Sivas Cumhuriyet University
Journal of Engineering Faculty

| cumfad.cumhuriyet.edu.tr | Founded: 2023 Available online, ISSN Publisher: Sivas Cumhuriyet Üniversitesi

Using the Random Forest Model to Improve Lung Cancer Diagnosis and Prognosis: A Pilot Study

Esin Bilgin[1,a], Michelle M. Zhu[1,b]

[1]*Montclair 07043, NJ, USA*

Research Article

History

Received: 28/11/2024
Accepted: 13/12/2024

ABSTRACT

There have been increasing research interests and efforts on the application of IoT network to healthcare systems. Our study primarily highlights the effectiveness of the Random Forest (RF) model in analyzing Lung cancer datasets and recognizing trends and patterns. Our study can detect different stages of lung cancer and predict the disease progression. Our work can offer automated and personalized diagnosis using a machine learning approach. The effectiveness of the random forest algorithm was used in the detection of lung cancer, which is an important problem area in the healthcare system, and successful results were obtained. Our approach can provide useful advice to healthcare professionals as a decision support method. Performance metrics such as accuracy, precision, recall and F1 scores were collected to measure the effectiveness of the model. This research has great potential in the diagnosis and treatment of lung cancer patients with low errors and fast speed.

Keywords: Lung cancer prediction, lung cancer classification, machine learning, random forest.

[a] bilgine2@montclair.edu ORCID | [b] zhumi@montclair.edu ORCID

How to Cite: Bilgin E, Zhu MM (2024) Using the Random Forest Model to Improve Lung Cancer Diagnosis and Prognosis: A Pilot Study, Journal of Engineering Faculty, 2(2): 197-204

Introduction

Cancer is a disease that men and women have long sought to treat, unfortunately, it often leads to fatal outcomes. Lung cancer ranks as the most fatal cancer, responsible for more deaths than any other type, largely due to its late-stage diagnosis. While treatment is crucial, early detection is essential for survival rate. According to a study published in Statista [1], cancer is the second leading cause of death in the United States. Lung cancer ranks first numerically in cancer-related deaths in both men and women. There are two main types of lung cancer: i) small cell and ii) non-small cell. Non-small cell lung cancer is the most common type of lung cancer, and small cell lung cancer occurs most frequently in heavy smokers. Of course, this type of cancer has basic symptoms, like other diseases, but the thing to pay attention to is to check whether the cancer has spread to other organs.

Machine learning and its methods are rapidly finding a place in every aspect of our lives, and their importance is also increasing in the field of health. Especially, their use as a decision support system has enabled disease detection and diagnosis to produce successful results. One of these is the detection of lung cancer, which humanity has been struggling with for many years and trying to find a solution for.

In this study, unlike other studies, lung cancer prediction and classification were made using real data obtained from real-life individuals instead of synthetic, that is, produced data.

The paper is organized as follows. In Section II, we provide background information on Lung Cancer and Machine Learning, introduce Random Forest algorithm and Lung Cancer types, and give an overview of Integration of the artificial intelligence into lung cancer classification and prediction. Section III explores related studies in literature. The RF algorithm that generates predictions about lung cancer classification is detailed in Section IV. Section V and Section VI outline the experimental settings and discuss the results. A discussion stating the strengths and downsides of proposed approach is given in Section VII. Finally, Section VIII concludes the study.

Background

Lung Cancer

Lung cancer primarily develops in the cells that line the airways within the lung tissues [3]. The lungs' essential role in delivering oxygen to the body is disrupted by the unchecked proliferation of atypical cells in one or both lungs [4]. Lung cancer, as seen in Figure 1, can be classified according to the size of the tumor cells into different subtypes which is adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell carcinoma (LCC), and other variants. All these NSCLC subtypes occupy approximately 85% of all lung cancer cases. LUAD is now the most frequently diagnosed NSCLC with histopathologic feature of glandular origin in the peripheral zone of the lung, and the histological type with a stronger association with non-smokers and younger people. The second subtype, which is LUSC, develops from flat cells in the lining of the airways in the central zones of the lungs and is directly associated with smoking. The third subtype is LCC characterized by aggressive growth; tumor cells are large and exhibit no features of the other types and finds rare [5]. Besides these, the Other Variants of NSCLC include adenocarcinoma with squamous differentiation (adenosquamous carcinoma), spindle-shaped or sarcomatoid carcinoma, and carcinoid tumors, which together contribute less than 5% of NSCLC cases [6]. Another subtype of lung cancer is small cell lung cancer (SCLC) contributing 15% of the lung cancer cases, is closely related to smoking, has a high rate of metastasis, and demonstrates neuroendocrine

