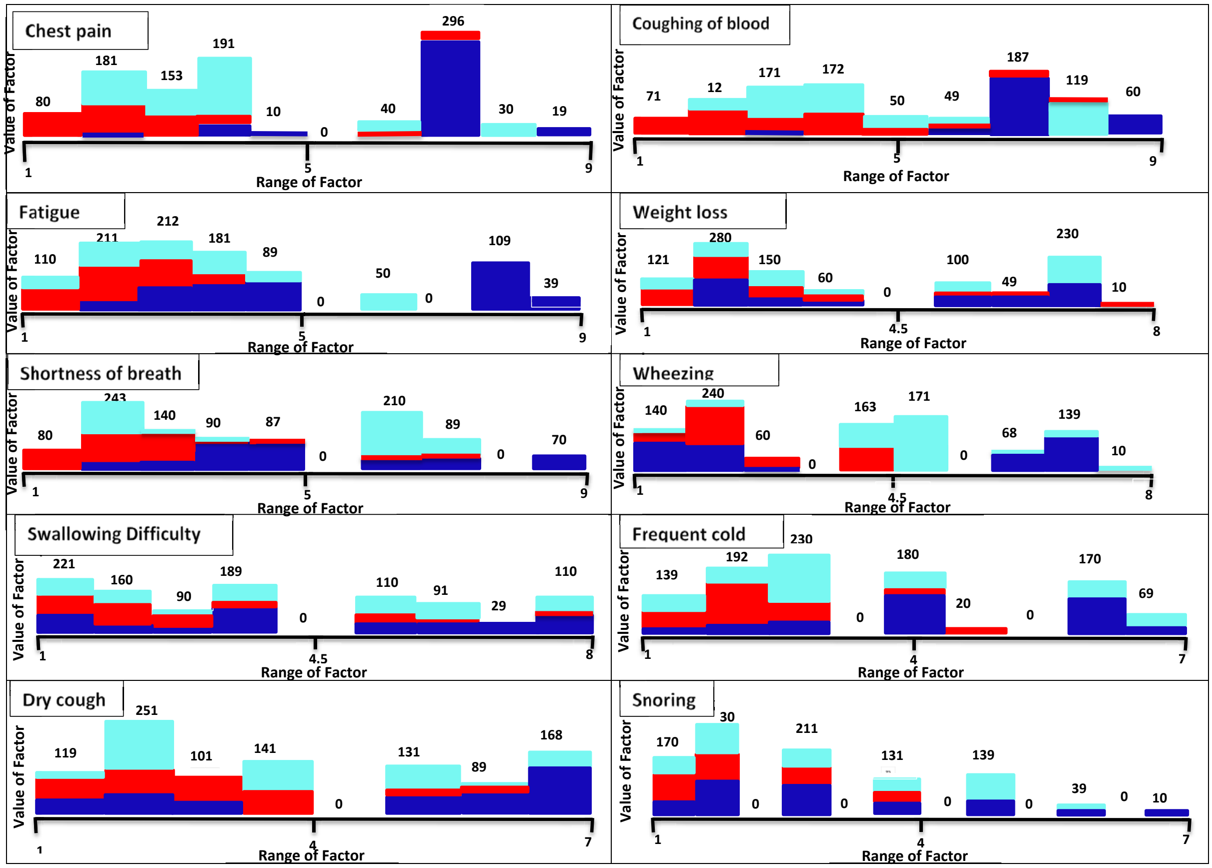

Figure 2. Analysis of Symptoms of the Studied Database (colors represent risk levels: blue for high, cyan for medium and red for low).

- 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 materialize. 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.
- There are cases where snoring can be dangerous especially on the scale of (5–7). This can happen in the presence of other factors that indicate cancer.

2.2.3. 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.

2.3. Questionnaire analysis

Medical questionnaires based on either predictive factors (smoking, pollution, genetic risks, etc.), or on symptoms (chest pain, coughing blood, etc.) are prepared. The questionnaires are distributed among physicians specialized in various fields, such as internal medicine, thoracic surgery, general surgery and oncology.

