Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Using Kohonen Artificial Neural Network to Cluster and Visualize Risk Factors for Lung Cancer

Emad Alsyed

Department of Nuclear Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia
E-mail: alsyed@kau.edu.sa

Received: 8 October 2024; **Revised:** 31 December 2024; **Accepted:** 10 February 2025

Abstract: Lung cancer is a complex disease with multiple risk factors, including smoking, exposure to second-hand smoke, alcohol consumption, air pollution, occupational exposures, genetic predisposition, age, and previous lung disease. Artificial Intelligence (AI) has emerged as a promising tool for clustering risk factors related to lung cancer, improving diagnostic efficiency, providing optimal treatment, and assessing prognosis. However, only a few studies have used clustering algorithms to identify modifiable risk factors for lung cancer. In this study, we have implemented the Kohonen Artificial Neural Network-also known as the Self-Organizing Map (SOM)-to cluster and analyze a lung cancer database of 1,000 records and six attributes (risk factors). The SOM methodology is a powerful tool for clustering and scaling up data, and it has proven to be a versatile and effective technique for solving complex problems and simplifying data. Our study demonstrates the potential of SOM to revolutionize how we understand and manage lung cancer risk factors, which is necessary to improve the prevention, early detection, and treatment of this fatal disease.

***Keywords*:** Kohonen Artificial Neural Network, Self-Organizing Map (SOM), Artificial Intelligence (AI), risk factors, lung cancer, data analysis

MSC: 68T07, 62H30

1. Introduction

Lung cancer is a complex disease with multiple risk factors contributing to its development. Among them, smoking is the most well-established risk factor responsible for most cases globally. However, a significant proportion of lung cancer patients comprises non-smokers, signifying the need for further exploration into other risk factors [1, 2]. These factors include exposure to alcohol consumption, air pollution, occupational hazards, genetic predisposition, age, previous non-carcinogenic lung disease, second-hand smoke, tobacco intake, etc. [1, 3]. Furthermore, socioeconomic status has been demonstrated to be inversely related to lung cancer incidence and mortality [4, 5]. A comprehensive understanding and evaluation of various risk factors for lung cancer is essential to ensure prevention, early detection, and treatment of this lethal disease.

In recent decades, the rapid growth of Artificial Intelligence (AI) and Machine Learning (ML) models has transformed the capability to detect intricate patterns in lung cancer data, which provides a level of quantitative analysis

DOI: https://doi.org/10.37256/cm.6520255854

that was impossible in the post-AI era [6, 7]. In addition, AI-based models have emerged as promising tools for classifying various risk factors related to lung cancer. As a result, AI can help improve the diagnostic efficiency of lung cancer, offer optimal treatment, and evaluate prognosis, thereby minimizing mortality [8–10]. Therefore, AI has immense potential to revolutionize how we comprehend and manage lung cancer risk factors. To date, only a limited number of studies have employed clustering algorithms, such as k-means and hierarchical clustering, to detect modifiable risk factors for lung cancer [11–13]. K-means clustering was employed to pre-process lung cancer patient data before applying frequent pattern discovery and decision tree algorithms to develop a lung cancer risk prediction system [14].

To the best of the authors' knowledge, using Kohonen Artificial Neural Network to produce a low-dimensional representation of the lung cancer risk factors data has been minimally addressed in the existing literature. Thus, in this study, we have implemented the Kohonen Artificial Neural Network-also known as the Self-Organizing Map (SOM) to cluster and analyze a lung cancer database of 1,000 records and six attributes (risk factors). SOM is a type of Artificial Neural Network (ANN) that is categorized as unsupervised learning [15, 16]. SOMs are utilized in several application areas for problem-solving, data visualization, and simplification. Ritter and Kohonen pioneered the SOM algorithm, which organizes features into spatially organized representations [17]. In recent years, it has become a powerful tool for clustering and upscaling data. The SOM has been particularly effective in clustering patients based on risk factors for Chronic Obstructive Pulmonary Disease (COPD) [18]. In addition, SOM was effective in classifying lung cancer stages using computed tomography images [19]. SOM has demonstrated high accuracy in identifying lung cancer stages [19]. Moreover, it is an advanced methodology for solving complex problems and simplifying data. In this research study, we have filled the literature gap by utilizing the SOM technique in classifying lung cancer data. More specifically, we have implemented the SOM algorithm to cluster and evaluate a lung cancer database comprising 1,000 records and six attributes, including age, air pollution, alcohol consumption, smoking, passive smoking, and obesity. Thus, we aim to uncover hidden patterns and identify distinct patient subgroups to inform more effective lung cancer prevention and treatment strategies.