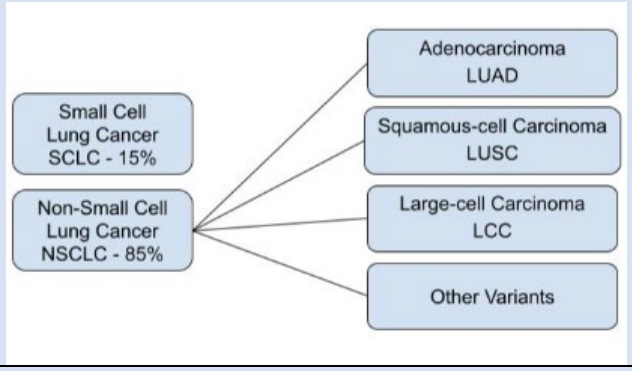

Figure. 1. Lung Cancer Types.

Primary risk factors associated with lung cancer include smoking, which is the most important factor, as well as exposure to second-hand smoke, air pollution, radon gas, asbestos, and various other environmental toxins. In addition, one of the factors that significantly increases the risk of contracting the disease is genetic predisposition. Symptoms of lung cancer usually occur in their advanced stages and include persistent cough, chest pain, difficulty

breathing, unexplained weight loss and fatigue. Lung cancer, unfortunately, has a high mortality rate due to the frequent occurrence of late-stage diagnosis. Its high mortality rate unfortunately leads to more deaths globally compared to other types of cancer [9]. On the other hand, advances in early diagnosis of the disease increase the patient's chances of survival and are promising in health science.

