International Journal of
Environmental Research and Public Health

MDPI

Article

Determining Association between Lung Cancer Mortality Worldwide and Risk Factors Using Fuzzy Inference Modeling and Random Forest Modeling

Xiu Wu [1], Blanchard-Boehm Denise [1], F.Benjamin Zhan [1] and Jinting Zhang [2,*]

[1] Department of Geography and Environmental Studies, Texas State University, San Marcos, TX 78666, USA
[2] School of Resource and Environmental Science, Wuhan University, Wuhan 430070, China
* Correspondence: whuzjt@whu.edu.cn

Abstract: Lung cancer remains the leading cause for cancer mortality worldwide. While it is well-known that smoking is an avoidable high-risk factor for lung cancer, it is necessary to identify the extent to which other modified risk factors might further affect the cell's genetic predisposition for lung cancer susceptibility, and the spreading of carcinogens in various geographical zones. This study aims to examine the association between lung cancer mortality (LCM) and major risk factors. We used Fuzzy Inference Modeling (FIM) and Random Forest Modeling (RFM) approaches to analyze LCM and its possible links to 30 risk factors in 100 countries over the period from 2006 to 2016. Analysis results suggest that in addition to smoking, low physical activity, child wasting, low birth weight due to short gestation, iron deficiency, diet low in nuts and seeds, vitamin A deficiency, low bone mineral density, air pollution, and a diet high in sodium are potential risk factors associated with LCM. This study demonstrates the usefulness of two approaches for multi-factor analysis of determining risk factors associated with cancer mortality.

Keywords: lung cancer mortality; lung cancer mortality-non-smoking factors; health; fuzzy inference modeling; random forest modelling

check for updates

Citation: Wu, X.; Denise, B.-B.; Zhan, F.; Zhang, J. Determining Association between Lung Cancer Mortality Worldwide and Risk Factors Using Fuzzy Inference Modeling and Random Forest Modeling. *Int. J. Environ. Res. Public Health* **2022**, *19*, 14161. https://doi.org/10.3390/ijerph192114161

Academic Editor: Paul B. Tchounwou

Received: 5 October 2022
Accepted: 26 October 2022
Published: 29 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lung cancer is defined by the World Health Organization's (WHO), "International Classification of Diseases", as a malignant neoplasm (tumor) of the trachea, bronchus, and/or lungs. About 98 to 99 percent of lung cancers are carcinomas, thus, this disease is often referred to as lung carcinoma [1]. Based on information from WHO's GLOBOCAN 2020 estimates of cancer incidence and mortality, lung cancer remains the leading cause of cancer-related deaths in the world, with an estimated 1.8 million deaths (18 percent) worldwide. Lung cancer has an incidence ratio of 1 out of 10 of newly diagnosed cancer cases, and a mortality rate of 1 out of 5 deaths worldwide [2,3].

More recently, the medical community continues to report that, worldwide, approximately 85 percent of lung cancer deaths are due to long-term smoking [4]. The etiology in this regard has been well-established over decades of research—it is the remaining 15 percent of deaths of non-smokers that has left the research community puzzled for the past six decades. Potential risk factors for lung cancer in this regard include passive smoking, radon gas, asbestos, aerosols from mining and metal processing, combustion (indoor emissions, exhaust, and petroleum processing), ionizing radiation, toxic gasses, rubber production, and silica processing. Further, Turner and colleagues also observe that, in more recent times, emissions from industry, power generation, transportation, and domestic burning, exceed considerably the WHO's health-based air-quality guidelines and subject the world's population to unsafe levels of air pollution. They suggest that outdoor air pollution is an urgent worldwide public health challenge, particularly in relation to cancers [5].

The question of how to compare the impacts of human diet, habits, and environmental risk factors on lung cancer is still a challenging task. To fill these knowledge gaps, we used fuzzy inference of weight and a Random Forest Tree (RFM) model to assess the weights of the aforementioned variables in association with lung cancer mortality (LCM). The common characteristics or determinants are revealed in the comparison of the two methods in the analysis. For decades, previous studies have mainly focused on the genetic and biological aspects of lung cancer, as well as the efficacy of medical surgeries and treatment protocols for lung carcinomas [6–8]. Moreover, the relationship between demographic and socioeconomic variables—gender, age structure, race, income, diet and food access, level of a country's development—have been extensively researched [9–14]. However, in more recent times, researchers such as Turner and colleagues have been focusing on the association between environmental risk factors and lung cancer. Yet, their focus is limited to one aspect of lung cancer's incidence and mortality (i.e., indoor and outdoor pollution). These diverse perspectives, no matter the focus, tend to be localized in one specific geographical region and, findings are hence limited to specific geographic regions [15–18]. Therefore, there is a need to investigate if other factors such as human diet and habits, and physical and mental health variables are associated with LCM. In addition, this study utilizes machine learning analysis and fuzzy inference rather than multivariate statistics to explore the association between these factors and LCM.

2. Materials and Methods

2.1. Data

We collected lung cancer mortality data and mortality data associated with 30 risk factors in 100 countries during the period from 2006 to 2016, from open datasets published by the Global Burden of Disease (https://vizhub.healthdata.org/gbd-results/ (accessed on 4 October 2022). The dataset contained 1097 observations. Mortality associated with a risk factor is simply the estimated number of deaths associated with each of the 30 risk factors. These 30 risk factors are listed in Table 1.

Table 1. List of risk factors of a country used as independent variables in the analyses.

NO.	Variable Name	Acronym	Description
1	Disease burden	DB	Mortality rate from disease burden
2	Tuberculosis	TB	Mortality rate from Tuberculosis
3	Unsafe water source	UWS	Mortality rate from unsafe water source
4	Unsafe sanitation	USN	Mortality rate from unsafe sanitation
5	No access to handwashing facility	NHF	Mortality rate from no access to handwashing facility
6	Household air pollution from solid fuels	HAP	Mortality rate from household air pollution from solid fuel
7	Non-exclusive breastfeeding	NEB	Mortality rate from non-exclusive breastfeeding
8	Discontinued breastfeeding	DBF	Mortality rate from discontinued breastfeeding
9	Child wasting	CW	Mortality rate from child wasting
10	Child stunting	CS	Mortality rate from child stunting
11	Low birth weight due to short gestation	LBW	Mortality rate from low birth weight due to short gestation
12	Secondhand smoke	SHS	Mortality rate from secondhand smoke
13	Alcohol use	AU	Mortality rate from alcohol use
14	Drug use	DU	Mortality rate from drug use
15	Diet low in fruits	DLF	Mortality rate from diet low in fruits

Table 1. *Cont.*

NO.	Variable Name	Acronym	Description
16	Diet low in vegetables	DLV	Mortality rate from diet low in vegetable
17	Unsafe sex	USX	Mortality rate from unsafe sex of a country
18	Low physical activity	LPA	Mortality rate from low physical activity
19	High fasting plasma glucose	HFP	Mortality rate from high fasting plasma glucose
20	High body-mass index	HBM	Mortality rate from high body-mass index
21	High systolic blood pressure	HBP	Mortality rate from high systolic blood pressure
22	Iron deficiency	IDY	Mortality rate from iron deficiency
23	Smoking		Mortality rate from smoking
24	Vitamin A deficiency	VAD	Mortality rate from Vitamin A deficiency
25	Low bone mineral density	LBD	Mortality rate from low bone mineral density
26	Air pollution	AP	Mortality rate from air pollution
27	Outdoor air pollution	OAP	Mortality rate from outdoor air pollution
28	Diet high in sodium	DHS	Mortality rate from diet high in sodium
29	Diet low in whole grains	DLG	Mortality rate from diet low in whole grains
30	Diet low in nuts and seeds	DLN	Mortality rate from diet low in nuts and seeds

2.2. Analysis Procedure

We treated LCM as the dependent variable and the 30 risk factors as independent variables in the analysis. We first computed the crude mortality rate for each variable in each of the 100 countries. The crude rate is the rate between the number of deaths associated with each risk factor divided by the average population size in each country from 2006 to 2016. We then classified the countries into five risk levels for each variable using quintiles. From the lowest quintile to the highest quintile, all 100 countries were classified into very low risk, low risk, medium risk, high risk, and very high risk in LCM, and in each of the other 30 variables. Figure 1 shows the distribution of LCM risk levels of these 100 countries in the world. Second, we performed analyses using the fuzzy inference modeling (FIM) and the random forest modelling (RFM) shown in Figure 2. In FIM, we implemented analysis using the Analytical Hierarchical Process (AHP), the RIDIT analysis, and the Chi-square analysis to search for the optimal lattice degree based on nearness. Third, we conducted the RFM. Fourth, we selected the optimal weighting group based on the results of the optimal lattice degree on nearness and compared the results from RFM. Fifth, we determined the similarity and dissimilarity in the comparison of the two approaches. Figure 2 illustrates the overall analysis framework. We provided a brief description of the fuzzy inference modeling below. Details of the RFM can be found in the article by Du and colleagues [19].