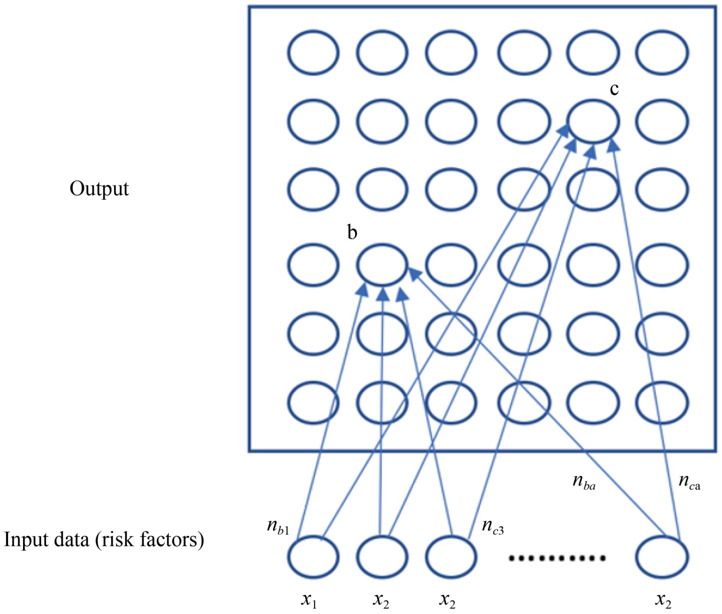


Figure 1. An illustrative diagram for SOM

As illustrated in Figure 1, the Kohonen SOM is a set of input data, such as A risk factors, onto a two-dimensional grid of neurons. Each neuron in the grid is given an initial weight vector $N = (n_{b1},\ n_{b2},\ ..,\ n_{bA})$ with the same dimensionality as the input data $b = (1,\ 2,\ \ ...\ A)$. $\sqrt{g = (1,\ 2,\ \ldots L)}$. The neuron matrix learned during training is utilized to map input data values onto a standardized two-dimensional grid. This image has been adopted from in Figure 1 [20]. During the model training process, a competitive learning methodology is deployed, where nodes compete against each other to respond to input data. During the training phase, the network is inputted with random training data, and the Euclidean distance to all weight vectors is computed using Equation (1) to update the neuron weights. Eventually, the neuron with a vector similar to the input is identified as the Best Matching Unit (BMU). Euclidean distance helps the SOM find the most similar neuron (BMU) to an input, and this drives the learning process, leading to a map that effectively organizes and represents the input data. Euclidean distance is simply a way to measure how similar two things are. In the case of a SOM, it tells us how close an input data point is to the "weights" of a neuron in the map. Smaller distance = more similar.

$$m_{t+1}^{ba} = m_t^{ga} + \eta h(b,\, c)\left(x_a - m_{ja}^{t}\right),\ \text{for } 1 \le a \le A \tag{1}$$

The neighborhood function $h(b,\, c)$ has a value of 1 at the winning neuron q, which decreases as the distance between g and q increases. The neighborhood function $h(b,\, c)$ plays a crucial role in shaping the topology of the map. It determines the extent to which neighboring neurons are also updated, ensuring that neurons with similar preferences are grouped together. As the training progresses, the neighborhood function becomes more localized, leading to the formation of distinct clusters. The learning rate parameter (η) controls the weight vector size. Consequently, each high-dimensional data point is embedded onto a single neuron, reproducing its structure. The neurons' weights serve as indicators of the input space and produce a distinct estimation of the distribution of the training samples. Through this process, the Kohonen SOM method creates more neurons pointing to areas with high concentrations of the training sample and fewer neurons pointing to regions with scarce samples. As described by Kohonen in 2001, it allows the SOM to compress information while maintaining the essential topological and/or metric correlations of the primary data elements on display. After the training phase, the neurons of the Kohonen SOM specify a low-dimensional representation of high-dimensional input data without altering the data distribution shape and the correlation between each input data element. Each unit in the grid has a distinct "codebook" vector representing the typical objects for that specific region in the map.