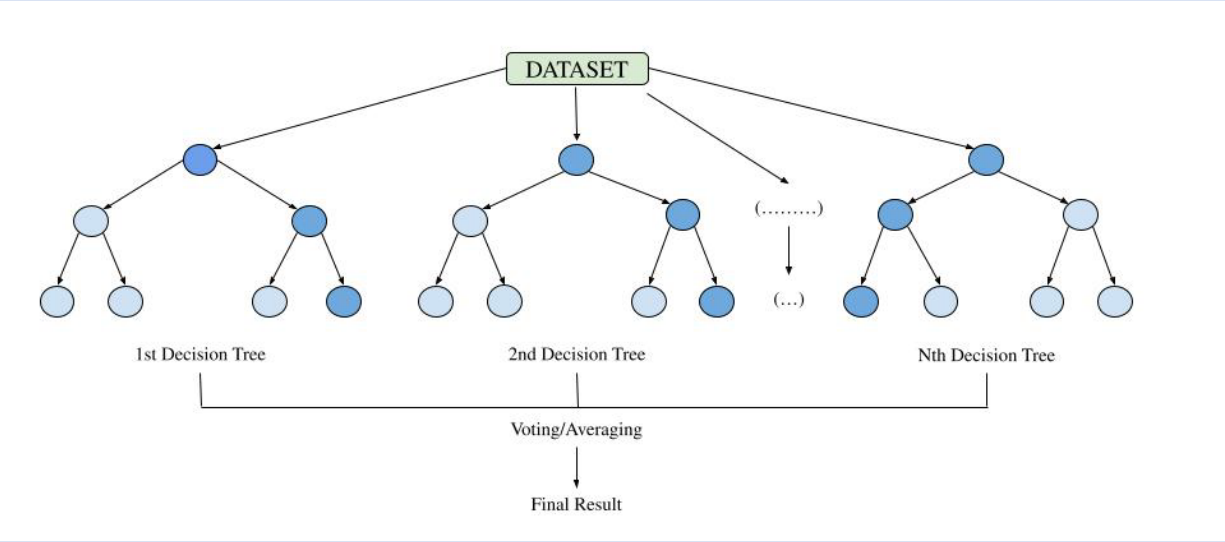

Figure. 2. Architecture of the random forest algorithm

Lung Cancer Prediction And Machine Learning

Due to the high mortality rate and complexity of patients with lung cancer [10], researchers have recently shifted their interests in early detection and diagnosis of the disease. The goals of lung cancer prediction are not only to improve early diagnosis but also to evaluate risk factors and even offer personalized treatment options. The advancement and integration of technology and data science and admirable developments in the field of deep learning have changed the approach and technical requirements for the prediction of lung cancer, as in many areas in the field of healthcare. These technologies allow the analysis of large amounts of patient data, examination images and other factors collected at different times, providing a more in-depth analysis of factors affecting lung cancer risk and progression.

Models used to predict lung cancer generally include risk factors such as family history of cancer and genetic pre-dispositions, as well as factors such as smoking, exposure to air pollution, occupational hazards [11]. Early screening and interventions can only be implemented by identifying individuals at risk [12] and improvements can be observed because of the application. Personal applications play a critical role in successful lung cancer prediction [13].

Machine learning models can predict certain predispositions, including genetic predisposition, to predict people's risk of developing lung cancer, as well as their response to certain treatments. Analysis and detection of genetic risk factors related to lung cancer enable effective treatments for this individual. Machine learning models are also used to predict the patient's survival rate based on many factors, such as understanding the stage of cancer, the characteristics of the cancer cells, the individual characteristics of the patient's body, and the individual's body's responses to the treatment administered to the individual. This allows the doctor or medical staff observing the patient to detect and manage the recovery process more effectively. Therefore, due to its great success, machine learning has become indispensable in lung cancer prediction [14].

ML can analyze diverse and complex data sets, such as patient demographics, possible history, density ratios, examination data, and genetic information. By identifying and describing patterns in these data sets, ML models provide more accurate predictions than classical prediction methods. ML algorithms have significantly accelerated the early detection of lung cancer. By analyzing examination data or imaging in a short time with deep learning, the presence or absence of cancer cells can be detected from different angles, even with small changes in the model.

ML models, such as Random Forest, Support Vector Machine (SVM), and deep learning approaches, are highly effective for cancer predictions [15],[16],[17]. These learning models can combine different types of data (e.g. imaging, genetic data) in the solution space to provide more accurate and personalized predictions for lung cancer recognition and detection. Additionally, these methods can help predict how patients will respond to certain treatments based on their unique genetic and examination profiles. This not only allows treatment plans to be more personalized, but it can also predict the subtype of cancer, if any, and even determine treatments appropriate for these subtypes, allowing the process to be faster and more efficient.

Like many healthcare datasets, lung cancer datasets are often known as unbalanced datasets [18]: for example, malignant cases outnumber benign cases. The imbalance here, or in other words, numerical inequality, can cause traditional models to make errors. On the other hand,

machine learning algorithms can easily overcome such challenges by developing models against imbalance or providing optimization, such as hyperparameter optimization, thereby improving model compatibility and accuracy. Unlike traditional and mostly static models, machine learning systems can be integrated into data augmentation and therefore continuously improve [19]. In other words, it is generalizable in terms of the amount of data. This feature allows the data obtained during research and inspection processes to be constantly updated and new data added.

Random Forest Algorithm

The Random Forest (RF) algorithm is a robust learning method used frequently and primarily for classification and regression tasks. Below is the mathematical expression of a basic equation for this algorithm:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(X) \quad (1)$$

The architecture of RF is based on creating many decision trees, thus optimizing the trees it creates and calculating the average prediction of individual trees. On the other hand, it has the ability to analyze complex and large data sets with many features and even reduces the overfitting is usually seen in single decision trees.