We used ArcGIS Pro 3.0 to create Figure 1 based on the World Map with Polyconic Projection with Meridional Interval on Same Parallel Decrease Away from Central Meridian by Equal Difference. ArcGIS Pro is a desktop GIS software developed by Esri, which replaces their ArcMap software generation. The product was announced as part of Esri's ArcGIS 10.3 release. ArcGIS Pro is notable in having a 64-bit architecture, combined 2-D and 3-D support, ArcGIS Online integration and Python 3 support.

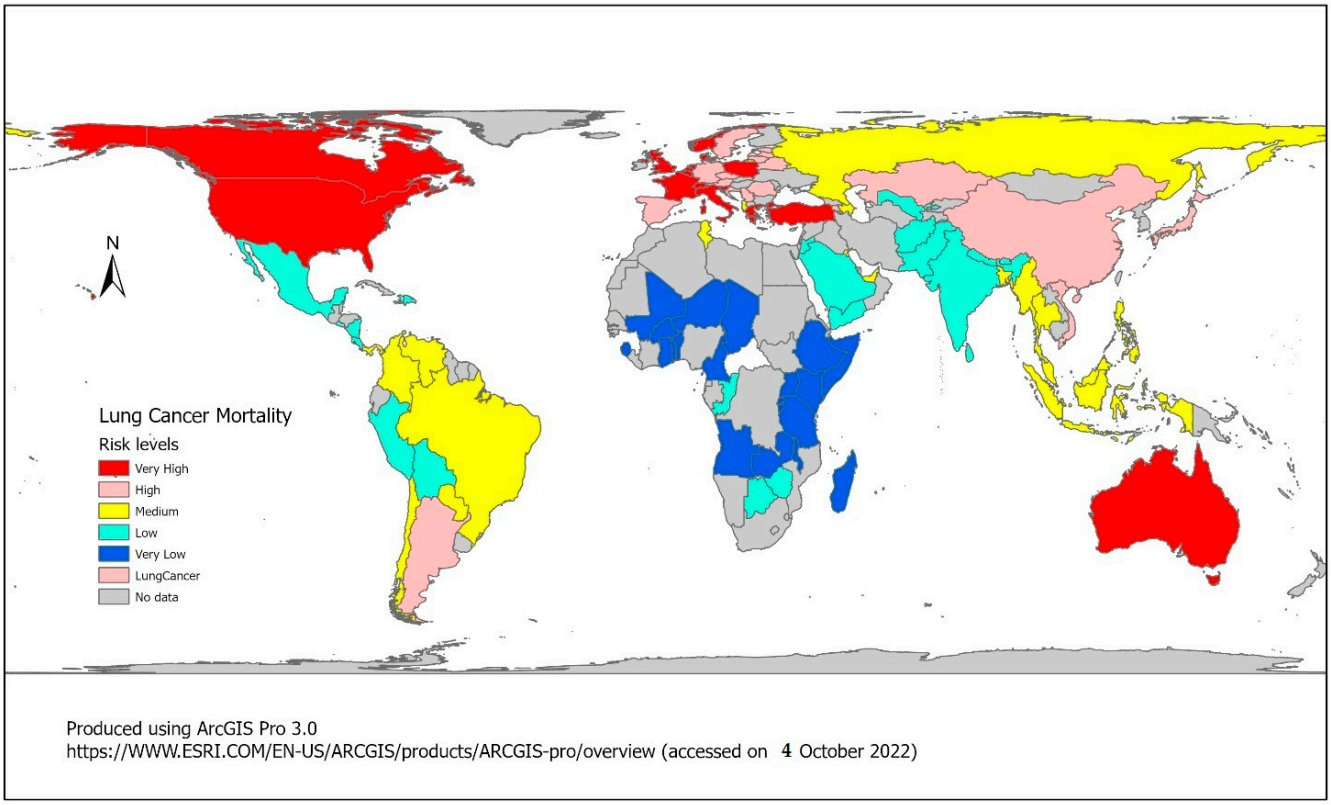


Figure 1. Risk levels of LCM in 100 countries from 2006 to 2016.

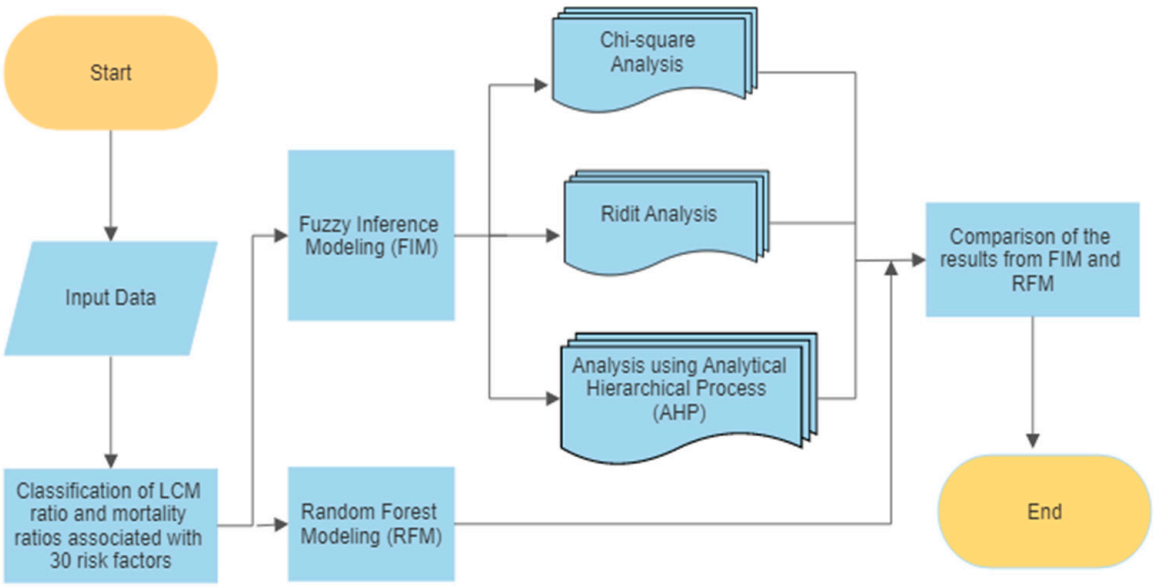


Figure 2. Analysis framework.

2.3. Fuzzy Inference Methods

The FIM is based on fuzzy logic that is used to make decisions on imprecise information. Since fuzzy logic originates from fuzzy set theory where reasoning is approximate, the fuzzy inference is used in the field of anomaly detection so that all variables are viewed as fuzzy variables [20]. The strengths of these methods include its capacity of modeling non-linearity efficiently, segregating normal and anomalous samples, and better predicting the inconsistencies [21]. The application of these methods in this study can be considered as the auxiliary validation of machine learning-based medical anomaly detection that is

related to the purpose of prediction and diagnosis, in addition to the medical data analyzed by machine learning. The procedure of fuzzy inference used in this study includes the 10 steps listed below [22].

(1). Filter out unrelated variables and accept Table 1 to establish the LCM index system. In this research, we selected the 30 risk factors as the variables related to LCM.

(2). Establish all factors as vector-matrix (U) and use the five risk levels of LCM as the five classes in the analysis (V).

(3). Generate the fuzzy similar matrix (R) among the 100 countries using the product of U times V based on the formula below.

$$\mathrm{R} = \begin{Bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{Bmatrix} \times \begin{Bmatrix} v_{11} & v_{12} & & v_{1n} \\ v_{21} & v_{22} & & v_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ v_{m1} & v_{m2} & & v_{mn} \end{Bmatrix} \quad (1)$$

(4). Implement the Chi-square test, generate chi-square value and weight, and normalize the weights to obtain the first set of weights-A1.