2. Methods

2.1 *Dataset*

The dataset for this study was retrospectively collected from a well-known lung cancer database-data.world [21], originally compiled by Ahmad et al. [9] through a combination of quantitative and qualitative approaches. In their study, they surveyed physicians across multiple specialties to identify key risk factors and symptoms associated with lung cancer. We extracted a dataset comprising records of 1,000 lung cancer patients, each with six risk factors or attributes. These attributes include:

Age of the patient: The risk of developing various types of cancerincluding lung cancer-increases with growing age.

Air pollution (on a scale of 1-8): It is among the most significant long-term factors contributing to lung cancer development.

Level of alcohol consumption (on a scale of 1-8): The probability of developing lung cancer is higher in alcoholic individuals.

Obesity level (on a scale of 1-8): Obese individuals are at a greater risk of developing cancer than those with normal weight.

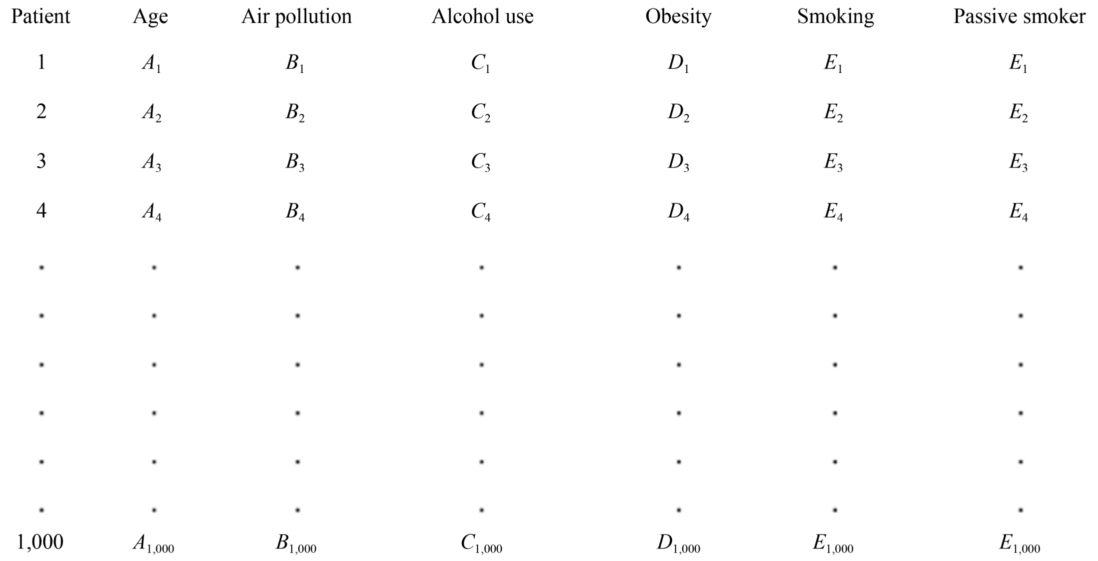
Level of smoking (on a scale of 1-8): The likelihood of developing lung cancer increases substantially with increased smoking or exposure to second-hand smoke.

Level of passive smoking (on a scale of 1-8): Non-smoking adults exposed to second-hand smoke have an elevated risk of developing lung cancer.

2.2 *Clustering and visualization*

We performed the following steps during this research study:

a) We stored the lung cancer patients' data in a table in which each row specifies a single patient, and each column represents the level of different risk factors for that patient. Therefore, each table cell contains the value for a specific risk factor in a particular patient.