Mechanism and advantages of Random Forest

RF randomly mirrors a subset of the training data from the solution space in order to create different training sets for each decision tree and therefore increase efficiency. This allows each tree to learn from a different subset of data and increases the diversity among the resulting land trees. For each tree, a random subset of features is selected while discriminating across nodes in the forest. This randomness allows our model to run without any dependencies and increases the robustness of the model. Decision trees are created without dependency, and each tree is grown to its maximum depth. Once all the trees are trained, RF makes various predictions by either adding the individual predictions made by each of the individual trees to the majority vote, or by calculating the average of the predictions. This aggregation process reduces noise, reducing the risk of overfitting and increasing prediction accuracy. Thanks to the ensemble nature of the algorithm, Random Forest is less likely to overfit than single decision trees when working with noisy data. The algorithm can effectively deal with missing values by using existing data from different trees. RF provides information on the importance of features, allowing practitioners to determine which features contribute most to model predictions. Figure 2 illustrates the basic architecture of random forest algorithm.

Related Work

In recent years, the intersection of healthcare and advanced technologies has led to a notable advance in machine learning techniques for lung cancer prediction, detection and prognosis [20][21]. This integration has not only empowered diagnostic practices but has also occupied huge research domain that explores the potential of ML to improve the efficiency of medical predictions. Hence, researchers are increasingly focused on leveraging these real-word technologies to address the complex challenges associated with lung cancer diagnosis and survival prediction [22][14]. A detailed review of the literature shows that studies in this domain can be classified into two groups. The first, which investigates the biological and clinical basis of lung cancer, and the second, which develops and applies machine learning-based models for prediction and classification. Computer-Aided Detection (CAD) based systems aiding radiological evaluations have led to the growth of machine learning as a powerful tool for improving lung cancer detection. Early on, CAD applications were focused on lung nodules in chest radiographs [23]. Soon, the technology was applied to computed tomography (CT), adding new detection methods. These recent CAD systems with selection node enhancement and automatic rule-based classification enable the effective detection of lung nodules in thin section CT scans [24]. By isolating nodules from their normal anatomical structures such as blood vessels, which are significant contributors to false positive, these systems attain sensitivities above 86%, while maintaining the level of false positives in a controlled manner. CAD offers several advantages over digital mammography in early lung cancer screening prospects due to incorporation of ML techniques with three-dimensional (3D) imaging modalities that improve diagnostic accuracy [24][25]. On the other hand, it has been demonstrated that low dose computed tomography (LDCT) can reduce lung cancer mortality, studies have reported risk reductions of 20 to 43% [26][27]. LDCT screening suffers from variability between graders, as well as high rates of false positives and false negatives leading to screening inconsistency and unreliability [28][7]. In this work, Ardila et al. [25] addressed these issues with an advanced deep learning model for lung cancer risk assessment that uses both previous and current LDCT volumes. On data from the National Lung Cancer Screening Trial (NLST, n = 6717), their model had high accuracy with an area under the curve (AUC) of 94.4% and exhibited consistent behavior on a separate validation set (n = 1139). This model was tested in controlled reader studies where it performed 11 percent better at decreasing false positive rates and 5 percent better at decreasing false negative rates compared to six radiologists when prior imaging was not available, and when prior imaging was available, performed equivalently to the radiologist assessments. The results suggest that automation would probably boost both precision and accessibility of lung cancer screening by orders of magnitude and could help pave the way for more reliable screening worldwide in the era of deep learning and automated diagnostics.

Although CAD systems have been the subject of pioneering studies in lung nodule detection, recent investigations indicate that combining categorical clinical data into machine learning models may further improve diagnostic capabilities. Additional clinical data, like patient

history, risk factors, and demographic information offer additional complimentary insight, which can improve model specificity, and therefore how predictions are tailored more specifically. A lung cancer risk estimation tool was developed by Benveniste et al. [6], through integrating clinical factors like age, gender, smoking history, and comorbidities, using gradient boosting algorithms. The Brier score and ROC-AUC of 0.044 and 82% on the PLCO dataset and 0.70 and 70% on NLST dataset indicated a well calibrated performance on their model which underscores the aptitude of clinical data in lung cancer risk estimation Similarly, Aslani et al. [29] presented a time series deep learning model that combined longitudinal imaging data with patient specific clinical variables such as follow-up imaging results and smoking history. Combining temporal clinical with imaging data improved malignancy prediction: their approach reduced a false positive rate by 12%, and a false negative rate by 9% over traditional CAD systems.