(5). Perform RIDIT analysis to obtain the second set of weights-A2. RIDIT values are generally based on the observed distribution of a response variable for a specified set of individuals [23]. This approach is very closely related to distribution-free methods based on ranks such as the Wilcoxon Test [24]. RIDIT possesses two very important properties. First, it assigns a rank value to each class proportional to the relative frequency of observations in that class. Second, it standardizes the rank values to vary between 0 and 1. The latter property eliminates the problem of variation in the relative positions with respect to the number of ranks. RIDIT technique appears to suppress the differences in distributional shape [25].

(6). Perform analysis using the AHP method to obtain the third set of weights-A3.

(7). Use B1 = A1 × R, B2 = A2 × R, B3 = A3 × R to obtain the values of B1, B2, and B3, and then normalize these values.

(8). Compute the lattice degree of nearness σ using Equation (2) below based on the original weights C.

$$\sigma = \frac{1}{2}[B_i \otimes C + (1 - B_i \odot C)] \quad (2)$$

where $B_i \otimes C = \bigvee_{\chi \in v} (\mathrm{C(x)} \wedge B(x))$ and $B_i \odot C = \bigwedge_{\chi \in v} (\mathrm{C(x)} \vee B(x))$, i = 1, 2, 3.

(9). Obtain an optimal lattice degree of nearness using fuzzy pattern recognition and

$$\text{the formula : } \sigma_{optimal} = max(\sigma_1, \sigma_2, \sigma_3) \quad (3)$$

(10). Select the optimal weight group based on the results from the three analytic approaches.

3. Results

3.1. Fuzzy Inference Modeling

3.1.1. Chi-Square Analysis

Initially, 30 variables were used and passed the chi-square test. We used disease burden as an example in Table 2.

H_0: *disease burden with very low, low, middle, high, and very high risk are independent from the LCM rate,*

H_α: *they are dependent.*

We calculated the observed and expected values to obtain five levels of disease burden mortality rates and examine the Chi-square test. The *p*-value was less than 0.001 after we compared observed results with expected results, leading to the rejection of H_0. We hence concluded that different risk levels of LCM caused by disease burden were not independent

of each other (H_a). Once the disease burden passed the test, we computed the χ^2 values of the disease burden to standardize the χ^2 weights, which were shown in Table 3.

Table 2. Process of the Chi-square analysis.

Observed Values						
variable	very low risk	low risk	medium risk	high risk	very high risk	total rate
DB	7.199	31.0088	63.8118	109.3561	169.6175	380.9932
non-DB	49.1659	109.6643	156.3428	188.1021	212.7317	716.0068
total	56.3649	140.6731	220.1546	297.4582	382.3492	1097
Expected Values						
variable	very low risk	low risk	medium risk	high risk	very high risk	total
DB	19.5758	48.8564	76.4607	103.3086	132.7917	380.9932
non-DB	36.7891	91.8167	143.6939	194.1496	249.5575	716.0068
total	56.3649	140.6731	220.1546	297.4582	382.3492	1097
chi-square test						<0.001
χ^2 Values						
variable	very low risk	low risk	medium risk	high risk	very high risk	total
DB	7.8252	6.5199	2.0925	0.3540	10.2125	27.0041
non-DB	4.1639	3.4693	1.1134	0.1884	5.4342	14.3691
Total	11.9891	9.9891	3.2059	0.5424	15.6467	41.3733

Table 3. Results of Chi-square analysis.

Variable	χ^2 Value	χ^2 Weight
Disease burden	41.37	0.0030
Tuberculosis	516.68	0.0372
Unsafe water source	523.82	0.0377
Unsafe sanitation	533.08	0.0384
No access to handwashing facility	532.60	0.0383
Household air pollution from solid fuels	497.60	0.0358
Non-exclusive breastfeeding	553.70	0.0399
Discontinued breastfeeding	581.37	0.0418
Child wasting	646.50	0.0465
Child stunting	545.02	0.0392
Low birth weight due to short gestation	581.04	0.0418
Secondhand smoke	522.60	0.0376
Alcohol use	451.03	0.0325
Drug use	531.80	0.0383
Diet low in fruits	491.40	0.0354
Diet low in vegetables	506.60	0.0365
Unsafe sex	449.78	0.0324
Low physical activity	514.90	0.0371
High fasting plasma glucose	246.80	0.0178

Table 3. *Cont.*

Variable	χ^2 Value	χ^2 Weight
High body-mass index	331.50	0.0239
High systolic blood pressure	160.50	0.0116
Iron deficiency	281.30	0.0202
Smoking	586.03	0.0422
Vitamin A deficiency	595.18	0.0428
Low bone mineral density	470.50	0.0339
Air pollution	370.06	0.0266
Outdoor air pollution	453.13	0.0326
Diet high in sodium	473.13	0.0341
Diet low in whole grains	429.40	0.0309
Diet low in nuts and seeds	475.24	0.0342
Total	13,893.66	1.0000

3.1.2. RIDIT Analysis

First, we determined the frequencies of the five levels associated with the 30 risk factors. The frequencies in the five levels from very low risk through very high risk are 219, 221, 217, 221, and 219 among the 1097 values. Next, we calculated the mean of each risk level. These five means are 109.5, 110.5, 108.5, 110.5 109.5, respectively. Third, we computed the RIDIT value related to each of the five levels. These values are 0.0998, 0.3004, 0.5000, 0.6996, and 0.9002 from the lowest risk level to the highest risk level (Table 4). In the fourth and final step, we calculated the RIDIT values of the 30 risk factors and standardized weights shown in Table 5.

Table 4. Process of RIDIT analysis.

Level	(1) Frequency	(2) Average	(3) Cumulative	(4) Sum	RIDIT Value
				(2) + (3)	(4)/Total Ratio
Very low risk	219	109.5	0	109.5	0.0998
Low risk	221	110.5	219	329.5	0.3004
Medium risk	217	108.5	440	548.5	0.5000
High risk	221	110.5	657	767.5	0.6996
very high risk	219	109.5	878	987.5	0.9002
Total ratio	1097				

Table 5. Results of RIDIT analysis.

Variable	RIDIT Value	RIDIT Weight
Disease burden	0.247	0.244
Tuberculosis	0.018	0.018
Unsafe water source	0.020	0.020
Unsafe sanitation	0.015	0.015
No access to handwashing facility	0.014	0.014
Household air pollution from solid fuels	0.024	0.024
Non-exclusive breastfeeding	0.004	0.004

Table 5. *Cont.*

Variable	RIDIT Value	RIDIT Weight
Discontinued breastfeeding	0.001	0.000
Child wasting	0.028	0.028
Child stunting	0.006	0.006
Low birth weight due to short gestation	0.021	0.021
Secondhand smoke	0.011	0.011
Alcohol use	0.034	0.034
Drug use	0.006	0.006
Diet low in fruits	0.022	0.022
Diet low in vegetables	0.015	0.015
Unsafe sex	0.041	0.041
Low physical activity	0.014	0.014
High fasting plasma glucose	0.076	0.076
High body-mass index	0.057	0.057
High systolic blood pressure	0.107	0.106
Iron deficiency	0.072	0.071
Smoking	0.001	0.001
Vitamin A deficiency	0.007	0.007
Low bone mineral density	0.003	0.003
Air pollution	0.043	0.043
Outdoor air pollution	0.027	0.026
Diet high in sodium	0.024	0.024
Diet low in whole grains	0.032	0.032
Diet low in nuts and seeds	0.022	0.022
Total	1.012	1.000

3.1.3. AHP Analysis

We used the Delphi method in Table 6 to investigate the roles of the 30 independent variables in LCM and produced the AHP weights for LCM in Table 7. In the results, living environment qualities such as outdoor air pollution and air pollution were classified as the first class. Smoking and secondhand smoke related to human behavior were viewed as the second class. Diet health of patients was the third-class, including diet high in sodium, low bone mineral density, diet low in whole grains, and diet low in vegetables and fruits. Patient physical and mental health indices (e.g., disease burden) were assigned as the fourth class. Childhood health, such as low birth weight due to short gestation, was assigned as the class with the least impact. The results of the AHP analysis show the importance of environmental variables on LCM (Table 7).

Table 6. Results of the Delphi method.