Patient	Age	Air pollution	Alcohol use	Obesity	Smoking	Passive smoker
1	A_1	B_1	C_1	D_1	E_1	E_1
2	A_2	B_2	C_2	D_2	E_2	E_2
3	A_3	B_3	C_3	D_3	E_3	E_3
4	A_4	B_4	C_4	D_4	E_4	E_4
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1,000	$A_{1,000}$	$B_{1,000}$	$C_{1,000}$	$D_{1,000}$	$E_{1,000}$	$E_{1,000}$

Figure 2. An illustrative example of a table representing the patient data sorted to perform SOM

The example in Figure 2 includes 1,000 lung cancer patients in rows and six different risk factors in columns (A = Age, B = Air pollution, C = Alcohol consumption, D = Smoking, E = passive smoking, F = Obesity).

b) We used Equation (2) to standardize each risk factor using z-score normalization across the risk factors distribution acquired for their inter-comparison.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Where; x specifies the risk factor value, μ is the mean, and σ represents the standard deviation.

c) A SOM model was trained using unsupervised learning to produce a lower dimensional representation of the input data, i.e., risk factors.

d) Finally, we employed SOM to create a code plot and visualize the data for lung cancer patients. It displays each cluster and the node weight vectors or 'codes' associated with each node.

These codes comprised the original normalized values used to produce the map in SOM and specify the clustering patterns. The fan chart at the clusters' center reveals the characteristics that define how the risk factors were grouped into each cluster. For instance, if a cluster had a large fan piece for age, it implies that most of the patients in this cluster were grouped based on similar normalized risk factor values in the age region. Heatmaps can be generated to display counts or distribution plots illustrating the number of data samples meeting a particular criterion for a specific codebook, which is visualized by pie charts depicting the representative vectors for the grid.

3. Results and discussion

Figure 3 depicts the SOM-generated codes plot, with each color representing one of the tested risk factors. For instance, the yellow color specifies alcohol consumption, while the orange color represents obesity. The color coding related to each risk factor is given in Figure 3. All thirty-six (36) nodes contain information about the 1,000 patients and six attributes (tested risk factors). Figure 3 is pie charts depicting the representative grid vectors. These code plots visually represent the clusters, which are the node weight vectors associated with each node.

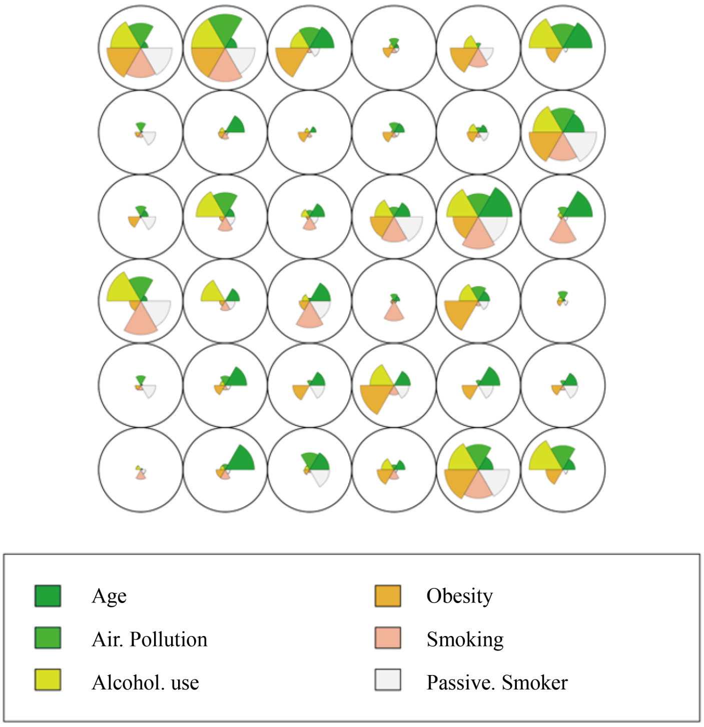


Figure 3. The risk factors are represented by code plots

Patients in the 1st node (lower left) have relatively low normalized values for most risk factors, while patients in the 31st node (upper left) have high normalized values for five risk factors, including air pollution, alcohol use, obesity, smoking, and passive smoking. Furthermore, patients in the same 31st node have relatively lower age values in Table 1.