Proposed Method

Random Forest was chosen for this study due to its excellent performance on classification tasks. The fact that it allows generalization in harmony with real-world data and its suitability to create harmony with the data set were also effective in choosing this algorithm. RF's ensemble learning approach improves classification accuracy by combining predictions from multiple decision trees, while also reducing the risk of overfitting rules with individual decision trees. This feature is particularly useful when developing complex datasets containing different features because it offers the ability to better generalize unseen. And it has been widely used for its secure and accurate results in vary fields including health domain [30]. Due to its ensemble structure and proven success in classification tasks, the random forest model is chosen for this study to classify lung cancer severity levels.

The model was implemented using a combination of clinical, demographic and lifestyle factors. These factors include age, gender, smoking history, environmental exposures, and genetic characteristics. RF's ability to detect the importance of defined features ensures that the classification process progresses successfully. This makes the model more understandable. In fact, it provides healthcare professionals with a wide range of information to understand the key factors that influence disease diagnosis and detection and, of course, treatment decisions.

Experiment

Dataset Description and Splitting

The dataset [31] was used in this study contains detailed information on patients with lung cancer, covering a variety of demographic, environmental, lifestyle and health-related factors that contribute to the risk and severity of lung cancer. Key variables include patient age, gender, air pollution exposure, alcohol use, dust allergies, occupational hazards, genetic risk, chronic lung disease, diet, obesity, smoking status (active and passive) and various symptoms of lung cancer such as chest pain, coughing of blood, fatigue, weight loss, shortness of breath, and difficulty swallowing.

Lung cancer remains one of the leading causes of cancer death worldwide. Environmental factors such as smoking and air pollution cause this disease. Recent research indicates that air pollution can increase the risk of lung cancer even in nonsmokers. The data set allows the examination of such relationships. In this way, the risk factors that lead to lung cancer can be revealed and how environmental influences and health factors affect cancer progression can be analyzed.

Before training the Random Forest model and for the other model to comparison, the dataset was carefully examined for inconsistencies or missing values. Since the dataset does not contain missing values, the model training process is facilitated without the need for estimation or additional preprocessing. Notably, the dataset does not contain missing values which allows the Random Forest algorithm (as well as for the other models for comparison) be applied directly.

In this study, the dataset was split into 70% for training and 30% for testing. Since the data was collected from real-life.

Performance Evaluation

After splitting the data, we classified these data using a random forest classifier. First, our proposed model was tested with classification accuracy. After the random forest algorithm produced a satisfactory accuracy value, other performance metrics were calculated along with accuracy to understand whether our model had a robust mechanism: Precision, Recall, F1 Score. The performance values obtained are shown in Table 1.

Table 1. Performance Metrices of the Random Forest Model for Lung Cancer Severity Prediction

	Accuracy%	Precision%	Recall%	F1 Score%
RF. Algorithm	95.01%	95.00%	95.01%	94.99%

Experimental Results

In order to better understand the performance of the Random Forest model, a comparison with some popular machine learning models was made. In addition to Random Forest, this comparison includes Decision Tree, Logistic Regression and Naive Bayes algorithms. Table 2 summarizes the performances of these models according to accuracy, precision, recall and F1-Score metrics

Table 2. Performance Comparison of Machine Learning Models for Lung Cancer Severity Prediction

	Accuracy%	Precision%	Recall%	F1 Score%
RF. Algorithm	95.01%	95.00%	95.01%	94.99%
Decision Tree	93.90%	93.84%	93.90%	93.86%
Logistic Regression	94.18%	94.18%	94.18%	94.16%
Naive Bayes	72.58%	75.35%	72.57%	71.28%
RF. Algorithm	95.01%	95.00%	95.01%	94.99%

The Random Forest classifier was trained and tested on a lung cancer severity dataset. Our proposed model was perfectly and easily accurate on the test set with an accuracy of 95.01%. As mentioned in the previous section and presented in Table 1, the values of our performance metrics of each class comfortably reached an apex point that generates a hyperplane. Here, for a fair comparison and to test the robustness of this hyperplane, the confusion matrix of 300 test samples shows no misclassifications, with all samples correctly classified according to their severity level i) Low, ii) Medium and iii) High, respectively. These findings suggest that the present model performs better than other models, effectively capturing the dynamics within the dataset. Figure 3 illustrates the importance of the predictors (features) of the adjusted RF algorithm that affects the severity of lung cancer. All these features were ranked high importance in the model, with Chest Pain, Smoking and Shortness of Breath having the highest importance.

