A Decision Support System for Improving Lung Cancer Prediction Based on ANN

Yen Nhu Thi Phan[1], Lam Son Quoc Pham[1], Sinh Van Nguyen[1](✉),
and Marcin Maleszka[2]

[1] School of Computer Science and Engineering, International University,
Vietnam National University HCM City, Ho Chi Minh City, Vietnam
nvsinh@hcmiu.edu.vn

[2] Department of Applied Informatics, Wroclaw University of Science and Technology,
Wrocław, Poland

https://it.hcmiu.edu.vn, https://kis.pwr.edu.pl/en/

Abstract. Recent advancements in artificial intelligence (AI) and big data analysis have shown great potential for improving the diagnosis of lung cancer. Early detection of lung cancer is crucial for increasing patient survival rates. This paper analyze the data BRFSS (Behavioral Risk Factor Surveillance System), conducted from 2017 to 2020 to identify risk factors and symptoms of lung cancer. We develop a decision support system (DSS) based on data mining technique to assist healthcare practitioners and users in early diagnosis of lung cancer. Thirteen risk factors and demographic data are selected as predictors. The ANN and a logistic regression (LR) model are performed to predict the probability of lung cancer and to serve as a prognostic index respectively. The ANN model shown an accuracy of 84.79%, a sensitivity of 79.8%, and a specificity of 89.76%, a 93% of the ROC (AUROC) curve. While the LR model obtained an accuracy of 80.2%, a sensitivity of 80%, and a specificity of 72.2%, with a 76.1% AUROC. The models are trained with a batch size of 100, using stochastic gradient descent (SGD) optimizer. By using data analysis and mining techniques, we discovered new patterns in the health behavioral risk data that are previously unknown. Overall, our proposed method has a potential to significantly improve the early detection and treatment of lung cancer.

Keywords: Lung cancer · ANN Model · Data analysis · Decision support system

1 Introduction

Lung cancer is one of the most deadly diseases in the world, affecting people of all ages and becoming more prevalent in every year. The three stages of analyzing the BRFSS data [1] includes training an LR and ANN models, and creating a web application for the final diagnosis of patients. The input data of the system requires 13 risk factors and demographic attributes from the user and produces

N. T. Nguyen et al. (Eds.): ACIIDS 2023, LNAI 13995, pp. 346–357, 2023.
https://doi.org/10.1007/978-981-99-5834-4_28

a diagnosis within a second. According to cancer statistics 2018 in [2], there is an estimation of 234,030 new cases of lung cancer in the United States and 154,050 cancer-related deaths. In India, there are approximately 70,000 cases per year, with a number is go on to rise. Early diagnosis of lung cancer is critical in improving a patient's chances of survival and recovery, highlight the need for a Machine Learning (ML) based support system.

Computer aided diagnosis (CAD) systems have proven to be useful in supporting medical professionals in the detection and diagnosis of diseases [3]. These systems serve as a second opinion to the physician, offering additional support in the diagnostic process [4]. The AI-based expert systems convert knowledge of an expert in a specific field into software that have been used for many years in the medical field. These systems can answer questions and have ability to accommodate new knowledge, either through traditional rule-based techniques or through more advanced approaches like ANN and neuro-fuzzy systems. Despite of the numerous studies that have been conducted on the treatment and diagnosis of lung cancer, our research focus on an early prediction of the disease through an analysis of risk factors. A decision aid system will be built using a multi-parameter neural network that is fed with the most important risk factors and trained to classify lung cancer. The development of AI-based applications in healthcare has become necessary due to the limited availability of medical expertise globally. In many countries, the problem in recent years is lack of medical professionals, who can serve a large number of citizens. For this reason, a computer-based intelligent system is very important and necessary to support. The expert systems can help address human inconsistencies, lack of qualified experts, and a large amount of easily accessible data to provide better decision-making support for patients [5]. One of the biggest challenges in the diagnosis of lung cancer is that the symptoms in the early stage is often asymptomatic, make it difficult to detect [6]. This is why early detection is critical, as the survival rate of lung cancer is the second lowest of all cancers, with only 18%. The goal of this study is to improve the early diagnosis of lung cancer and provide a more efficient and accessible tool for patients and medical professionals alike.