Table 1. Standardized values (*z*-scores) of risk factors for six nodes (N1, N12, N17, N26, N30 and N31)

Node #	Age	Air pollution	Alcohol consumption	Smoking	Passive smoking	Obesity
N1	-1.68	-1.39	-0.97	-1.63	-0.37	-0.94
N12	-0.18	-0.90	-1.35	-0.68	-0.78	-0.08
N17	-0.52	0.07	0.16	1.19	-0.78	-0.51
N26	0.41	-1.23	-0.92	-1.049	-0.61	-1.14
N30	0.67	1.06	1.0	1.19	1.17	1.56
N31	-1.02	1.31	1.03	1.19	1.27	1.33

The Kohonen ANN helps visualize the count of how many patients are mapped to each node on the map. Figure 4 illustrates that most nodes contain about twenty (20) patients or less. However, around 100 patients were classified to be on the 5th node.

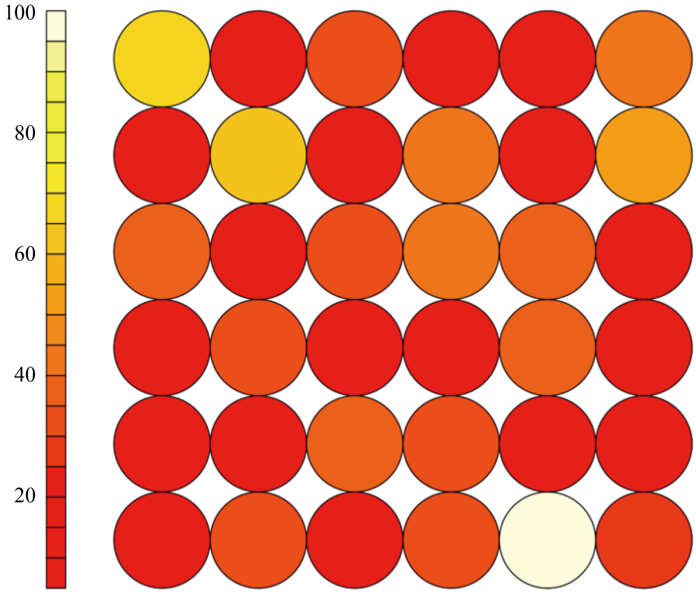


Figure 4. The distribution of the number of patients regarding the self-organized risk factors. This heatmap illustrates the distribution or patient frequency over the nodes of the SOM

From the findings, it can be observed that Nodes 26 and 31 contained the second-highest number of patients, as approximately eighty (80) patients were classified to be on each of these nodes.

Nodes 30 comprised approximately sixty (60) patients. This node represents high values of all of the six tested risk values. Table 1 presents the standardized risk factor values for six nodes (N1, N12, N17, N26, N30, and N31). For example, in the first node, fan sizes are based on the numbers presented in the first row.

Lung cancer is a complicated disease with several risk factors which cause its development. The most pronounced risk factor for lung cancer is smoking-which is responsible for most lung cancer incidences across the world. However, many non-smokers are also impacted by lung cancer, emphasizing the dire need for advanced research into various risk factors, such as exposure to second-hand smoke, air pollution, low consumption of fruits and vegetables, occupational exposures, genetic predisposition, age, and past lung disease.

Although many studies have explored various risk factors, there is very little scientific assessment of the impact of a combination of multiple risk factors [22]. For instance, while cigarette smoking is the primary risk factor for lung cancer, all smokers do not develop lung cancer, suggesting that many other risk factors may play a vital role in lung carcinogenesis. Therefore, this scenario underlines the need to evaluate many lung cancer patients with various risk factors. A comprehensive understanding of these risk factors for lung cancer is essential to promote prevention strategies, early detection, and treatment of this fatal disease.

The primary objective of this research was to utilize clustering methodologies to visualize various risk factors to reduce the high-dimensional lung cancer patient dataset. Therefore, we analyzed a dataset comprising 1,000 lung cancer patients and six risk factors (i.e., Age, Air pollution, Alcohol consumption, Smoking, passive smoking, Obesity) using a Kohonen ANN technique called SOM. Using SOM is beneficial in this context, as it allows for the data partitioning and offers a simple two-dimensional visualization of expression patterns. Unlike traditional clustering methods that simply group similar data points together, SOMs create a 2D map where neighboring neurons represent similar data points. This topological property allows for a visual representation of complex patterns and relationships between different risk factors, making it easier to identify clusters and outliers.