The results of the questionnaire analysis indicate that the four main risk factors are smoking, exposure to radiation, air pollution and genetic factors, as illustrated in Figure 3A. These factors may be slightly different from those recognized internationally due to the local nature of our environment (a high level of pollution caused by an oil refinery, a power station and a cement plant). Figure 3B presents the statistical results of the symptoms. It shows that coughing blood, fatigue, shortness of breath and weight loss are the most common symptoms of lung cancer.

2.4. 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 [4]. 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 [30]. 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 [4]. 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].

There are many other lung cancer risk factors (e.g. alcohol consumption, age, obesity and type of food [1, 26]), but they have not been considered as important as the above-mentioned ones.

The symptoms that indicate the presence of lung cancer are coughing blood, chest pain, frequent cough, swallowing difficulties, weight loss, wheezing and hoarseness [1, 4]. Some studies have reported that consumption of pickled food could also increase the risk of pulmonary nodules [29].

2.5. The proposed lung cancer prediction tool

Based on the knowledge obtained from previous analyses, a Lung Cancer Prediction Tool (LCPT) is designed and coded.

The designed tool determines the Lung Cancer Risk Degree (LCRD) as Eq. (1) shows.

$$LCRD = \sum_{i=1}^{N} W_i{*}(R_i/R_{max})\ \% \tag{1}$$

where N is the total number of factors. W_i is the weight of i^{th} risk factor, and it is computed in a different way for each factor. The results of database analysis, questionnaires and international medical studies were used to determine the R_i value which defines the number of occurrence of the i^{th} factor during all analysis. R_i is computed as Eq. (2) determines. R_{max} is the total number of occurrence for all factors.

$$R_i = \text{the number of "high" occurrences } + 0.5{*} \text{ the number of "medium" occurrences} \tag{2}$$

To compute each factor's weight, LCPT asks the user about the smoking factor. If the user is a smoker or a passive smoker, then the user is asked about the number of cigarettes per day, the duration of smoking and the period of passive smoking. These three factors are the input to a fuzzy inference system that computes the "smoking weight degree" as a value on a scale of (0–1). The next question to the user is about environmental pollution. If the user's answer is yes, the system asks the user to determine the amount of pollution based on a scale of (0–1) scale (the pollution weight degree). The third question is about exposure to radiation, which determines the number of hours a person is exposed to radiation on a scale of (0–1) (The radiation weight degree). The last three questions are about the family history of lung cancer, alcohol consumption and chronic lung inflammation. If the answer approves the presence of the risk factor, its weight will be equal to one.

2.6. 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.