The remaining of the paper is structured as follows: Sect. 2 presents the related works. Our proposed method is presented in Sect. 3. Section 4 shows the implementation and evaluation of the proposed method. The last section (Sect. 5) presents conclusion and future work.

2 Related Works

With the development of modern technologies, the field of medical diagnosis has been significantly impacted by AI and ML. While the image processing techniques combined with the ML methods are widely studied and applied to recognize objects (e.g. face recongnition [7]) and medical diagnosis on medical image datasets [8,9], this section discuss several related works which focus on lung cancer diagnosis and DSS. The most noteworthy work in the area of lung cancer diagnosis was conducted by Ardila et al. (2019) which demonstrated the potential to improve the consistency of lung cancer diagnoses using deep-learning models [10]. The method is developed based on deep-learning model for lung cancer screening on the CT scanner image. The method achieved higher accuracy to the state-of-the-art methods when compared to radiologists. It showed an absolute reduction of 11% in false positives and 5% in false negatives. The method proposed by Linh et al. [8] to determine the tumor region on the brain MRI images based on 3D generative adversarial network (3D-GAN). The method proved the performance and accuracy are better than the existing methods. Sharmila et al. [9] presented an effective approach for predicting and classifying lung cancer using machine learning and image processing techniques. The method proved superior accuracy for lung cancer prediction. However, further details on classifier selection and future research directions would be improved. Sinh et al. [11,12] suggested a method for building 3D models of MRI image objects based on geometric modeling. This methods not only visualize the 3D medical image objects but also support for doctors and medical staffs in diagnostic and treatment. Singla et al. (2013) [13] built a DSS for lung disease diagnosis using a rule-based inference engine to diagnose various lung diseases. The system components includes a knowledge base, a fact base, an inference engine, an explanation module, and an interface for both users and developers. Although the system obtained promising results in diagnosing certain lung diseases, it also highlighted some limitations of using rule-based expert systems, such as the lack of human inconsistencies and the limited knowledge base. Even so, the study set the stage for future research in this field and offered a fundamental perspective on how expert systems can support decision-making. Because of the capacity to handle uncertainty and imprecise information of the existing methods, fuzzy logic has been applied in medical diagnosis recently. Rodiah et al. (2020) presented a web-based fuzzy logic inference engine that is implemented to support lung cancer diagnosis [14]. The system used a knowledge base contained rules for assisting decision-making and the output was determined by a defuzzification step. The method has addressed to overcome the drawbacks of previous works. The study showed promising results in classifying patients based on their age, anamnesis, and heavy smoking. A research for lung cancer prediction using data mining techniques suggested by E. Yatish et al. (2019) [15]. The various data mining techniques like decision trees, k-nearest neighbor, logistic regression, random forest, and SVM are used to predict lung cancer tumor. The study found that KNN and logistic regression had a higher rate of prediction accuracy compared to other techniques. In recent researches, ANN is widely used to analyze patient data obtained from a web survey related to lung cancer symptoms such as age, gender, and six risk factors. The "Age" attribute is found as the greatest impact on the results. However, the imitation of a small dataset and incompleted consideration of all the risk factors and symptoms could impact the generalizability of the findings (2019) [22]. Adrian Cassidy et al. (2007) [16] built a tool for early detection and treatment of lung cancer. The valuable review is to compare the proposed risk models and investigate the works of others. The paper shown the importance of system validation for accurate risk prediction models and identify risk factors beyond age and smoking. G. Chada [17] proposed a method for lung cancer risk prediction based on a multi-parameterized ANN. The study used the datasets from 1997–2015 of National Health Interview Survey to create an ANN based on personal health clinical and demographic information. Despite the limited data is used, the model performed very well in predicting patients at risk of lung cancer, with a sensitivity of 75.3% and specificity of 80.6%. The model identified 649 cancer and 488,418 non-cancer cases using attributes such as age, gender, BMI, diabetes, smoking status, heart disease, and history of stroke. The AUC of 0.86 (0.85–0.88) for the training and validation sets indicates a promising tool for lung cancer risk prediction. Another method for identifying lung cancer risk factors based on the deep neural networks that is presented in [18]. The method is implemented on the web-based using Behavioral Risk Factor Surveillance System data source from 1997 to 2017 to realize the lung cancer risk factors. They analyzed more than 7 million records and excluded missing factors to obtain over 230,000 records for further selection. By leveraging the weights of DNN models, they identified notable lung cancer risk factors and quantitatively analyzed their degree of influence on lung cancer incidence in the elderly. For men at age of 65 years or older, the risk factor was smoking frequency, followed by time since quitting, use of e-cigarettes, and having smoked at least 100 cigarettes in their lifetime.