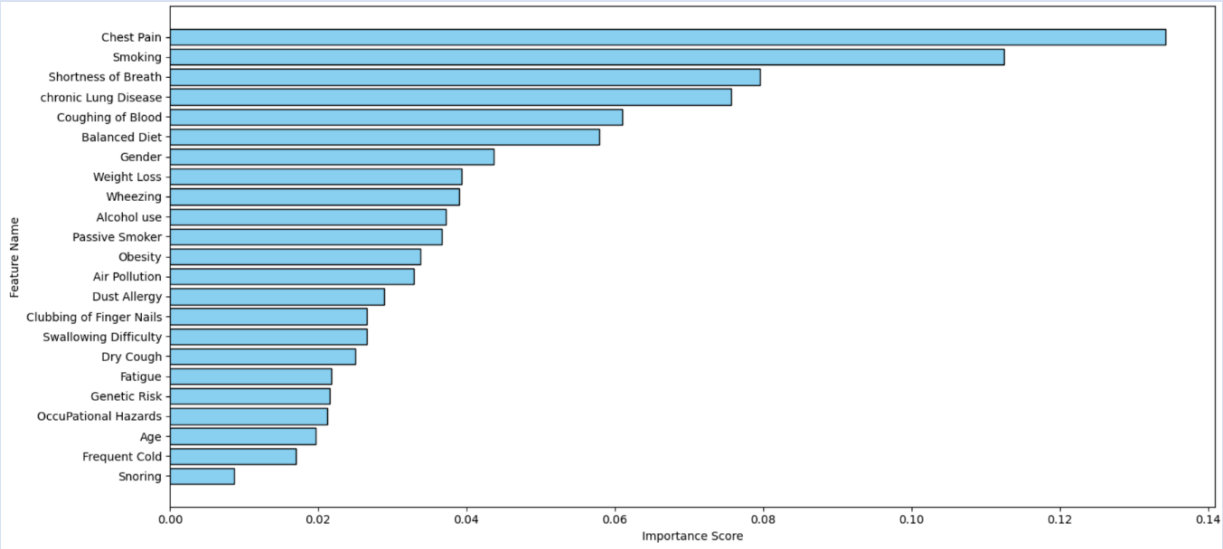

Figure 3. Feature Importance for Lung Cancer Severity Prediction

These results suggest that these features probably offer important clues to lung cancer severity. Balanced Diet, Chronic Lung Disease and Coughing Up Blood are also other important features, which signifies that the model

considers the input from a variety of clinical symptoms and life-style factors to make its prediction. Snoring and Age, lower ranked features, have little impact on the model's decision-making process; which means they are not so informative to classify severity levels in this dataset. Thus, this feature of importance insight is a nice plug for the interpretability of the model, since it indicates symptoms and risk factors that are associated with lung cancer severity and replicates common medical knowledge that is the transparent to the model

7. Strength, Weakness and Limitations

In this study, the use of the RF algorithm, which has a running structure based on the combination and comparison of many decision trees, ensured that the accuracy obtained in the classification is quite high. In other words, the classifier and the hyperplane created thus work quite well.

The one potential weakness of the model may be the use of data collected from people living in a particular region. However, it is super important that the data used in the study are obtained from real life and fortunately, it reduces such concerns. However, the model can be generalized and applied to different regions.

Additionally, the Random Forest model is known to be resistant against overfitting. This strength is due to the model using multiple decision trees to reduce noise and making predictions through majority voting or averaging. The dataset used in our study was obtained from a regional focus. This may create a potential risk of overfitting. However, in the study, the k=5 cross-validation method was applied to generalize the performance of the model. This method strengthened the separation of the training and test sets and prevented overfitting. This reliability was further supported by careful hyperparameter selection, ensuring the robustness of the results.