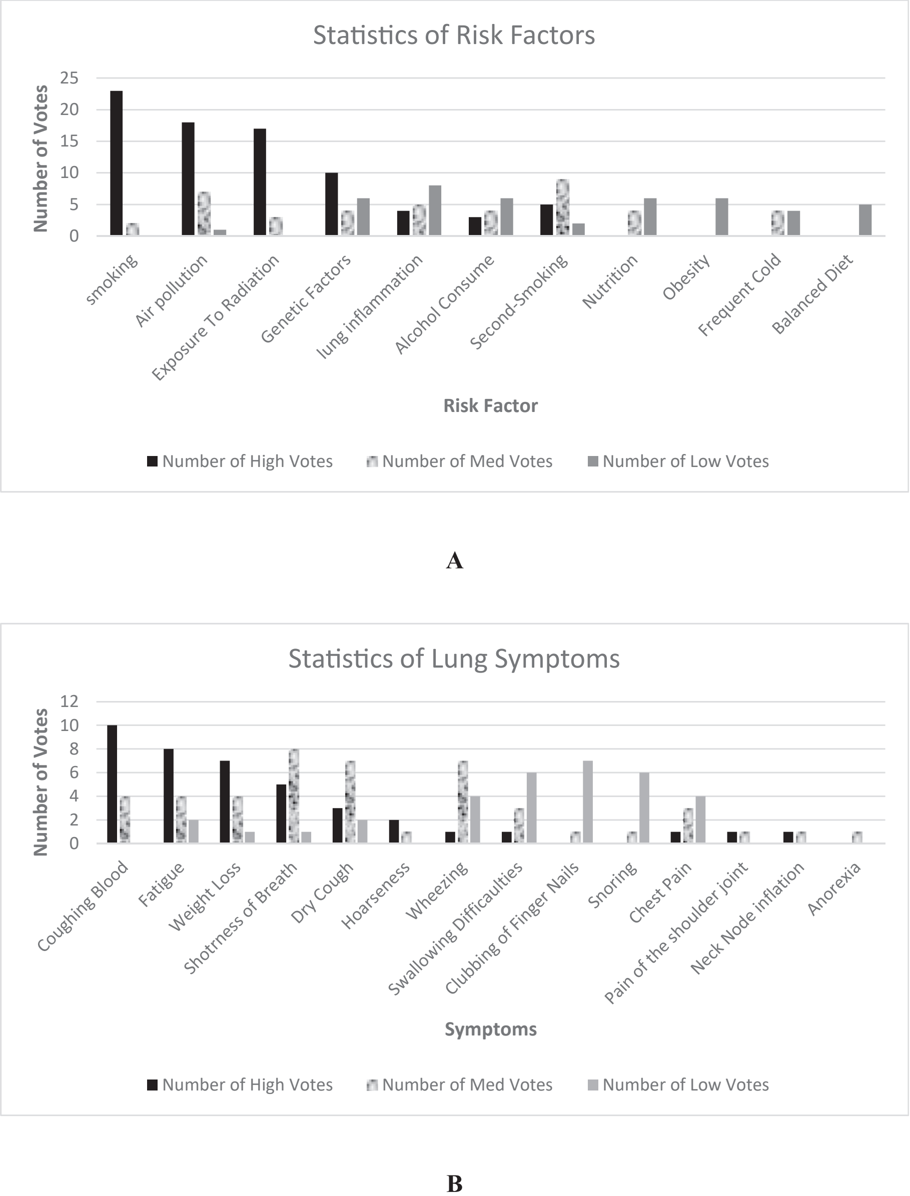

Figure 3. Statistics of symptoms and factors according to local medical questionnaires (A) For factors (B) for symptoms.

2.6.1. *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 [32].

The second idea in the training process is to choose the most important attribute to be the root of the tree from which the search process is proceeded. Therefore, to choose the right attribute, information theory is used, specifically the principle of entropy, which measures the amount of randomness or uncertainty in a statistical distribution. In the case of n class for the studied problem, the entropy is given in terms of the probability p(c) of each attribute (i.e. risk factors and symptoms), illustrated in Eq. (3) [33].

$$Entropy(S) = H(s) = -\sum_{c=1}^{n} p(c)\ \log(p(c))\Big) \tag{3}$$

By using Eq. (3), we could define the Information Gain (IG) which represents the value of initial entropy minus the value of entropy after the distribution of probability on the branches (i.e. after splitting the samples). Therefore, in any tree-based classification issue, entropy can be used to calculate the profit rate of the IG for each branch and, then, the branch that achieves the highest profit rate is chosen.

Table 1. Examples of our LCPT Test Inputs and the Corresponding Results.