In general, the several studies focused on the development of expert systems and rule-based inference engines; the other methods used ML techniques such as deep learning, fuzzy logic, and data mining to improve diagnosis accuracy. Additionally, the models varied in their inputs, with some studies used patient data that obtained from the surveys, while others used imaging data.

3 Proposed Method

3.1 Data Selection and Analysis

Our method aim to build a model that could serve as a DSS for individuals at high risk of lung cancer. The BRFSS data is a join project of all 50 states participating in the US territories, as well as the CDC and Prevention. It aimed at collecting uniform state-specific data on health risk behaviors, chronic diseases and conditions, access to health care, and utilization of preventive medical services related to the prominent causes of illness and death in the US. The survey is conducted in both landline and mobile phone-based surveys with individuals over the age of 18, and in 2020 (see Fig. 1). It assessed health status and healthy days, exercise, sleep problems, chronic diseases, dental health, tobacco use, cancer screenings, and healthcare accessibility as general factors. An important part of building a model with predictive accuracy is the used data for training, testing and evaluating. Therefore, data selection is a critical step in our research. Our data selection process is started with the exclusion of individuals who did not satisfy the age requirement of being 18 or older. We then excluded missing demography data and behavioral factors that might harm the model performance. To address the issue of imbalanced data, we excluded non-lung cancer

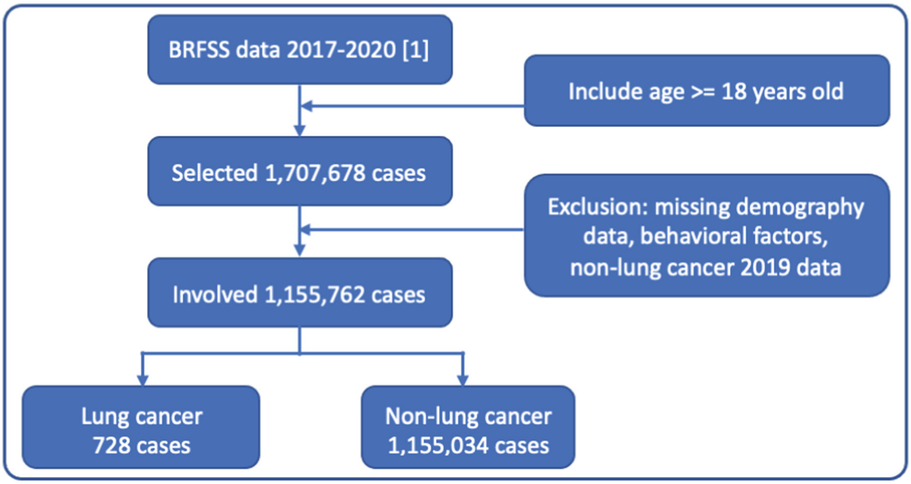
BRFSS data 2017-2020 [1]
Include age >= 18 years old
Selected 1,707,678 cases
Exclusion: missing demography data, behavioral factors, non-lung cancer 2019 data
Involved 1,155,762 cases
Lung cancer 728 cases
Non-lung cancer 1,155,034 cases