Conclusion and Future Directions

Technologies based on machine learning, such as machine learning and deep learning, have recently begun to be used extensively in the field of healthcare. It shows that machine learning is very important for the field of health, especially in matters such as obtaining successful results, protecting human health with early diagnosis and diagnosis. Nevertheless, in the context of intricate problems such as classification, the application of machine learning algorithms has demonstrated the capacity to achieve extraordinarily high degrees of accuracy. However, these advances bring with them different problems such as data quality issues and the need to analyze the results.

Data collected from a specific region may present limitations. These limitations may introduce potential biases that could impact applicability to broader audiences. Such limitations highlight the need for more diverse and representative datasets enhanced with other data types (e.g. like images, laboratory test results, etc.) in future studies to ensure broader applicability of the results.

Our proposed model used a classifier-based algorithm random forest, to detect lung cancer early through severity detection. In this way, selecting the optimal classifier among the different trees created by the random forest architect during the learning process increased the success of the model. Due to nature as an ensemble method which uses several decision trees and majority voting to make predictions, the Random Forest model is considered to be resistant to overfitting. K=5 cross-validation was used during model training to further reduce the risk of overfitting, guaranteeing the separation of training and test sets and improving the dependability of the outcomes. The proposed RF-based model achieved satisfactory performances (over 90%) i) accuracy, ii) precision, iii) recall and iv) F1-score.

Future research will concentrate on integrating data from many locations and people in order to overcome the shortcomings of the regional dataset utilized in this investigation.

In addition, the use of different and advanced machine learning techniques is planned for future studies. Among these, Convolutional Neural Networks (CNNs) stand out, as they are particularly effective in image classification tasks and in addition to some different machine learning techniques, this study is planned to be conducted as LSTM based which is especially well-suited for time-series models. It is prospected to be especially useful in cases involving patient monitoring and temporal dependencies.

References

[1] Statista, "Lung cancer - Statistics & Facts" https://www.statista.com/topics/8909/lung-cancer-in-the-us/#topicOverview.

[2] Hurriyet, "Akciğer kanseri evreleri yaşam süresi ve tedavisi — hurriyet.com.tr." https://www.hurriyet.com.tr/aile/akciger-kanseri-evreleri-yasam-suresi-ve-tedavisi-430064.

[3] A. F. Gazdar, P. A. Bunn, and J. D. Minna, "Small-cell lung cancer: what we know, what we need to know and the path forward," Nature Reviews Cancer, vol. 17, no. 12, pp. 725–737, 2017.

[4] E. L. O'Dowd, T. M. McKeever, D. R. Baldwin, S. Anwar, H. A. Powell, J. E. Gibson, B. Iyen-Omofoman, and R. B. Hubbard, "What characteristics of primary care and patients are associated with early death in patients with lung cancer in the uk?," Thorax, vol. 70, no. 2, pp. 161–168, 2015.

[5] W. D. Travis, E. Brambilla, A. G. Nicholson, Y. Yatabe, J. H. Austin, M. B. Beasley, L. R. Chirieac, S. Dacic, E. Duhig, D. B. Flieder, et al. "The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification," Journal of thoracic oncology, vol. 10, no. 9, pp. 1243–1260, 2015.

[6] Benveniste et al., "A gradient boosting model for lung cancer risk estimation using clinical data." https://arxiv.org/abs/2308.12188, 2023.

[7] C. S. White et al., "Lung cancer screening with low-dose ct: a white paper of the society of thoracic radiology and the american college of radiology," J. Thorac. Imaging, vol. 28, no. 5, pp. 295–306, 2013.

[8] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, "Progress and prospects of early detection in lung cancer," Open biology, vol. 7, no. 9, p. 170070, 2017.

[9] R. Sharma, "Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050," International Journal of Clinical Oncology, vol. 27, no. 4, pp. 665–675, 2022.

[10] S. Wankhade and S. Vigneshwari, "A novel hybrid deep learning method for early detection of lung cancer using neural networks, Healthcare Analytics, vol. 3, p. 100195, 2023.

[11] J. C. Laguna, M. Tagliamento, M. Lambertini, J. Hiznay, and L. Mezquita, "Tackling non–small cell lung cancer in young adults: From risk factors and genetic susceptibility to lung cancer profile and outcomes," American Society of Clinical Oncology Educational Book, vol. 44, no. 3, p. e432488, 2024.