A	B	C	D	E	F	G	H	I	J	K
1.	0	0	0	5	0	No	0	0	30	10.1143%
2.	0	0	0	5	0	No	0	0	28	10.1143%
3.	0	0	0	5	0	No	0	0	63	15.1143%
4.	20	20	0	7	0	No	1	0	40	36.4325%
5.	20	20	0	7	7	No	1	0	60	50.5563%
6.	40	40	0	7	4	Yes	1	0	60	66.7183%
7.	15	20	0	7	0	No	1	0	45	30.1998%
8.	40	40	0	7	0	Yes	1	0	65	58.5606%
9.	40	40	0	7	0	Yes	1	1	70	63.5606%
10.	0	0	10	7	0	Yes	1	1	61	52.3547%
11.	10	35	0	5	0	No	0	0	35	22.4141%
12.	5	20	0	7	7	No	1	0	20	30.2949%
13.	0	0	0	7	0	No	0	1	45	19%
14	50	40	0	7	0	No	0	0	67	40.7183%
15.	0	0	5	7	8	No	0	0	62	41.604%
16.	0	0	5	7	8	Yes	1	1	23	61.604%
17.	0	0	10	7	8	Yes	0	1	50	48.3547%
18.	0	0	0	5	0	No	1	0	61	20%
19.	35	20	0	5	5	No	1	1	55	45.8117%
20.	50	40	0	7	0	No	1	1	66	50.5469%
21.	0	0	0	2	0	No	1	0	25	4%
22.	10	40	0	0	0	No	1	0	33	12.7605%
23.	5	24	0	0	4	No	0	0	23	17.7024%
24.	17	20	0	3	0	No	0	0	37	19.5111%
25.	0	0	2	0	0	No	0	0	57	6.37287%
26.	0	0	10	2	0	No	0	0	41	12.3547%
27.	2	8	0	2	0	Yes	0	0	18	26.7227%
28.	0	0	0	0	0	Yes	0	1	62	10%
29.	0	0	0	0	5	No	1	0	27	15%
30.	5	12	0	0	0	No	1	0	46	13.7525%

A: Individual's number, B: Period of smoking (Years), C: Number of cigarettes (per day), D: Period of passive smoking (Hours per day), E: Pollution degree (1–10), F: Expose to radiation (Hours per day), G: Genetic Factor, H: Alcohol consumption (1 for yes, 0 for no), I: Inflammations of lung (1 for very frequent, 0 for little), J: Age, K: LCPT Output.

RF depends on decision trees and has two basic steps: bagging and randomized node optimization (RNO) [33]. At the bagging stage, the training set is randomized, so we get $S_0^t \subset S_0$, that is the randomly sampled subset of training data is made available for the tree t.

In the RNO step, the decision at each node is selected by a randomized procedure. For each node, a set of randomly sampled features is selected T_j T. This process is called "feature bagging". So, if many features (lung-cancer factors) are very good predictors for the target output (lung-cancer risk degree), those features will be chosen in many trees causing them to be correlated. The forest output probability p(c|v) will be the average of all trees' predictions, as Eq. (4) describes [34].

$$p(c|v) = {}^{1}/_{T} \sum_{t=1}^{T} p_t(c|v) \tag{4}$$

where T is the number of all trees and p_t is the probability of output classes given some input test v.

3. 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.

Table 2. The number of occurrence through trees (NOTT) and degree of importance (DOI) of each risk factor and symptom.

Risk Factor	Number of Occurrence Through Trees	Degree of Importance
Smoking	5	0.76
Age	6	0.72
Passive smoking	8	0.67
Balanced diet	3	0.62
Occupational hazards	2	0.61
Air pollution	5	0.61
Genetic risk	12	0.55
Alcohol consumption	10	0.51
Chronic lung disease	5	0.33
Gender	0	0
Obesity	0	0
Dry cough	5	0.54
Wheezing	4	0.49
Snoring	3	0.45
Coughing Blood	3	0.44
Fatigue	7	0.41
Swallowing difficulty	5	0.39
Shortness of breath	6	0.38
Clubbing of finger nails	7	0.37
Chest pain	5	0.34
Weight loss	1	0.17

Table 3. The Sensitivity, Specificity and Accuracy of LCPT results.

	TP	TN	FP	FN	Specificity	Sensitivity	Accuracy
Expert Opinion	8	18	3	1	85.71%	88.88%	86.66%
Our LCPT	9	19	2	0	90.47%	100%	93.33%

3.1. 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.

3.2. 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.

To compute the Final Degree of Importance (FDI) Eq. (5) is used. The FDI that represents both symptoms and risk factors is depicted in Figure 4.