Fig. 1. Data selection plan of BRFSS from 2017–2020 [1]

data in year 2019, since the ratio of respondents with and without lung cancer is 1:30000, potentially causing significant imbalance issues in our final data. We use the Synthetic Minority Oversampling Technique (SMOTE) [19], as a powerful pre-processing technique to address the issue of imbalanced data and enhance the performance of our predictive models. The final dataset contains information unique to survey participants. We selected the 13 most referenced lung cancer risk factors out of more than 279 different features. These features include demographic data such as age, sex, and body mass index (BMI), as well as smoking history, general health status, and lung cancer screening history. The age feature is categorized into two levels, $18 \leq AGE \leq 64 : 0$ and $65 \leq AGE \leq 99 : 1$. The sex feature is represented by two levels, male: 1 and female: 0. The BMI feature is computed based on four categories: underweight, healthy weight, overweight, and obese. Notably, the most frequent category in this column is overweight, which is consistent with the trend observed in the US population from 2017 to 2020. To provide an additional feature for the intelligent support system, we presented a BMI calculator. Patients who do not know their exact body mass index can input their respective weight and height. The system will automatically calculate and categorize them into one of the four ranges. The BMI encoding will then entered the model as another feature, encoded as 1 to 4, respectively, at the backend.

The input values for underlined variables have been modified from using 1 and 2 to binary encoding with 0 and 1, as it offers certain advantages in terms of interpretation and model fitting. While changing a small dataset from 0 to 2 and keeping 1 as it is may not result in significant changes in model performance, consistent encoding becomes crucial for larger datasets. This approach facilitates better recall and benefits loss functions like binary cross-entropy, which is used to evaluate the neural network model in this study. For instance, to consider the height data of women and men with a regression equation of $height = a + b * gender + residual$. With the dummy variable of 0 and 1, the average height of women can be estimated to be 170 and the difference between the average height of men and women is 10. However, with the dummy variable of 1 and 2, estimating the average height of women would be harder, resulting in a less interpretable model. The general health feature consists of five categories: excellent, very good, good, fair, and poor. The smoking history features include two fields: "smoked at least 100 cigarettes" and "four-level smoker status." Finally, we selected all the questions in the lung cancer screening section added after the year 2016, as they are the only lung-specific questions available in the BRFSS data.

3.2 Learning Methodology

BMI Calculator and Classification: The first function of the system is simple to automatically calculate the BMI of patients through their weight and height inputs. They are classified into four ranges for encoding feature feeding model training and serving purpose. The formula to compute BMI and the logic to classify them are presented as in [21]: $BMI = weight/((height * 0.01) * (height * 0.01))$. If (BMI $<$ 18.5) then BMI = 1 (you are underweight); If (BMI $\geq$ 18.5 and BMI $<$ 24.9) then BMI = 2 (you are healthy weight); If (BMI $\geq$ 25 and BMI $<$ 29.9) then BMI = 3 (you are overweight); If (BMI $>$ 30) then BMI = 4 (you are obese); else BMI = 0.