The most significant finding from this research study is that Node # 5 contains the highest number of lung cancer patients. The reason could be the relatively high levels of the tested risk factors, including air pollution, alcohol consumption, smoking, passive smoking, and obesity on the 5th node, barring age. This outcome implies that a high level of these risk factors may increase the probability of lung cancer incidence, even at a relatively young age. Future research studies must more closely examine the association between these identified five factors and lung cancer incidence.

The results show that two other nodes (26 and 31) contain the second-highest number of patients. The concentration of patients in Nodes 26 and 31 may be linked to the elevated levels of obesity and air pollution observed in these areas. Thus, these two factors could have higher impact which potentially contributing to the higher lung cancer patient numbers.

Node 31 has relatively higher levels of air pollution, alcohol consumption, smoking, passive smoking, and obesity but a very low age. This outcome provides additional evidence to the aforementioned point that high exposure to risk factors increases the probability of lung cancer incidence, even in youngsters. In contrast, Node # 26 has a very high age but very low levels of the other five factors. As a result, it is possible to hypothesize that lung cancer incidences are more likely to occur in elderly people with some exposure to air pollution, alcohol consumption, smoking, passive smoking, and obesity. Therefore, future studies focusing on elderly individuals with diversified risk factors are suggested.

This research study also has some limitations. First, it does not consider many other risk factors like family history of lung cancer, past radiation therapy, and genetic predisposition. However, future research studies involving these risk factors are imperative to offer a better understanding of the most at-risk individuals. Second, our study is based on a dataset of only 1,000 patients. Therefore, caution must be applied while interpreting the findings, as they cannot be generalized to the clinic. Nonetheless, involving a larger sample size will enhance our understanding of the riskiest factors. Therefore, in future studies, a larger sample should be considered.

This study did not explore the potential of SOM to visualize gene expression patterns. Thus, future research could investigate the application of SOM to identify patient subgroups, understand the molecular mechanisms of CRISPR-Cas9, and optimize treatment strategies [23]. In addition, the method used in this research (SOM) could be integrated with multi-objective evolutionary algorithms, as proposed by Abdulrahman et al. [24]. Therefore, we can enhance the identification of complex patterns and relationships within biological networks, such as gene regulatory networks or protein-protein interaction networks. This combination could lead to more accurate and efficient community detection, providing valuable insights into the underlying mechanisms of these networks.

4. Conclusion

In this study, we have introduced a novel approach by employing a Self-Organizing Map (SOM) to assess risk factors associated with lung cancer. The main objective of this work was to utilize clustering and visualization techniques to evaluate data for 1,000 lung cancer patients and their related risk factors. Furthermore, the study has demonstrated the SOM's ability to detect complex patterns of risk factors in lung patients' data, including Age, Air pollution, Alcohol consumption, Smoking, passive smoking, and Obesity. Furthermore, we conclude that by expanding this study to incorporate some additional risk parameters like genetic risks, dust allergy, coughing blood, chronic inflammation, fatigue, balanced diet, weight loss, and asthma, we can offer a more comprehensive approach for thoroughly assessing the possibility of developing lung cancer. Moreover, we have also proposed that the SOM could be utilized with outcome data to predict the incidence of lung cancer, with the learned representations of self-organized risk factors serving as prediction attributes. We suggest an integrated approach like this can offer a promising area for future research studies. This study's limitations include the exclusion of certain risk factors and a small sample size. Future research should address these by considering more factors and a larger sample to improve accuracy and generalizability. Moreover, the future direction of this research involves expanding the feature set to include additional risk factors, experimenting with different SOM architectures, and combining SOM with advanced machine learning techniques. Validating the model on larger and more diverse datasets, developing a user-friendly interface, and exploring the impact of early intervention are also key areas of focus.

Conflict of interests

The author declares no conflicts of interest.

References

[1] Corrales L, Rosell R, Cardona AF, Martin C, Zatarain-Barrón ZL, Arrieta O. Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Critical Reviews in Oncology/Hematology*. 2020; 148: 102895.

[2] Smolle E, Pichler M. Non-smoking-associated lung cancer: a distinct entity in terms of tumor biology, patient characteristics and impact of hereditary cancer predisposition. *Cancers*. 2019; 11: 204.

[3] Wang N, Mengersen K, Tong S, Kimlin M, Zhou M, Hu W. Global, regional, and national burden of lung cancer and its attributable risk factors, 1990 to 2017. *Cancer*. 2020; 126: 4220-4234.