[12] A. Issanov, A. Aravindakshan, L. Puil, M. C. Tammemägi, S. Lam, and T. J. Dummer, "Risk prediction models for lung cancer in people who have never smoked: a protocol of a systematic review," Diagnostic and Prognostic Research, vol. 8, no. 1, p. 3, 2024.

[13] S. Srivastava, N. Jayaswal, S. Kumar, P. K. Sharma, T. Behl, A. Khalid, S. Mohan, A. Najmi, K. Zoghebi, and H. A. Alhazmi, "Unveiling the potential of proteomic and genetic signatures for precision therapeutics in lung cancer management," Cellular Signalling, vol. 113, p. 110932, 2024.

[14] S. N. A. Shah and R. Parveen, "An extensive review on lung cancer diagnosis using machine learning techniques on radiological data: state-of-the-art and perspectives," Archives of Computational Methods in Engineering, vol. 30, no. 8, pp. 4917–4930, 2023.

[15] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using support vector machine, multilayer perceptron, and random forest algorithms," Healthcare Analytics, vol. 3, p. 100185, 2023.

[16] S. Khandakar, M. A. Al Mamun, M. M. Islam, K. Hossain, M. M. H. Melon, and M. S. Javed, "Unveiling early detection and prevention of cancer: Machine learning and deep learning approaches," Educational Administration: Theory and Practice, vol. 30, no. 5, pp. 14614–14628, 2024.

[17] M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," Healthcare Analytics, vol. 3, p. 100125, 2023.

[18] S. Padma, S. S. Kumar, and R. Manavalan, "Performance analysis for classification in balanced and unbalanced data set," in 2011 6th International Conference on Industrial and Information Systems, pp. 300–304, IEEE, 2011.

[19] M. Imran, H. U. R. Siddiqui, A. Raza, M. A. Raza, F. Rustam, and I. Ashraf, "A performance overview of machine learning-based defense strategies for advanced persistent threats in industrial control systems," Computers & Security, vol. 134, p. 103445, 2023.

[20] H. T. Gayap and M. A. Akhloufi, "Deep machine learning for medical diagnosis, application to lung cancer detection: a review," BioMedInformatics, vol. 4, no. 1, pp. 236–284, 2024.

[21] S. R. Quasar, R. Sharma, A. Mittal, M. Sharma, D. Agarwal, and I. de La Torre Díez, "Ensemble methods for computed tomography scan

images to improve lung cancer detection and classification," Multimedia Tools and Applications, vol. 83, no. 17, pp. 52867–52897, 2024.

[22] L. Bertolaccini, M. Casiraghi, C. Uslenghi, S. Maiorca, and L. Spaggiari, "Recent advances in lung cancer research: unravelling the future of treatment," Updates in Surgery, pp. 1–12, 2024.

[23] M. L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields.," Medical physics, vol. 15 2, pp. 158–66, 1988.

[24] Q. Li, F. Li, and K. Doi, "Computerized detection of lung nodules in thin-section ct images by use of selective enhancement filters and an automated rule-based classifier," Acad. Radiol., vol. 15, pp. 165–175, Feb 2008.

[25] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," Nat. Med., vol. 25, no. 6, pp. 954–961, 2019.

[26] C. I. Henschke et al., "Early lung cancer action project: overall design and findings from baseline screening," Lancet, vol. 354, no. 9173, pp. 99–105, 1999.

[27] D. R. Aberle et al., "National lung screening trial research team: baseline characteristics of participants in the randomized national lung screening trial," J. Natl. Cancer Inst., vol. 102, no. 23, pp. 1771–1779, 2010.

[28] G. Chassagnon et al., "Differentiation of subsolid pulmonary nodules by use of histogram analysis in contrast-enhanced ct imaging," Radiology, vol. 286, no. 3, pp. 1086–1096, 2018.

[29] Aslani et al., "Enhancing cancer prediction in challenging screen-detected incident lung nodules using time-series deep learning." https://arxiv.org/abs/2203.16606, 2022.

[30] Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[31] "Lung Cancer Prediction — kaggle.com." https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data.