Variable	DB	TB	UWS	USN	NHF	HAP	NEB	DBF	CW	CS	LBW	SHS	AU	DU	DLF	DLV	USX	LPA	HFP	HBM	HBP	Smoking	IDY	VAD	LBD	AP	OAP	DHS	DLG	DLN
DB	1.00	0.20	0.14	0.09	0.50	0.13	0.25	0.33	0.50	1.00	3.33	0.11	1.00	3.33	0.33	0.17	0.17	3.33	3.33	3.33	3.33	0.14	3.33	3.33	0.25	0.20	0.14	0.33	1.00	0.33
TB	5.00	1.00	0.11	0.10	0.50	0.25	1.00	1.00	2.00	2.00	3.33	0.13	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0.20	2.00	2.00	0.25	0.20	0.14	0.33	0.33	0.33
UWS	7.00	9.00	1.00	8.00	8.00	8.00	1.00	1.00	1.00	1.00	1.00	0.17	0.50	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.11	1.00	1.00	0.25	0.14	0.10	0.33	1.00	0.33
USN	11.00	10.00	1.00	1.00	2.00	0.50	1.43	1.43	1.43	1.43	1.43	0.20	0.33	0.33	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.13	0.33	0.33	0.25	0.13	0.10	0.33	0.14	0.33
NHF	2.00	2.00	1.50	0.50	1.00	0.50	1.00	1.00	1.00	1.00	1.00	0.11	0.50	1.00	0.20	0.20	0.20	0.20	1.00	1.00	1.00	0.11	0.33	1.00	0.25	0.14	0.10	0.33	2.00	2.00
HAP	8.00	4.00	3.00	2.00	2.00	1.00	0.33	0.33	0.50	1.00	2.00	0.14	1.00	2.00	0.50	1.00	0.33	0.50	1.00	0.33	0.20	0.11	2.00	2.00	0.25	0.20	0.11	0.33	0.33	2.00
NEB	4.00	1.00	1.00	0.70	1.00	3.00	1.00	1.00	1.00	1.00	1.00	0.20	0.50	1.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.33	2.00
DBF	3.00	1.00	1.00	0.70	1.00	3.00	1.00	1.00	1.00	1.00	1.00	0.14	0.50	1.00	0.50	0.50	2.00	2.00	2.00	2.00	1.00	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.50	1.43
CW	2.00	0.50	1.00	0.70	1.00	2.00	1.00	1.00	1.00	1.00	1.00	0.14	0.50	1.00	0.50	0.50	0.11	0.20	0.20	0.33	1.00	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.50	1.43
CS	1.00	0.50	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.14	0.50	1.00	0.50	0.50	0.14	0.14	0.14	0.14	0.14	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.50	1.43
LBW	0.30	0.30	1.00	0.70	1.00	0.50	1.00	1.00	1.00	1.00	1.00	0.11	0.25	0.33	0.25	0.25	0.11	0.11	0.11	0.11	0.11	0.09	0.33	0.33	0.25	0.11	0.09	0.33	0.33	0.33
SHS	9.00	8.00	6.00	5.00	9.00	7.00	5.00	7.00	7.00	7.00	9.00	1.00	2.00	10.00	3.33	3.33	10.00	10.00	10.00	10.00	10.00	0.50	10.00	10.00	0.25	1.00	0.50	0.33	0.33	10.00
AU	1.00	1.00	2.00	3.00	2.00	1.00	2.00	2.00	2.00	2.00	4.00	0.50	1.00	1.00	0.50	0.50	0.14	0.14	0.14	0.14	0.14	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.50	1.43
DU	0.30	1.00	2.00	3.00	1.00	0.50	1.00	1.00	1.00	1.00	3.00	0.10	1.00	1.00	0.50	0.50	0.14	0.14	0.14	0.14	0.14	0.11	1.00	1.00	0.25	0.14	0.11	0.33	0.50	1.43
DLF	3.00	0.50	1.00	2.00	5.00	2.00	2.00	2.00	2.00	2.00	4.00	0.30	2.00	2.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	0.11	1.00	1.00	1.00	0.11	0.09	0.33	1.00	1.00
DLV	6.00	0.50	1.00	2.00	5.00	1.00	2.00	2.00	2.00	2.00	4.00	0.30	2.00	2.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	0.11	1.00	1.00	1.00	0.11	0.09	0.33	1.00	1.00
USX	6.00	0.50	1.00	2.00	5.00	3.00	1.00	0.50	9.00	7.00	9.00	0.10	7.00	7.00	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.09	0.50	0.50	0.50	0.09	0.08	0.33	1.00	1.00
LPA	0.30	0.50	1.00	2.00	5.00	2.00	1.00	0.50	5.00	7.00	9.00	0.10	7.00	7.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.09	0.50	0.50	0.50	0.09	0.08	0.33	1.00	1.00
HFP	0.30	0.50	1.00	2.00	1.00	1.00	1.00	0.50	5.00	7.00	9.00	0.10	7.00	7.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.09	0.50	0.50	0.50	0.09	0.08	0.33	1.00	1.00
HBM	0.30	0.50	1.00	2.00	1.00	3.00	1.00	0.50	3.00	7.00	9.00	0.10	7.00	7.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.09	0.50	0.50	0.50	0.09	0.08	0.33	1.00	1.00
HBP	0.30	0.50	1.00	2.00	1.00	5.00	1.00	1.00	1.00	7.00	9.00	0.10	7.00	7.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.09	0.50	0.50	0.50	0.09	0.08	0.33	1.00	1.00
Smoking	7.00	5.00	9.00	8.00	9.00	9.00	9.00	9.00	9.00	9.00	11.00	2.00	9.00	9.00	9.00	9.00	11.00	11.00	11.00	11.00	11.00	1.00	20.00	20.00	20.00	1.30	1.00	2.50	2.00	2.00
IDY	0.30	0.50	1.00	3.00	3.00	0.50	1.00	1.00	1.00	1.00	3.00	0.10	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	0.05	1.00	0.50	0.50	0.09	0.08	0.33	1.00	1.00
VAD	0.30	0.50	1.00	3.00	1.00	0.50	1.00	1.00	1.00	1.00	3.00	0.10	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	0.05	2.00	1.00	0.50	0.09	0.08	0.33	1.00	1.00
LBD	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	0.05	2.00	2.00	1.00	0.09	0.08	0.33	1.00	1.00
AP	5.00	5.00	7.00	8.00	7.00	5.00	7.00	7.00	7.00	7.00	9.00	1.00	7.00	7.00	9.00	9.00	11.00	11.00	11.00	11.00	11.00	0.77	11.00	11.00	11.00	1.00	1.00	20.00	100.00	100.00
OAP	7.00	7.00	10.00	10.00	10.00	9.00	9.00	9.00	9.00	9.00	11.00	2.00	9.00	9.00	11.00	11.00	13.00	13.00	13.00	13.00	13.00	1.00	13.00	13.00	13.00	1.00	1.00	20.00	100.00	100.00
DHS	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.40	3.00	3.00	3.00	0.05	0.05	1.00	1.11	1.11
DLG	1.00	3.00	1.00	7.00	0.50	3.00	3.00	2.00	2.00	2.00	3.00	3.00	2.00	2.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00	1.00	0.01	0.01	0.90	1.00	1.00
DLN	3.00	3.00	3.00	3.00	0.50	0.50	0.50	0.70	0.70	0.70	3.00	0.10	0.70	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00	1.00	0.01	0.01	0.90	1.00	1.00
Sum	101.40	74.00	67.75	89.19	92.00	79.88	64.51	63.80	85.13	98.13	136.10	19.60	86.28	100.20	52.62	53.45	71.85	76.27	77.57	77.04	76.57	7.06	83.83	83.00	59.00	7.30	5.82	53.30	223.42	239.92

Table 7. Results of the AHP Analysis.

Variables	Priority Vector	Weights
Disease burden	0.0159	0.0159
Tuberculosis	0.0184	0.0184
Unsafe water source	0.0253	0.0253
Unsafe sanitation	0.0175	0.0175
No access to handwashing facility	0.0107	0.0107
Household air pollution from solid fuels	0.0168	0.0168
Non-exclusive breastfeeding	0.0131	0.0131
Discontinued breastfeeding	0.0144	0.0144
Child wasting	0.0103	0.0103
Child stunting	0.0090	0.0090
Low birth weight due to short gestation	0.0067	0.0067
Secondhand smoke	0.0827	0.0827
Alcohol use	0.0143	0.0143
Drug use	0.0108	0.0108
Diet low in fruits	0.0206	0.0206
Diet low in vegetables	0.0207	0.0207
Unsafe sex	0.0262	0.0262
Low physical activity	0.0220	0.0220
High fasting plasma glucose	0.0202	0.0202
High body-mass index	0.0202	0.0202
High systolic blood pressure	0.0205	0.0205
Iron deficiency	0.1288	0.1288
Smoking	0.0148	0.0148
Vitamin A deficiency	0.0146	0.0146
Low bone mineral density	0.0368	0.0368
Air pollution	0.1420	0.1420
Outdoor air pollution	0.1665	0.1665
Diet high in sodium	0.0384	0.0384
Diet low in whole grains	0.0257	0.0257
Diet low in nuts and seeds	0.0160	0.0160

3.1.4. Overall Results of Fuzzy Inference Modeling

Because the lattice degree of nearness from the Chi-square method has the highest value of 69.36% (Table 8), the Chi-square method is considered the best weight method.