Logistic Regression: LR is a statistical method used to analyze and model the relationship between a categorical dependent variable and one or more independent variables. In medical applications, it can be used as a prognostic or diagnostic index to classify patients based on their risk of developing a certain disease. The LR model is based on the *logit* transformation, which computes the probability of output with the explanatory variables taken into account. The transformation is written as $logit(p)$, where p is the proportion of objects that have a certain characteristic. In medical applications, p can be seen as the probability of a patient having a specific disease. The following formula [23] shall be used in order to transform the value of the dependent variable into a binary one: $logit(p) = ln\frac{p}{1-p} = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_m * x_m$. The above formula obtains the value α so as to calculate for p basic algebra is used to resolve the equation: $p = \frac{1}{1+e^{-\alpha}}$. In order to compute the intercept and the regression coefficients, the matrix form of the model is used: $logit(p) = X * b + \epsilon$, where:

$$logit(p) = \begin{pmatrix} ln(\frac{p_1}{1-p_1}) \\ \vdots \\ ln(\frac{p_n}{1-p_n}) \end{pmatrix} \quad and \quad X = \begin{pmatrix} 1 & x_1^1 & \dots & x_m^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^n & \dots & x_m^n \end{pmatrix} \tag{1}$$

thus: $b = (X' * X)^{-1} * X' * ln(\frac{p}{1-p})$. Logistic regression is often used as a binary classification algorithm, meaning it classifies a patient as either having or not having a certain disease. In the case of lung cancer, for example, the logistic regression model can be used to classify patients who are at higher risk of developing lung cancer compared to others within the group.

To compute the risk of lung cancer in a patient, the logistic regression model takes into account various risk factors such as age, gender, smoking habits, body mass index, and medical history. These risk factors are selected based on their association with lung cancer. The first step in developing the model is to construct the model and define its intercept and coefficients after training the model to get the complete equation.

Interpreting the results of the logistic regression model involves understanding the equation and the values of the coefficients. For example, a patient who is 55 years old, male, a daily smoker, has a healthy weight range, started smoking at age 18, stopped smoking at age 0, smokes an average of 20 cigarettes per day, has fair general health status, has smoked more than 5 packs of cigarettes in their lifetime, attempted to quit smoking, and does not have any respiratory diseases or asthma would have a $logit(p)$ of -2.437. This means that the probability of this patient having lung cancer is about 8.32%.

3.3 Artificial Neural Network

The ANN model in [24] is a powerful machine learning algorithm used for the early detection of lung cancer. Our proposed process of design and implementation is presented as in Fig. 2. We uses the feed-forward backpropagation algorithm, which adjusts the weights in the network through each iteration to reduce errors. The architecture of the feed-forward backpropagation consists of 13 input neurons in the input layer representing significant symptoms of lung diseases, 100 hidden neurons in the first and second hidden layers, and one output neuron in the output layer representing the presence or absence of lung disease (Table 1).

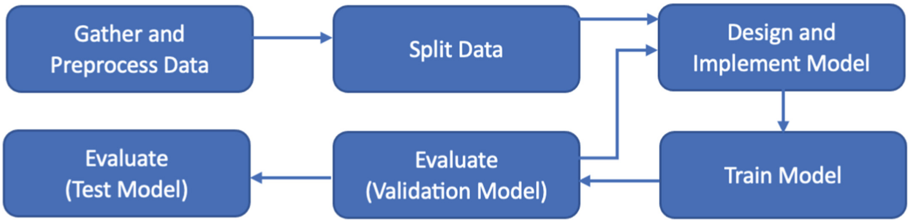
Gather and Preprocess Data
Split Data
Design and Implement Model
Train Model
Evaluate (Validation Model)
Evaluate (Test Model)

Fig. 2. Our process of design and implementation

After obtaining and pre-processing data, we design and setup information for the ANN architecture as follow:

Table 1. Configuration of ANN architecture.