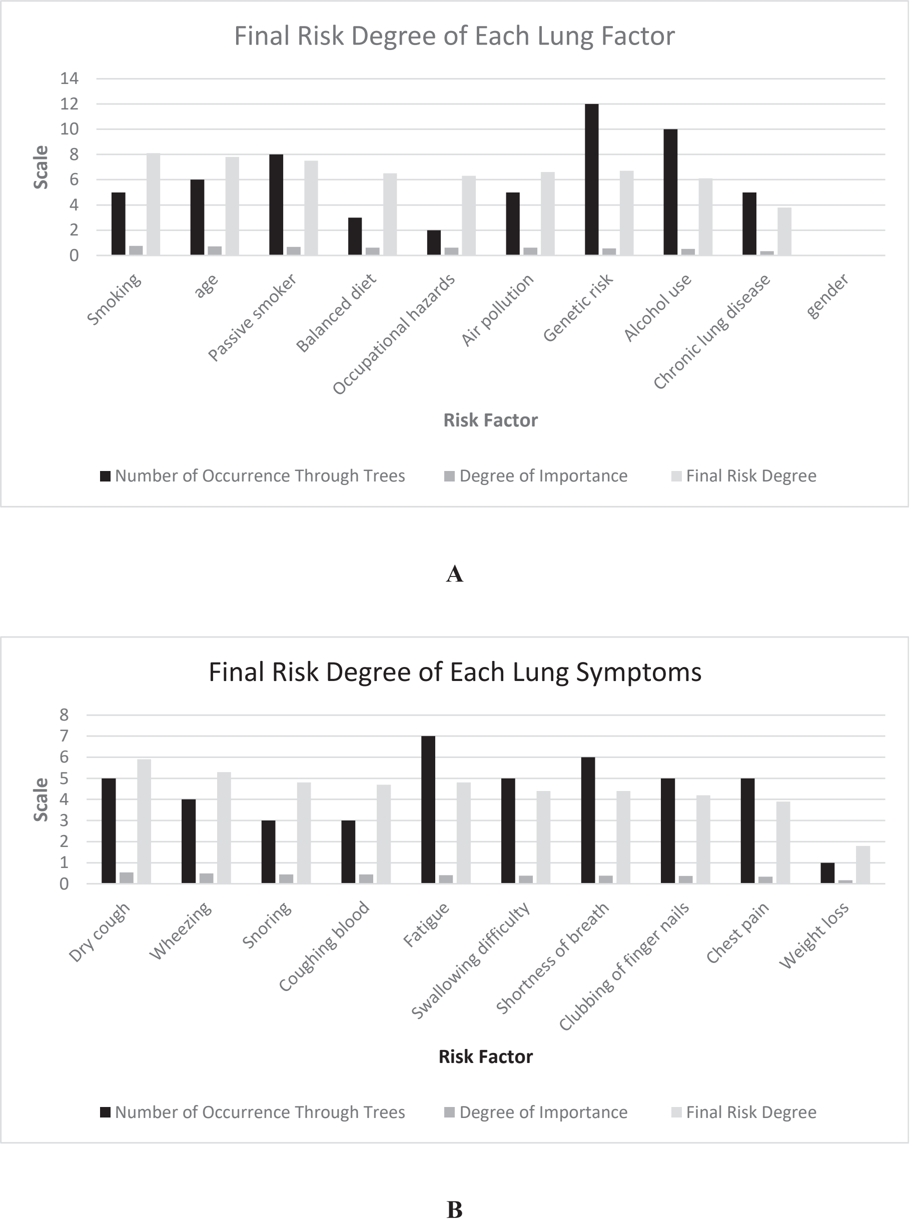

Figure 4. Final risk degree: For factors (A), for symptoms (B).