3.2. Results of the RFM

We utilized the free application of Google Colab to perform the RFM. Google Colab uses Python 3.6 and allows researchers to share codes (Figure 3). The tree structure of the RFM is illustrated in Figure 4, and the results are given in Figure 5. In Figure 3, the maximum tree depth was 5, the minimum number of cases in the parent node was 10, and the minimum number of cases in a child node was 5. We obtained 20 nodes, 10 terminal nodes, and 5 levels of depth of the tree. The predicted accuracy of LCM was 96.17%. The tree structure of LCM in Figure 3 broke down the factor of No access to handwashing facility (NHF) by disease burden of 680 observations into two parts of 532 samples and

148 samples. A total of 663 observations were selected as training data and 434 observations were used as testing data. The results of the RFM modeling are shown in Figure 5. The top 7 risk factors associated with LCM are Smoking, Low physical activity (LPA), Child wasting (CW), Tuberculosis (TB), Low birth weight due to short gestation (LBW), Iron deficiency (IDY), and Diet low in nuts and seeds (DLN).

Table 8. Lattice Degree on Nearness.

Method	Original Weights C	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	Fuzzy Inference	Lattice Degree of Nearness
		2.92%	**7.18%**	**15.41%**	**30.66%**	**43.83%**		**σ**
Chi-square	$B1 \otimes C$	2.92%	7.18%	14.48%	23.44%	43.83%	43.83%	
	$B1 \odot C$	5.11%	9.52%	15.41%	30.66%	47.44%	5.11%	
	$C \cdot B1$							69.36%
RIDIT Analysis	$B2 \otimes C$	2.92%	7.18%	15.41%	26.65%	42.35%	42.35%	
	$B2 \odot C$	3.98%	9.97%	17.05%	30.66%	43.83%	3.98%	
	$C \cdot B2$							69.18%
AHP	$B3 \otimes C$	2.92%	7.18%	15.41%	24.12%	40.16%	40.16%	
	$B3 \odot C$	6.74%	11.95%	17.02%	30.66%	43.83%	6.74%	
	$C \cdot B3$							66.71%

```
import pandas as pd
def error_matrix(y, p):
  """
  Populates an error matrix
  y: the ranking of lung cancer
  p: the predictions
  """
  p = p.astype('int64')
  y = y.values.astype('int64')
  labels = np.sort(np.unique(y.tolist()))
  labels = [str(l) for l in labels]
  n = len(labels)
  em = pd.DataFrame(np.zeros((n, n)), dtype='int64'), index=labels, col-
umns=labels)
  for mp, my in zip(p, y):
    em[str(my[0])][str(mp)] += 1
    pass
  return em
Direct_df=pd.read_csv('lungcancerForRD.csv')
X = Direct_df[['DB', 'TB', 'UWS', 'USN', 'NHF', 'HAP',
       'NEB', 'DBF', 'CW', 'CS', 'LBW', 'SHS', 'AU', 'DU', 'DLF', 'DLV','USK',
'LPA', 'HFP', 'HBM', 'HBP', 'Smoking ', 'IDY', 'VAD', 'LBD',
       'AP', 'OAP', 'DHS', 'DLG', 'DLN']]
feature_names=['DB', 'TB', 'UWS', 'USN', 'NHF', 'HAP',
       'NEB', 'DBF', 'CW', 'CS', 'LBW', 'SHS', 'AU', 'DU', 'DLF', 'DLV','USK',
'LPA', 'HFP', 'HBM', 'HBP', 'Smoking ', 'IDY', 'VAD', 'LBD',
       'AP', 'OAP', 'DHS', 'DLG', 'DLN']
y = Direct_df[["level"]][Direct_df['level']>0]
clf = rfc(max_depth=5, random_state=0)
from sklearn.inspection import permutation_importance
clf.fit(X, y)
perm importance = permutation importance(clf, X, y)
prediction = clf.predict(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.35,
random_state=0)
from sklearn.ensemble import RandomForestRegressor
rf1 = RandomForestRegressor(n_estimators=100)
rf1.fit(X train, y train)
rf1.feature_importances_
plt.rcParams.update({'figure.figsize': (12.0, 8.0)})
plt.rcParams.update({'font.size': 10})
plt.barh(feature_names,rf1.feature_importances_)
print(f"The scikit-learn prediction accuracy is {clf.score(X, y) * 100}%")
my_er = error_matrix(y, prediction)
for i in range(5):
  print(clf.estimators_[i])
fig = plt.figure(figsize=(80,70))
annotations = tree.plot_tree(clf.estimators_[0],
                             feature names=['DB', 'TB', 'UWS', 'USN', 'NHF',
'HAP',
       'NEB', 'DBF', 'CW', 'CS', 'LBW', 'SHS', 'AU', 'DU', 'DLF', 'DLV','USK',
'LPA', 'HFP', 'HBM', 'HBP', 'Smoking ', 'IDY', 'VAD', 'LBD','AP', 'OAP',
'DHS', 'DLG', 'DLN'],
                             class_names=["Diseaseburden","Tuberculosis","un-
safewatersource", "safewatersource","safesanitation","Noaccesstohandwash-
ingfacility", "household airpollutionfromfuels","nobreastfeeding", "Discon-
tinuedbreastfeeding", "Childwasting", "Childstunting", "lowbirthweight",
"Secondhandsmoke", "Alcoholuse", "Druguse", "Dietlowinfruits",
"Dietlowinvegetables","Unsafesex","Lowphysicalactivity","Highbody-massin-
dex","Highsystolicbloodpressure","Irondeficiency","Smoking","VitaminAdefi-
ciency","Lowbonemineraldensity","Airpollution","Outdoorairpollution","Di-
ethighinsodium","Dietlowinwholegrains","Dietlowinnutsand-
seeds"],filled=True)
fig.savefig("decision_tree.png")
```

Figure 3. Python codes used in the FRM.