Layers	Parameter	Value
1 (Input)	Neuron	13
2–5 (Hidden)	Neuron	24
	Activation	Relu
	Dropout	50%
6 (Output)	Neuron	1
	Activation	Sigmoid
Training times = 50 Batch size = 100 Optimizer = SGD Learning rate = 0.05 Train type = Feed forward Loss = Binary crossentropy		

As we known, the deep neural network is a modern variant of the ANN, which replaces the step activation function with something better, such as ReLU, and applies Sigmoid to the output. The neural net consists of three layers, the input layer, the hidden layer, and the output layer. Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value that is improved through each epoch. The decision is made as to whether the user suffered from lung cancer or not based on the output. The activation function of each layer is computed by applying a ReLU and a sigmoid function. When the sigmoid activation function is applied to the output, the final output of the neural network is converted into probabilities, from which one can choose the classification with the highest probability. The neural network's training method is a variant of gradient descent called backpropagation, which enables the adjustment of all weights at the same time. Binary cross-entropy is used as the loss function, as it is appropriate for binary classification. The cost function can be calculated for each epoch as the whole dataset is fed to the neural network, and the weights are updated until the goal is reached, which is to minimize the cost function. Stochastic Gradient Descent, on the other hand, takes the row one by one and then runs the neural network, and then adjusts the weights right after. This technique helps in finding the global minimum rather than getting stuck in the local one and is faster than batch gradient descent since there is no need to load all data into memory. The binary cross-entropy loss function will be used for system evaluation, where unseen data will be fed into the models ten times to see how the system will perform on real-world data. TensorFlow 2 will be used for model training, and visualization will be made using popular libraries such as seaborn and matplotlib. Backpropagation is driven by very interesting and sophisticated mathematics that enable the adjustment of all the weights at the same time, making it an efficient method for training neural networks.

4 Implementation and Evaluation

Data pre-processing and analysis is a critical phase in any data-driven research, and Spark has proven to be a valuable tool in this regard. Its application in this work allowed for the implementation of big data processing techniques, which led to a modern, interactive, and user-friendly approach to data pre-processing. The chosen database covers various behavioral risk factors for general health, including lung cancer. However, when pre-processing for logistic regression and neural network models for early prediction purposes, an issue of imbalance between respondents with and without lung cancer emerged. To address this, various techniques were used to balance the data and improve model performance. The data in the database comprises multiple explanatory variables of different levels of data, each encoded according to the status of each respondent. Therefore, mode imputation was employed during the pre-processing stage to fill in missing data and improve model performance. The process of data integration for each dataset was unique, with the inferSchema tool automatically defining the datatype for each field, leading to inconsistencies between datasets. For instance, the 2017 and 2018 datasets required the conversion of null values to numbers, which helped standardize the data types across all datasets. Compared to the original dataset, the processed data was cleaned, balanced, and ready for training and fitting machine learning models. This dataset served as the foundation for the models that produced the predictions in this study. Hyperparameter tuning is a critical step in optimizing ANN models, as the performance of the model significantly depends on the values of hyperparameters. However, determining the optimal values for hyperparameters is a challenging task, and trying all possible combinations manually can be time-consuming and resource-intensive. To automate this process, we used GridSearchCV, which is an iterative tuning process that helps to determine the optimal hyperparameters for ANN models. In our study, it took more than three hours to find the optimal hyperparameters for the ANN model using GridSearchCV. We also learned from our previous attempts to optimize the model manually. The first step in our predictive model is to input the significant symptoms from the subject's test data into the neural network and compute the net output value from the output layer. If the output value is less than the predefined threshold, the output neuron inhibits and displays "Low risk of lung cancer" along with the probability of lung cancer presence. If the output value is greater than the threshold, the output neuron fires and displays "High risk of lung cancer". As a result, ROC curves were developed and used in various applications and research. The ROC curve is useful because one point in space is considered better than another if it is positioned more to the northwest of the square. Predictions are better if the false positives rate is low and true positives rate is high, and the area under the curve (AUC) was introduced as a measure of performance. The AUC value falls within a specific range and shows the quality of classification. For example, a range of 0.9 to 1.00 is considered excellent classification, 0.8 to 0.9 is considered good, and 0.7 to 0.8 is considered fair. In contrast, the rate is considered a failure if it is below 0.5 [24]. To visualize the models, we used the ROC curve and the accuracy of the training and validation data. First, we split the data into training and testing sets and then performed imputation and oversampling on the training data (see Fig. 3) (Table 2).