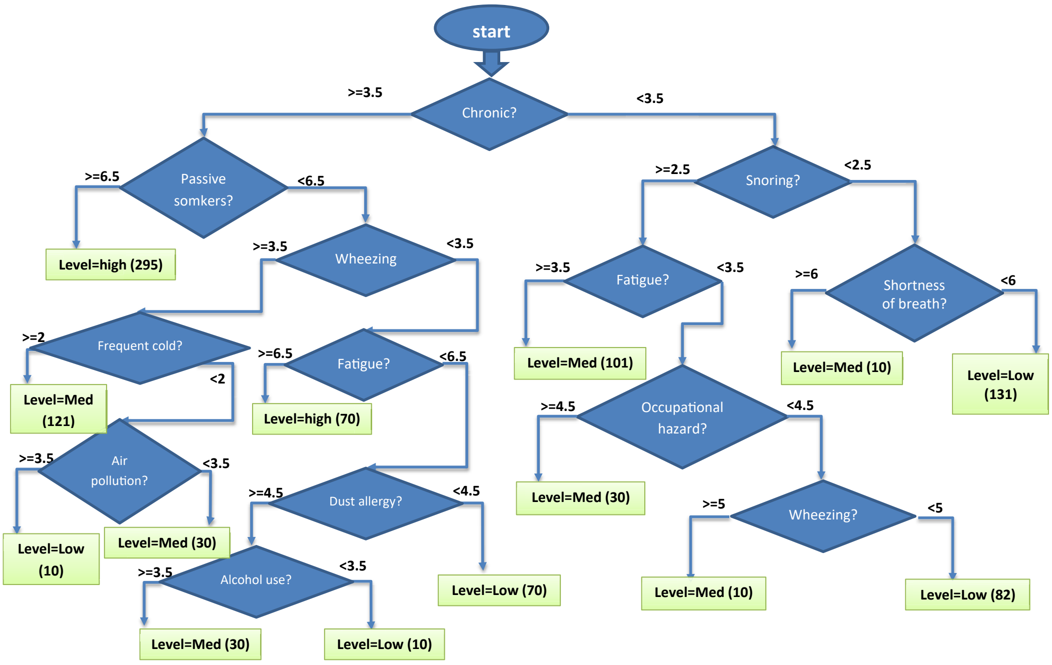

Figure 5. Random tree generated by RF algorithm according to the analysis of the dataset.

$$FDI = DOI + (NOPT)*0.01 \tag{5}$$

The idea behind Eq. (5) is that some factors have high values of Degree of Importance (DOI) but medium or low values of Number of Occurrence Through Trees (NOPT), or vice versa. Therefore, FDI is calculated to mix them correctly. Figure 4A provides information about the most significant risk factors. It shows that smoking, air pollution, genetic risk and occupational hazards have a high final risk degree, which is very close to the results obtained by the proposed tool. Figure 4B information about the most significant symptoms that indicate the presence of lung cancer.

Figure 5 shows a random tree generated by an RF algorithm containing some factors and symptoms with the corresponding risk levels. It also describes the DOI of some factors, such as smoking, fatigue, chronic inflammation, air pollution, and alcohol use.

To increase the reliability of the algorithm, a hospital-based study was considered. The study could only be performed in the presence of lung cancer. The subject of our case was a patient who had been diagnosed with non-small cell lung cancer. The subject had been smoking more than 50 cigarettes a day for more than 40 years, exposed to a highly-polluted environment (he worked in a refinery) and was a heavy drinker. He had neither a history of cancer nor had frequent inflammations, and he was not exposed to any type of radiation. Using the proposed LCPT, the cancer risk of the subject was 52.566%, which is a high-risk level.

There are many cases like this one and, using LCPT, patients can be warned to do proper medical checks when their risk levels are high; the tool, therefore, can rescue people's lives.

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 Eqs. (6), (7), and (8) [35, 36]. 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.

$$Specificity = TN/(\ TN\ +\ FP)*100 \tag{6}$$

$$Sensitivity = \ TP\ /\ (\ TP\ +\ FN)\ *\ 100 \tag{7}$$

$$Accuracy = ((TP + TN)\ /\ (\ TP\ +\ TN +\ FP +\ FN))*100 \tag{8}$$

The best result of any designed system is to get 100% sensitivity in which all individuals from the cancer-presence category are diagnosed as sick and not diagnosing any normal individual as cancer-presence (100% specificity). Accuracy, on the other hand, is a very important metric that refers to the ability of the system to get prediction results with the minimum error rate. In our system, as Table 3 shows, we had 90.47%, 100% and 93.33% for specificity, sensitivity and accuracy respectively, which indicates a high level of performance. From a comparative point of view, our LCPT have performed better than the experts' opinion by 4.76%, 11.12% and 6.67% of specificity, sensitivity and accuracy respectively.