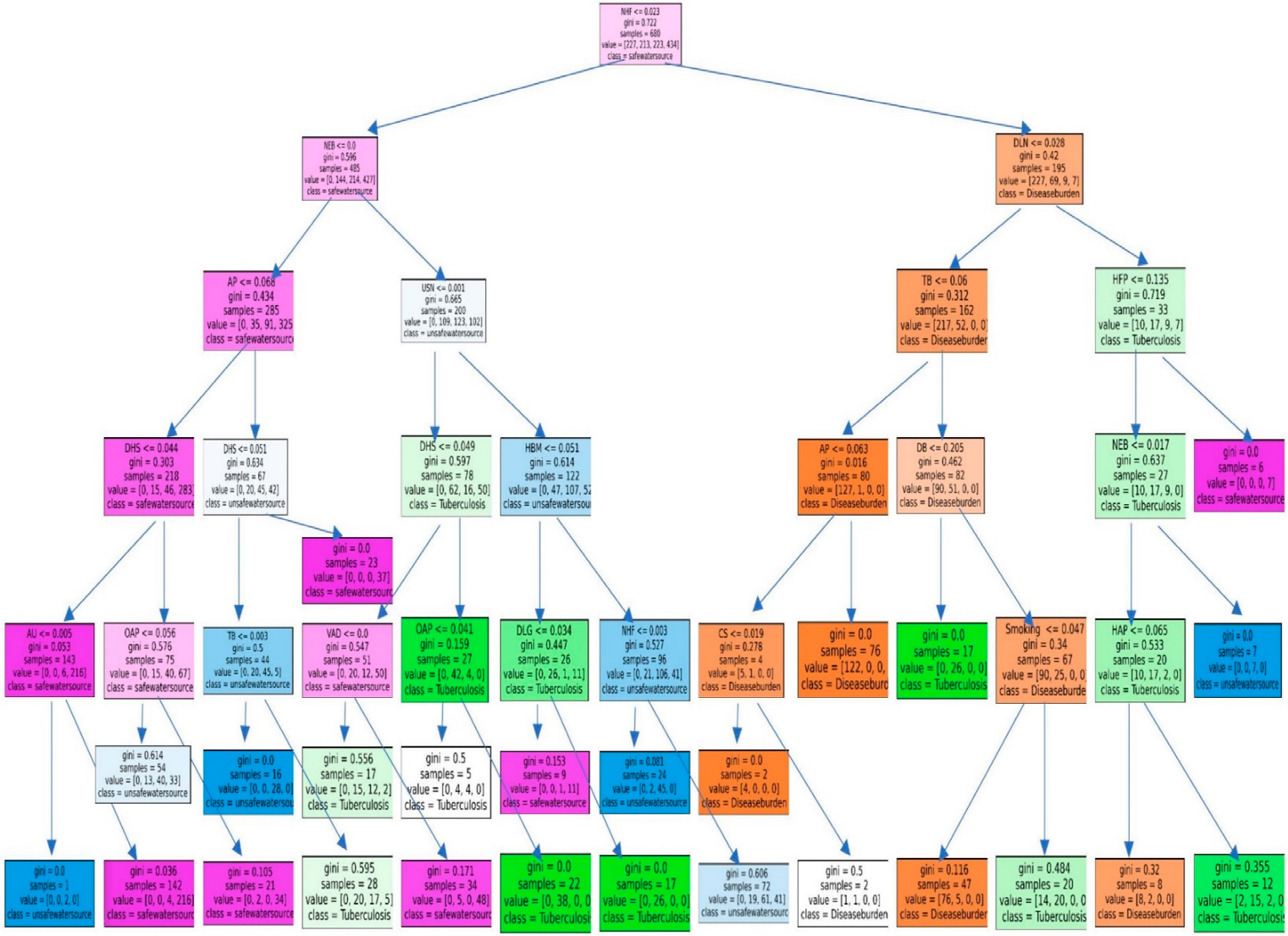


Figure 4. The tree structure of the RFM.

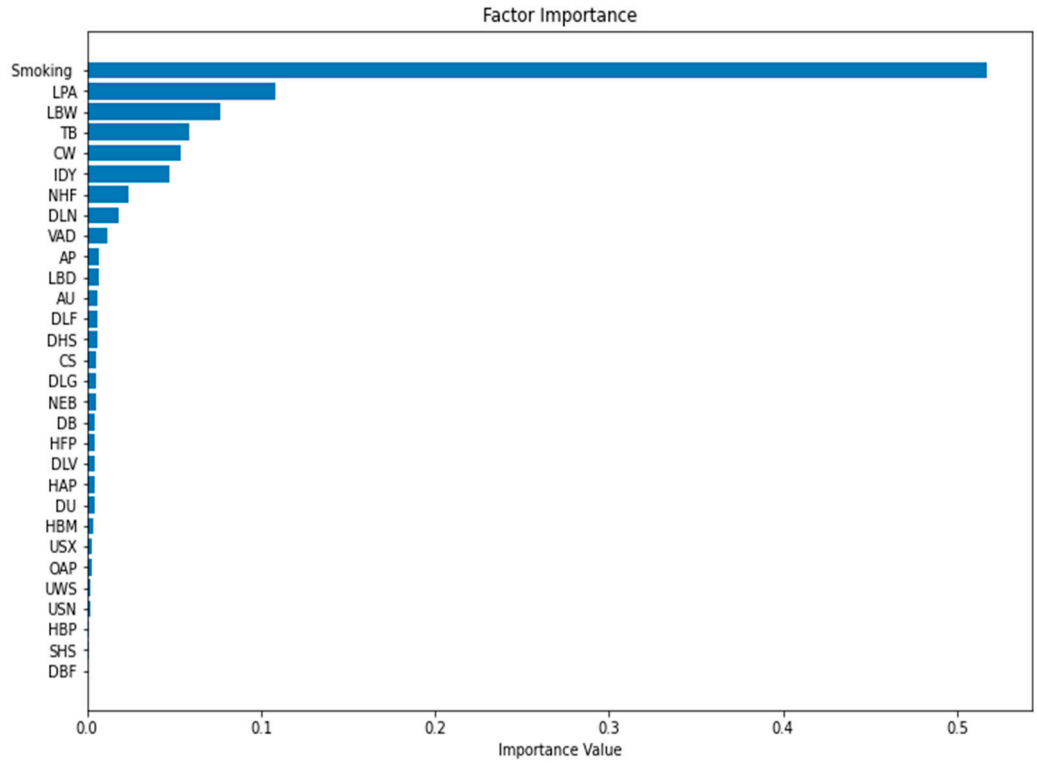


Figure 5. Importance of different risk factors determined by the FRM.

4. Discussion

This research took advantage of empirical analysis to compare the importance of lung cancer's impact factors. Table 9 is a summary of the analysis results for the random forest modeling (RFM) and fuzzy inference modeling (FIM). The top 10 risk factors detected by each method are highlighted in yellow in Table 9. The top 10 risk factors identified by RFM are: smoking, low physical activity, child wasting, low birth weight due to short gestation, iron deficiency, diet low in nuts and seeds, vitamin A deficiency, low bone mineral density, air pollution, and diet high in sodium. Since smoking is a well-known risk factor associated with lung cancer [26–30], the results from RFM appear to be more meaningful compared with the results from FIM. The RFM findings that shed light on the smoking habit are highly superior to environmental factors, which is the first killer of lung cancer, pregnancy, and heart disease [31]. The main reason is related to the immune system's impairment in the recruitment of white cells that release free radicals to kill off the pathogens. These free radicals could provoke an inflammatory overload when combined with those in cigarette smoke, stimulating the activated leukocytes that emit an array of cytokines, resulting in the generation of more inflammatory cells [32]. Meanwhile, the RFM results are based on the machine-learning bagging algorithm and use the ensemble learning technique [33]. It created as many decision trees as possible on the subset of the data and congregated the output of all decision trees. It reduced overfitting problems in decision trees and variances so that it substantially improves the accuracy in the terminated comparison. Importantly, the RFM evidence portrays that the LCM research belongs to machine learning-based medical anomaly detection that aims to predict and diagnose illnesses [34]. The RFM, therefore, is a generally advanced application of emergent disease detection.

Table 9. Comparison of the analysis results from different methods.

Main Factors	RFM Ranking	AHP Weight	AHP Ranking	Chi-Square Weight	Chi-Square Ranking	RIDIT Weight	RIDIT Ranking
Smoking	1	0.015	3	0.042	3	0.002	24
Low physical activity	2	0.022	10	0.037	9	0.013	18
Child Wasting	3	0.010	22	0.047	1	0.032	8
Low birth weight due to short gestation	4	0.007	24	0.042	4	0.023	13
Iron deficiency	5	0.129	18	0.020	22	0.067	3
Diet low in nuts and seeds	6	0.016	16	0.034	13	0.021	16
Vitamin A deficiency	7	0.015	19	0.043	2	0.008	20
Low bone mineral density	8	0.037	6	0.034	15	0.003	23
Air pollution	9	0.142	2	0.027	20	0.040	6
Diet high in sodium	10	0.038	5	0.034	14	0.024	12
Household air pollution from solid fuels	11	0.017	15	0.036	11	0.025	11
Diet low in fruits	12	0.021	12	0.035	12	0.021	15
Disease burden	13	0.016	17	0.003	24	0.245	1
Diet low in vegetables	14	0.021	11	0.037	10	0.014	17
Alcohol use	15	0.014	20	0.033	17	0.033	7
Drug use	16	0.011	21	0.038	6	0.006	22
High body-mass index	17	0.020	13	0.024	21	0.054	4
Child stunting	18	0.009	23	0.039	5	0.007	21
Unsafe water source	19	0.025	9	0.038	7	0.023	14
High fasting plasma glucose	20	0.020	14	0.018	23	0.070	2
Diet low in whole grains	21	0.026	8	0.031	19	0.031	9
Outdoor air pollution	22	0.167	1	0.033	16	0.025	10
Unsafe sex	23	0.026	7	0.032	18	0.048	5
Secondhand smoke	24	0.083	4	0.038	8	0.010	19

The fuzzy inference provided an effective and quantitative weighting method to search the primary impact factors on LCM. On the one hand, its strength is capable of modeling non-linearity efficiently, segregating normal and anomalous samples, and better predicting the inconsistencies [35]. On the other hand, the FIM outcomes strengthened the RFM detections. Albeit the results from FIM are mixed, smoking was identified to be the third most important risk factor by both the AHP analysis and the Chi-square test, in accordance with the RFM outcome. Smoking, in the FIM, was not picked up by RIDIT analysis as an important risk factor, implying that the validity of the RIDIT analysis for the data used in this study is problematic. However, it does not have an impact on the FIM results. This is because the most optimal weight group was computed by the Chi-square test. The air pollution and outdoor air pollution were detected as the top two most important risk factors by AHP analysis, which were verified by Turner and colleagues, who applied and compared the association between outdoor air pollution and lung cancer to account for the global spatial variability of lung cancer [5]. These findings and other risk factors identified by AHP analysis warrant additional research. It is also important to note that four risk factors, smoking, low physical activity, child wasting, and air pollution, are the common risk factors identified by FRM and two other methods from FIM. Additional research is needed to examine the association between these four risk factors and lung cancer mortality.