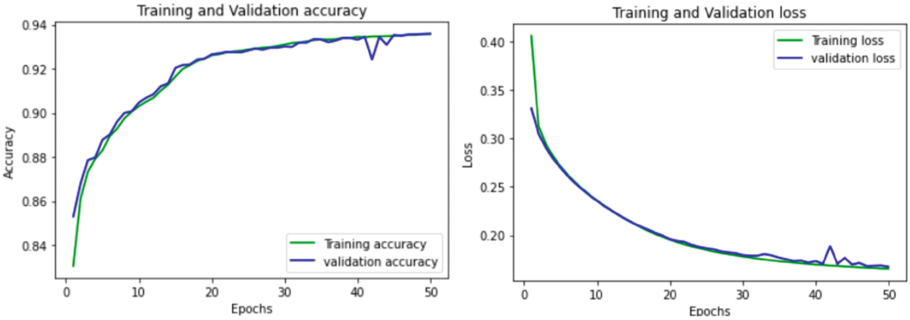

Fig. 3. The obtained results of our training and validation ANN model

Table 2. Comparison of the metric index between the models.

Model	Accuracy	Sensitivity	Specificity	AUROC
Our ANN Model	84.79%	79.80%	89.76%	93.00%
Our logistic regression Model	80.20%	80.00%	72.20%	76.10%
Multi-parameterrized ANN [20]	No	75.30%	80.60%	86.00%

The models are trained 50 times with a batch size of 100 and optimized using stochastic gradient descent. Our ANN model showed (accuracy: 84.79%, sensitivity: 79.80%, specificity: 89.76% and AUROC: 93.00%). While the LR model obtained (accuracy: 80.20%, sensitivity: 80.00%, specificity: 72.20% and AUROC: 76.10%). Comparing to the Multi-parameterrized ANN (the accuracy is not shown, sensitivity: 75.30%, specificity: 80.60% and AUROC: 86.00%).

5 Conclusion and Future Work

This study developed the ANN and LR models to aid an early detection of lung cancer. The proposed models utilized data mining techniques to discover new patterns in health behavioral risk data that had not been previously discovered. The contribution in this study is answered the questions: how the data mining techniques can be used to support detecting lung cancer; and how to increase the interpretability and trustworthiness of the developed models. A decision support system is developed to aid user in decision-making and recommended next steps in possible treatment plans. To address the training problem, the lightweight model with high accuracy is recommended and training on the computer with limited resources. The visual insights are utilized to address model failure and improve the architecture. A web application is developed for lung cancer detection through risk factors, utilizing deep algorithms and state-of-the-art evaluation metrics. The obtained results have been shown better data for more synthetic prediction, dealing with skewed data, improving pre-processing steps for better calculations. Future work includes improving the dataset of the lung cancer centric; setting up a complete database architecture, and optimizing hyperparameters for more precise diagnostics and recommendations.