According to RF results, as Table 4 shows, the selected factors of LCPT (which are 8) have a bigger average DOI than the entire twelve factors and the dropped factors. This means that only 66.66% of factors have the highest effect on the detection of lung cancer, which is a good selection of factors provided by our LCPT.

Table 4. Average DOI for each group of factors.

	Entire Factors	Dropped	Selected
Number of Factors	12 (100%)	4 (33.33%)	8 (66.66%)
Average DOI	0.448	0.155	0.595

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.

Declarations

Author contribution statement

A. Mayya: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

A. Ahmad: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Data associated with this paper can be found at the Lung cancer database. URL: https://data.world/cancerdatahp/lung-cancer-data.

Acknowledgements

Part of our work is supported by Al Andalus University and its hospital. The LCPT will be used by them.

References

[1] Chiefs of Ontario, Lung Cancer in First Nations People in Ontario. Ontario, Cancer Care Ontario and Institute for Clinical Evaluative Sciences, 2017.
[2] D.S. Ettinger, D.E. Wood, L.D. Aisner, W. Akerley, J. Bauman, A.L. Bazhenova, et al., Non–small cell lung cancer, version 1.2017, J. Natl. Compr. Canc. Netw. 2016 (2016 October 14).
[3] M. Kennedy, P. Beddy, J. Bruzzi, J. Bruzzi, J. Murray, K. O'Regan, et al., Diagnosis, Staging and Treatment of Lung Cancer (NCEC National Clinical Guideline, sixteenth ed., Department of Health, Dublin, 2017. Available at: http://health.gov.ie/national-patient-safety-office/ncec/national-clinical-guidelines.
[4] D. Shead, A. Corrigan, S. Kidney, L. Hanisch, R. Clarke, K. Williams, Lung Cancer Screening, first ed., National Comprehensive Cancer Network, Washington, 2017.
[5] D.E. Wood, E.A. Kazerooni, S.L. Baum, G.A. Eapen, D.S. Ettinger, L. Hou, et al., Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology, J. Natl. Compr. Canc. Netw. 16 (4) (2018 Apr 1) 412–441.
[6] A.M. Deal, M.I. Milowsky, Tools to improve clinical trial design in urothelial cancer, Cancer 119 (16) (2013) 2950–2952.
[7] A. Cassidy, S. Duffy, J. Myles, T. Liloglou, J. Field, Lung cancer risk prediction: a tool for early detection, Int. J. Canc. 120 (1) (2006) 1–6.
[8] S. Tiwari, N. Walia, H. Singh, A. Sharma, Effective analysis of lung infection using fuzzy rules, Int. J. Bio-Sci. Bio-Technol. 7 (6) (2015) 85–96.
[9] M. Billah, N. Islam, An early diagnosis system for predicting lung cancer risk using adaptive neuro fuzzy inference system and linear discriminant analysis, J. MPE Mol. Pathol. Epidoemiol. 1 (3) (2016) 1–4.
[10] K. Ahmed, A. Al-Emran, T. Jesmin, R. Mukti, Z. Rahman, F. Ahmed, Early detection of lung cancer risk using data mining, Asian Pac. J. Cancer Prev. APJCP 14 (1) (2013) 595–598.
[11] P. Ramachandran, N. Girija, T. Bhuvaneswari, Early detection and prevention of cancer using data mining techniques, Int. J. Comput. Appl. 97 (13) (2014) 48–53.
[12] P. Thangaraju, G. Barkavi, T. Karthikeyan, Mining lung cancer data for smokers and NonSmokers by using data mining techniques, Int. J. Adv. Res. Comput. Commun. Eng. 3 (7) (2014) 7622–7626.
[13] T. Christopher, J. Jamera, Study of classification algorithm for lung cancer prediction, Int. J. Innovat. Sci. Eng. Technol. 3 (2) (2016) 42–49.
[14] T. Manikandan, N. Bharathi, M. Sathish, V. Asokan, Hybrid neuro-fuzzy system for prediction of lung diseases based on the observed symptom values, J. Chem. Pharmaceut. Sci. 8 (2017) 69–76.