Most importantly, the difference between FIM and RFM draws focus to the Chi-square results and RFM results. This could be due to the weighing discrepancies of environment, nutrition, diet and sex. Apart from the common factors, there are five different factors that should be noticed in the top 10 factors of both results. Iron deficiency, diet low in nuts and seeds, low bone mineral density, air pollution, diet high in sodium in the RFM results were concentrated in a poorly balanced nutrition, except for air pollution. Malnutrition caused 35% of the incidences of cancer worldwide, estimated by the World Cancer Research Fund

(WCRF) Report 2007 [36]. Air pollution leading to lung cancer is reported by the Lancet October 2022 [37]. The reason is that air pollution stimulated inactive cells with cancer-causing mutations to generate tumors. Simultaneously, the Chi-square findings depicted diet low in vegetables, child stunting, drug use, unsafe water source, and secondhand smoke as the five carcinogenic factors. Diet low in vegetables belongs to diet risk factors. Child stunting is chronic malnutrition, as the same as child wasting. Diet and nutrition, as two of modifiable lifestyle factors, were associated with reduced total cancer-specific mortality, updated by the WCRF and the American Institute for Cancer Research (AICR) (2018) [38–41]. Drug use was positively correlated with sexual behaviors [42], impacting on individual HIV-infection, ultimately resulting in the increased risk of developing lung cancer for the general population [43]. Unsafe water source is present in various food products, including mutagenic and carcinogenic compounds [36]. Secondhand smoke is strongly associated with small cell lung cancer [44], causing a 25% increased risk of lung cancer for non-smokers (American Cancer Society Report). Indeed, the root of the difference between the two results is the biases of the two methods. Due to overfitting, RFM results might have unacceptably high variance and consequently poor predictions on unseen data. FIM results depend on the lattice degree of nearness, which might from subjective judgment by experts.

In addition to smoking, this study suggests that future research should examine risk factors such as low physical activity, child wasting, low birth weight due to short gestation, iron deficiency, diet low in nuts and seeds, vitamin A deficiency, low bone mineral density, air pollution, and diet high in sodium. Despite the fact that this study established two robust models to classify LCM to determine the most sensitive impact factors, some limitation should be noted. First, albeit two models in the application of LCM are novel for pinpointing etiology and pathogenesis on LCM, overfitting in the RFM model might exist so that outcomes are changed. Second, the scale of this research is too big to model spatial-temporal regression. The scale might be narrowed down in future works to make models more robust. Finally, with the advent of the Big Data Era and the development of data mining techniques, deep learning-based medical anomaly detection draws more attention [19]. Some updated Artificial Intelligence (AI) algorithms such as convolution neural networks might improve the model's accuracy in future research [45].

5. Conclusions

This study demonstrates the feasibility of using Fuzzy Inference Modeling (FIM) and Random Forest Modeling (RFM) approaches to identify potential risk factors associated with Lung Cancer Mortality (LCM). The approaches may be useful and effective in exploring the association between a disease and its potential risk factors involving the analysis of large datasets. Future research efforts should expand the research to other diseases and their possible risk factors. In addition, further research is needed examine the effectiveness and difference of FIM and RFM in this type of analysis.

Author Contributions: All authors reviewed the manuscript. Conceptualization and methodology, X.W., J.Z., F.Z. and B.-B.D.; investigation, resources, and data curation, X.W.; writing—original draft preparation, X.W.; writing, reviewing, and editing, J.Z., F.Z. and B.-B.D.; supervision, F.Z. and B.-B.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We extend our thanks to the two anonymous reviewers for their valuable suggestions and comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. White, V.; Ruparelia, P. Chapter 28—Respiratory disease. *Kumar Clark's Clin. Med.* **2021**, *22*, 927–999. [CrossRef]
2. Hernandez, J.B.R.; Kim, P.Y.; NCBI. *Mortality and Morbidity*; National Center for Biotechnology Information; U.S. Laboratory of Medicine: Washington, DC, USA, 2021. Available online: https://www.ncbi.nlm.nih.gov/books/NBK547668/ (accessed on 4 October 2022).
3. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in ^{18}5 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
4. LENNARTSSON, C.; HEIMERSON, I. Elderly people's health: Health in Sweden: The National Public Health Report 2012. Chapter 5. *Scand. J. Public Health Suppl.* **2012**, *9*, 95–120. [CrossRef] [PubMed]
5. Turner, M.C.; Andersen, Z.J.; Baccarelli, A.; Diver, W.R.; Gapstur, S.M.; Pope III, C.A.; Prada, D.; Samet, J.; Thurston, G.; Cohen, A. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA Cancer J. Clin.* **2020**, *70*, 460–479. [CrossRef] [PubMed]
6. Yingxian, D.; Daojun, Z.H.U.; Guowei, C.H.E.; Lunxu, L.I.U.; Kun, Z.H.O.U.; Tao, Z.H.U.; Hongsheng, M.A. Clinical Effect of Day Surgery in Patients with Lung Cancer by Optimize Operating Process. *Chin. J. Lung Cancer* **2020**, *23*, 77–83. [CrossRef]
7. Shang, Y.; Zang, A.; Li, J.; Jia, Y.; Li, X.; Zhang, L.; Huo, R.; Yang, J.; Feng, J.; Ge, K.; et al. MicroRNA-383 is a tumor suppressor and potential prognostic biomarker in human non-small cell lung cancer. *Biomed. Pharmacother.* **2016**, *83*, 1175–1181. [CrossRef]
8. Ningning, D.; Yousheng, M. Advances in Lymph Node Metastasis and the Modes of Lymph Node Dissection in Early Stage Non-small Cell Lung Cancer. *Chin. J. Lung Cancer* **2016**, *19*, 359–363. [CrossRef]
9. Tan, A.S.L.; Potter, J. How the Expansion of the U.S. Preventive Services Task Force Lung Cancer Screening Eligibility May Improve Health Equity Among Diverse Sexual and Gender Minority Populations. *LGBT Health* **2021**, *8*, 503–506. [CrossRef]
10. Al Khayat, M.; Eijsink, J.; Postma, M.; van de Garde, E.; van Hulst, M. Cost-effectiveness of screening smokers and ex-smokers for lung cancer in the Netherlands in different age groups. *Eur. J. Health Econ.* **2022**, *23*, 1221–1227. [CrossRef]
11. Williams, R.M.; Li, T.; Luta, G.; Wang, M.Q.; Adams-Campbell, L.; Meza, R.; Tammemägi, M.C.; Taylor, K.L. Lung cancer screening use and implications of varying eligibility criteria by race and ethnicity: 2019 Behavioral Risk Factor Surveillance System data. *Cancer* **2022**, *128*, 1812–1819. [CrossRef]
12. Shi, J.; Yang, Y.; Xie, H.; Wang, X.; Wu, J.; Long, J.; Courtney, R.; Shu, X.; Zheng, W.; Blot, W.J.; et al. Association of oral microbiota with lung cancer risk in a low-income population in the Southeastern USA. *Cancer Causes Control Int. J. Stud. Cancer Hum. Popul.* **2021**, *32*, 1423–1432. [CrossRef] [PubMed]
13. Wei, X.; Zhu, C.; Ji, M.; Fan, J.; Xie, J.; Huang, Y.; Jiang, X.; Xu, J.; Yin, R.; Du, L.; et al. Diet and Risk of Incident Lung Cancer: A Large Prospective Cohort Study in UK Biobank. *Am. J. Clin. Nutr.* **2021**, *114*, 2043–2051. [CrossRef]
14. Wenwei, S.; Kwong, A.; Han, W.; Chia, Y.; Yao, W. Improved trends of lung cancer mortality-to-incidence ratios in countries with high healthcare expenditure. *Thorac. Cancer* **2021**, *12*, 1656–1661. [CrossRef]
15. Cheng, B.; Wang, C.; Zou, B.; Huang, D.; Yu, J.; Cheng, Y.; Meng, X. A nomogram to predict outcomes of lung cancer patients after pneumonectomy based on 47 indicators. *Cancer Med.* **2020**, *9*, 1430–1440. [CrossRef] [PubMed]
16. Felip, E. El cáncer de pulmón en mujeres. *Arbor Cienc. Pensam. Y Cult.* **2015**, *191*, a235. [CrossRef]
17. Timmermann, C. A History of Lung Cancer: The Recalcitrant Disease. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2014**, *48*, 122–125.
18. Jussawalla, D.; Jain, D. Lung cancer in Greater Bombay: Correlations with religion and smoking habits. *Br. J. Cancer* **1979**, *40*, 437–448. [CrossRef]
19. Du, S.; Wang, X.; Feng, C.C.; Zhang, X. Classifying natural-language spatial relation terms with random forest algorithm. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 542–568. [CrossRef]
20. Kaur, H.; Singh, G.; Minhas, J. A review of machine learning based anomaly detection techniques. *arXiv* **2013**, arXiv:1307.7286. [CrossRef]
21. Fernando, T.; Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Deep Learning for Medical Anomaly Detection—A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–37. [CrossRef]
22. Shi, B.; Chen, N.; Wang, J. A credit rating model of microfinance based on fuzzy cluster analysis and fuzzy pattern recognition: Empirical evidence from Chinese 2157 small private businesses. *J. Intell. Fuzzy Syst.* **2016**, *31*, 3095–3102. [CrossRef]
23. Panchal, J.; Majumdar, B.B.; Ram, V.V.; Basu, D. Analysis of user perception towards a key set of attributes related to Bicycle-Metro integration: A case study of Hyderabad, India. *Transp. Res. Procedia* **2020**, *48*, 3532–3544. [CrossRef]
24. Bross, I.D.J. How to use RIDIT analysis. *Biometrics* **1958**, *14*, 18–38. [CrossRef]
25. Bojian, Z.; Hong, Z. The Application of Fuzzy Mathematics in the Chronic diseases Research. *Dis. Surveill.* **1989**, *4*, 38–41. [CrossRef]
26. Hayashi, I. *Smoking: Health Effects, Psychological Aspects and Cessation*; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2012.
27. Slovic, P. *Smoking: Risk, Perception & Policy*; Sage Publications: New York, NY, USA, 2001.
28. Barnett, J.R.; Moon, G.; Pearce, J.; Thompson, L.; Twigg, L. *Smoking Geographies: Space, Place and Tobacco*; Wiley Blackwell: New York, NY, USA, 2017.
29. Rogers, T.J.; Criner, G.J.; Cornwell, W.D. *Smoking and Lung Inflammation: Basic, Pre-Clinical and Clinical Research Advances*; Springer: Berlin/Heidelberg, Germany, 2013.