References

1. Centers for Disease Control and Prevention. CDC - BRFSS, Centers for Disease Control and Prevention. Centers for Disease Control and Prevention (2023). https://www.cdc.gov/brfss/index.html. Accessed 11 Apr 2023
2. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics (2018). CA Cancer J. Clin. **68**, 7–30 (2018). https://doi.org/10.3322/caac.21442
3. Tiwari, A.: Prediction of lung cancer using image processing techniques: a review. Adv. Comput. Intell. Int. J. (ACII) **3**, 1–9 (2016). https://doi.org/10.5121/acii.2016.3101
4. Donoso, L.: Europe's looming radiology capacity challenge: a comparative study. For me the main threat is the shortage of radiologists, ESR President 2015/2016 Healthmanagement.org, vol. 16, no. 1 (2016)
5. Turban, E., Sharda, R., Delen, D.: Decision Support and Business Intelligence Systems, 9th edn. Pearson Education, New Jersey (2011). ISBN: 978-0136107293
6. Hamilton, W., Peters, T.J., Round, A., Sharp, D.: What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. Thorax **60**(12), 1059–65 (2005). PMID: 16227326; PMCID: PMC1747254. https://doi.org/10.1136/thx.2005.045880
7. Nguyen, L.D.V., Chau, V.V., Nguyen, S.V.: Face recognition based on deep learning and data augmentation. In: Dang, T.K., Küng, J., Chung, T.M. (eds.) FDSE 2022. CCIS, vol. 1688, pp. 560–573. Springer, Singapore (2022)
8. Phung, L.K., Nguyen, S.V., Le, T.D., Maleszka, M.: A research for segmentation of brain tumors based on GAN model. In: Nguyen, N.T., et al. (eds.) ACIIDS 2022. LNAI, vol. 13758, pp. 369–381. Springer, Cham (2022)
9. S Nageswaran, et al.: Lung cancer classification and prediction using machine learning and image processing. BioMed Res. Int. **2022**, Article ID 1755460, 8 pages (2022). https://doi.org/10.1155/2022/1755460
10. Diego, A., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. J. Nat. Med. **25**(6), 954–961 (2019)
11. Nguyen, V.S., Tran, M.H., Le, S.T.: Visualization of medical images data based on geometric modeling. In: Dang, T.K., Küng, J., Takizawa, M., Bui, S.H. (eds.) FDSE 2019. LNCS, vol. 11814, pp. 560–576. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35653-8_36
12. Nguyen, V.S., Tran, M.H., Vu, H.M.Q.: An improved method for building a 3D model from 2D DICOM. In: Proceedings of International Conference on Advanced Computing and Applications (ACOMP), pp. 125–131, IEEE (2018). ISBN: 978-1-5386-9186-1

13. Singla, J.: The diagnosis of some lung diseases in a prolog expert system. Int. J. Comput. Appl. **78**, 37–40 (2013)
14. Rodiah, E.H., Fitrianingsih, Susanto, H.: Web based fuzzy expert system for lung cancer diagnosis. In: International Conference on Science in Information Technology (ICSITech), p. 142 (2016)
15. Yatish Venkata Chandra, E., Ravi Teja, K., Hari Chandra Siva Prasad, M., Mohammed Ismail, B.: Lung cancer prediction using data mining techniques. Int. J. Recent Technol. Eng. (IJRTE) **8**(4), 12301–12305 (2019). ISSN: 2277–3878
16. Cassidy, A., Duffy, S.W., Myles, J.P., Liloglou, T., Field, J.K.: Lung cancer risk prediction: a tool for early detection. Int. J. Cancer **120**(6), 1–6 (2007)
17. Chada, G.: Using 3D convolutional neural networks with visual insights for classification of lung nodules and early detection of lung cancer (2019)
18. Songjing-Chen, S.W.: Identifying lung cancer risk factors in the elderly using deep neural networks: quantitative analysis of web-based survey data. J. Med. Internet Res. **22**(3), e17695 (2020)
19. Maldonado, S., López, J., Vairetti, C.: An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. **76**, 380–389 (2019)
20. Hart, G., Roffman, D., Decker, R., Deng, J.: A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS ONE **13**, e0205264 (2018). https://doi.org/10.1371/journal.pone.0205264
21. What Is Lung Cancer? The American Cancer Society. https://www.cancer.org/cancer/lung-cancer/about/what-is.html. Accessed April 2023
22. Nasser, I.: Lung cancer detection using artificial neural network. Int. J. Eng. Inf. Syst. (IJEAIS) **3**(3), 17–23 (2019). https://ssrn.com/abstract=3700556
23. S. Belciug and F. Gorunescu. Intelligent Decision Support Systems – Journal Smarter Healthcare, 1st edn. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-14354-1
24. Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., Doll, R.: Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ **321**(7257), 323–329 (2000)