[15] S. Senthil, B. Ayshwarya, Lung cancer prediction using feed forward back propagation neural networks with optimal features, Int. J. Appl. Eng. Res. 13 (1) (2018) 318–325.
[16] S. Durga, K. Kasturi, Lung disease prediction system using data mining techniques, J. Adv. Res. Dyn. Control Sys. 9 (5) (2017) 62–66.
[17] M. Markaki, I. Tsamardinos, A. Langhammer, V. Lagani, K. Hveem, O.D. Røe, A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a hunt study, EBioMedicine 31 (2018) 36–46.
[18] T. Hanai, Y. Yatabe, Y. Nakayama, et al., Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression, Canc. Sci. 94 (5) (2003) 473–477.
[19] P. Bach, M. Kattan, M. Thornquist, et al., Variations in lung cancer risk among smokers, J. Natl. Cancer Inst. 95 (6) (2003) 470–478.
[20] H. Bharathi, T.S. Arulananth, A review of lung cancer prediction system using data mining techniques and self organizing map (SOM), Int. J. Appl. Eng. Res. 12 (10) (2017) 2190–2195.
[21] A. Mishra, B. Ratha, Study of random tree and random forest data mining algorithms for microarray data analysis, Int. J. Adv. Elcetric. Comput. Eng. 3 (4) (2016) 5–7.
[22] M. Saii, A. Mia, Lung detection and segmentation using marker watershed and laplacian filtering, Int. J. Biomed. Eng. Clin. Sci. 1 (2) (2015) 29–42.
[23] C. Jeong, S. Jeong, S. Hong, et al., Nomograms to predict the pathological stage of clinically localized prostate cancer in Korean men: comparison with Western predictive tools using decision curve analysis, Int. J. Urol. 19 (9) (2012) 846–852.
[24] [Database] Lung cancer database. Available at URL: https://data.world/cancerdatahp/lung-cancer-data. [accessed January 10, 2018].
[25] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016 Oct 1.
[26] The Whole Report: World Cancer Research Fund/American Institute for Cancer Research. Available at URL dietandcancerreport.org. [Accessed March 15, 2018].
[27] Non-Small Cell Lung Cancer Report, National Comprehensive Cancer Network. Available at URL https://www.nccn.org/patients/[Accessed March 15, 2018].
[28] Lung Cancer Report/Non-Small Cell Lung Cancer, National Comprehensive Cancer Network. Available at URL https://www.nccn.org/patients/. [Accessed April 10, 2018].
[29] Y. He, Y. Zhang, G. Shi, et al., Risk factors for pulmonary nodules in north China: a prospective cohort study, Lung Canc. 120 (2018) 122–129.
[30] S. Garinet, P. Laurent-Puig, H. Blons, J. Oudart, Current and future molecular testing in NSCLC, what can we expect from new sequencing technologies? J. Clin. Med. 7 (6) (2018) 1–23.
[31] S. Hegelich, Decision trees and random forests: machine learning techniques to classify rare events, J EPA 2 (1) (2016) 98–120.
[32] C. Kingsford, S.L. Salzberg, What are decision trees? Nat. Biotechnol. 26 (9) (2008 Sep) 1011–1013.
[33] Zhao T, Lecture 6: Decision Tree, Random Forest, and Boosting, Schools of ISyE and CSE, Georgia Tech.
[34] Pulli K, Machine Learning for Vision: Random Decision Forests and Deep Neural Networks, VP Computational Imaging.
[35] Ch Kumar, Fuzzy Clustering-Based Formal Concept Analysis for Association Rules Mining, Applied Artificial Intelligence, 2012, pp. 274–301.
[36] W. Zhu, N. Zeng, N. Wang, Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations, in: NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland, 2010 Nov 14.