[4] Ebner PJ, Ding L, Kim AW, Atay SM, Yao MJ, Toubat O, et al. The effect of socioeconomic status on treatment and mortality in non-small cell lung cancer patients. *Annals of Thoracic Surgery*. 2020; 109: 225-232.

[5] Castro S, Sosa E, Lozano V, Akhtar A, Love K, Duffels J, et al. The impact of income and education on lung cancer screening utilization, eligibility, and outcomes: a narrative review of socioeconomic disparities in lung cancer screening. *Journal of Thoracic Disease*. 2021; 13: 3745.

[6] Svoboda E. Artificial intelligence is improving the detection of lung cancer. *Nature*. 2020; 587: S20.

[7] Al-Mistarehi A-H, Mijwil MM, Filali Y, Bounabi M, Ali G, Abotaleb M. Artificial intelligence solutions for health 4.0: overcoming challenges and surveying applications. *Journal of Healthcare Engineering*. 2023; 2023(1): 15-20.

[8] Zhou Y, Xu X, Song L, Wang C, Guo J, Yi Z, et al. The application of artificial intelligence and radiomics in lung cancer. *Precision Clinical Medicine*. 2020; 3: 214-227.

[9] Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. *Heliyon*. 2020; 6: e03402.

[10] Agarwal S, Yadav AS, Dinesh V, Vatsav KSS, Prakash KSS, Jaiswal S. By artificial intelligence algorithms and machine learning models to diagnosis cancer. *Materials Today: Proceedings*. 2023; 80: 2969-2975.

[11] Ramachandran P, Girija N, Bhuvaneswari T. Early detection and prevention of cancer using data mining techniques. *International Journal of Computer Applications*. 2014; 97: 48-53.

[12] Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. *EBioMedicine*. 2018; 31: 36-46.

[13] Chen S, Wu S. Identifying lung cancer risk factors in the elderly using deep neural networks: quantitative analysis of web-based survey data. *Journal of Medical Internet Research*. 2020; 22: e17695.

[14] Ahmed K, Kawsar AA, Kawsar E, Emran AA, Jesmin T, Mukti RF, et al. Early detection of lung cancer risk using data mining. *International Journal of Advanced Computer Science and Applications*. 2013; 14(1): 595-598.

[15] Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*. 1982; 43: 59-69.

[16] Kohonen T. An overview of SOM literature. *Self-Organizing Maps*. 2001; 30: 347-371.

[17] Ritter H, Kohonen T. Self-organizing semantic maps. *Biological Cybernetics*. 1989; 61: 241-254.

[18] Kramer AA, Lee D, Axelrod RC. Use of a Kohonen neural network to characterize respiratory patients for medical intervention. In: Malmgren H, Borga M, Niklasson L. (eds.) *Artificial Neural Networks in Medicine and Biology*. Berlin, Germany: Springer; 2000. p.192-196.

[19] Apsari R, Aditya YN, Purwanti E, Arof H. Development of lung cancer classification system for computed tomography images using artificial neural network. *AIP Conference Proceedings*. 2021; 2329(1): 050013.

[20] Köküer M, Naguib RNG, Janc P, Younghusband HB, Green R. Towards automatic risk analysis for hereditary non-polyposis colorectal cancer based on pedigree data. *Methods of Information in Medicine*. 2007; 319-337. Available from: https://doi.org/10.1016/B978-044452855-1/50014-3.

[21] Data.world. *Lung cancer database*. 2023. Available from: https://data.world/cancerdatahp/lung-cancer-data [Accessed 20th March 2006].

[22] Kale MS, Wisnivesky J, Taioli E, Liu B. The landscape of US lung cancer screening services. *Chest*. 2019; 155: 900-907.

[23] Dölarslan M. CRISPR-Cas9 mediated gene correction of CFTR mutations in cystic fibrosis: evaluating efficacy, safety, and long-term outcomes in patient-derived lung organoids. *SHIFAA Journal of Medical Sciences*. 2023; 2023: 41-47.

[24] Abdulrahman M, Niu Y. Multi-objective evolutionary algorithm with decomposition for enhanced community detection in signed networks. *Khwarizmia Engineering Journal*. 2023; 2023: 1-17.