30. Putrik, P.; Otavova, M.; Faes, C.; Devleesschauwer, B. Variation in smoking attributable all-cause mortality across municipalities in Belgium, 2018: Application of a Bayesian approach for small area estimations. *BMC Public Health* **2022**, *22*, 1699. [CrossRef] [PubMed]
31. Fox, K.R.; Hardy, R.Y.; Moons, P.; Kovacs, A.H.; Luyckx, K.; Apers, S.; Cook, S.C.; Veldtman, G.; Fernandes, S.M.; White, K.; et al. Smoking among adult congenital heart disease survivors in the United States: Prevalence and relationship with illness perceptions. *J. Behav. Med.* **2021**, *44*, 772–783. [CrossRef]
32. Cope, G.F. *Smoking: What All Healthcare Professionals Need to Know*; M & K Publishing: Stuart, FL, USA, 2016.
33. Spathis, D.; Vlamos, P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Inform. J.* **2019**, *25*, 811–827. [CrossRef]
34. Melnykova, N.; Kulievych, R.; Vycluk, Y.; Melnykova, K.; Melnykov, V. Anomalies Detecting in Medical Metrics Using Machine Learning Tools. *Procedia Comput. Sci.* **2022**, *198*, 718–723. [CrossRef]
35. Pedrycz, W. *Fuzzy Modelling: Paradigms and Practice*; Springer: Manhattan, NY, USA, 1996.
36. Lewandowska, A.M.; Rudzki, M.; Rudzki, S.; Lewandowski, T.; Laskowska, B. Environmental risk factors for cancer-review paper. *Ann. Agric. Environ. Med.* **2019**, *26*, 1–7. [CrossRef]
37. Gourd, E. New evidence that air pollution contributes substantially to lung cancer. *Lancet Oncol.* **2022**, *23*, e448. [CrossRef]
38. Hermans, K.E.P.E.; van den Brandt, P.A.; Loef, C.; Jansen, R.L.H.; Schouten, L.J. Adherence to the World Cancer Research Fund and the American Institute for Cancer Research lifestyle recommendations for cancer prevention and Cancer of Unknown Primary risk. *Clin. Nutr.* **2022**, *41*, 526–535. [CrossRef] [PubMed]
39. Romaguera, D.; Vergnaud, A.-C.; Peeters, P.H.; Van Gils, C.H.; Chan, D.S.; Ferrari, P.; Romieu, I.; Jenab, M.; Slimani, N.; Clavel-Chapelon, F.; et al. Is concordance with World Cancer Research Fund/American Institute for Cancer Research guidelines for cancer prevention related to subsequent risk of cancer? Results from the EPIC study. *Am. J. Clin. Nutr.* **2012**, *96*, 150–163. [CrossRef] [PubMed]
40. Vergnaud, A.-C.; Romaguera, D.; Peelers, P.H.; Van Gils, C.H.; Chan, D.S.; Romieu, I.; Freisling, H.; Ferrari, P.; Clavel-Chapelon, F.; Fagherazzi, G.; et al. Adherence to the World Cancer Research Fund/American Institute for Cancer Research guidelines and risk of death in Europe: Results from the European Prospective Investigation into Nutrition and Cancer cohort study. *Am. J. Clin. Nutr.* **2013**, *97*, 1107–1120. [CrossRef] [PubMed]
41. Shams-White, M.M.; Brockton, N.T.; Mitrou, P.; Romaguera, D.; Brown, S.; Bender, A.; Kahle, L.L.; Reedy, J. Operationalizing the 2018 world cancer research fund/american institute for cancer research (WCRF/AICR) cancer prevention recommendations: A standardized scoring system. *Nutrients* **2019**, *11*, 1572. [CrossRef] [PubMed]
42. Park, J.W.; Dobs, A.S.; Ho, K.S.; Palella, F.J.; Seaberg, E.C.; Weiss, R.E.; Detels, R. Characteristics and Longitudinal Patterns of Erectile Dysfunction Drug Use Among Men Who Have Sex with Men in the U.S. *Arch. Sex. Behav.* **2021**, *50*, 2887–2896. [CrossRef]
43. Sirera, G.; Videla, S.; Saludes, V.; Castellà, E.; Sanz, C.; Ariza, A.; Clotet, B.; Martró, E. Prevalence of HPV-DNA and E6 mRNA in lung cancer of HIV-infected patients. *Sci. Rep.* **2022**, *12*, 13196. [CrossRef]
44. Kim, C.H.; Lee, Y.C.A.; Hung, R.J.; McNallan, S.R.; Cote, M.L.; Lim, W.Y.; Chang, S.-C.; Kim, J.H.; Ugolini, D.; Chen, Y.; et al. Exposure to secondhand tobacco smoke and lung cancer by histological type: A pooled analysis of the International Lung Cancer Consortium (ILCCO). *Int. J. Cancer* **2014**, *135*, 1918–1930. [CrossRef]
45. Khalil, A.A.; E Ibrahim, F.; Abbass, M.Y.; Haggag, N.; Mahrous, Y.; Sedik, A.; Elsherbeeny, Z.; Khalaf, A.A.M.; Rihan, M.; El, S.W.; et al. Efficient anomaly detection from medical signals and images with convolutional neural networks for Internet of medical things (IoMT) systems. *Int. J. Numer. Methods Biomed. Eng.* **2022**, *38*, e3530. [CrossRef]