Machine Learning Approaches for Lung Cancer Prediction

Alpre Emre Celik
Department of Computer Engineering
Istanbul Aydin University
Istanbul, Turkey
alpercelik@stu.aydin.edu.tr

Jawad Rasheed
Department of Computer Engineering
Istanbul Aydin University
Istanbul, Turkey
0000-0003-3761-1641

Amani Yahyaoui
Department of Software Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0003-0603-6592

2022 12th International Conference on Advanced Computer Information Technologies (ACIT) | 978-1-6654-1050-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACIT54803.2022.9913114

***Abstract*—Cancer is a long-term, exhausting disease that requires changes in all living conditions of the patient and his/her environment. Although there are regional variations in deaths from all causes in the world, it is in the 3rd rank. Lung cancer is among the most frequent cancer kinds worldwide, regardless of male or female. Cancer is a preventable disease. To prevent a disease, it is necessary to know its causes and avoid them. The use of tobacco and tobacco products is the main risk factor for all cancers, especially lung cancer. Early diagnosis of cancer is lifesaving. According to the Turkish Respiratory Research Association, 200,000 people are diagnosed with cancer every year in our country. With the accelerated developments in technologies and the digitalization of health services, a large amount of cancer data has been collected and this data has been used by many researchers, especially in low and middle-income countries, to reduce the cost of tests used to predict different cancer types and to predict different cancer types. This article is exploited various machine learning algorithms for predicting lung cancer. Experimental results show that random forest performed better by attaining 96.08% accuracy.**

Keywords—diagnostic system, decision tree, k-nearest neighbors, linear regression, random forest

I. Introduction

Cancer is one of the most prevalent causes of mortality among humans [1][2]. Lung cancer is one of the most frequent and lethal cancers, capable of wreaking havoc on the human body. Early cancer detection is critical for effective cancer therapy. Many lives can be saved if lung cancer is detected in its early stages. Lung carcinoma is another term for lung cancer, which is a malignant tumor that grows uncontrollably in lung cells and is recognized by its uncontrolled cell proliferation. Individuals around the globe suffers from such cancer that may eventually cause mortality. If it went untreated, this cancer may spread to other regions of the body and progress more slowly thus may not be detected in early stages.

Recent technological advancements has enabled scientists to introduce various diagnostic tools based on artificial intelligence in medical domain; such as COVID-19 [3], brain tumor [4], and Diabetes [5]. Similarly, several cancer research studies focus on certain cancer treatment options, that includes artificial intelligence, image processing and datamining. For instance, authors in [6] proposed deep learning based prediction model to estimate the lung cancer stage using computed tomography (CT) images and secured 86% accuracy. Researchers in [7] investigated three machine learning algorithms (neural network, support vector machine and decision tree) to predict the recurrence of lung cancer and its survivability. Another research [8] proposed convolutional neural network based non-linear cellular automata lung cancer prediction model that can predict with an overall accuracy of 98.49%. Beside artificial intelligence based applications, researcher also exploited computer vision and image processing techniques to predict lung cancer, such as authors in [9] introduced Gabor filter to extract useful features along with stochastic diffusion search combined with machine learning algorithms (neural network, decision tree, and Naïve Bayes) to facilitate practitioners in predicting lung cancer.

Cigarette smoking is the most common cause of lung cancer in adults. Individuals who smoke are at the greatest risk of developing lung cancer; but it may occur in those who do not smoke. Other things to consider are the duration and number of cigarettes you smoke, anxiety, alcohol intake, chronic diseases, and allergies [10]. The careful examination of side effects and hazard variables is essential in developing a strategy for identifying lung cancer patients. The use of data mining methods may make it feasible to anticipate the development of cancer tumors. To forecast and compare on an existing dataset, this study utilizes a variety of classification algorithms including decision tree [11], random forest [12], k-nearest neighbors [13], and linear regression [14]. Fig. 1 shows the proposed workflow of the suggested study.

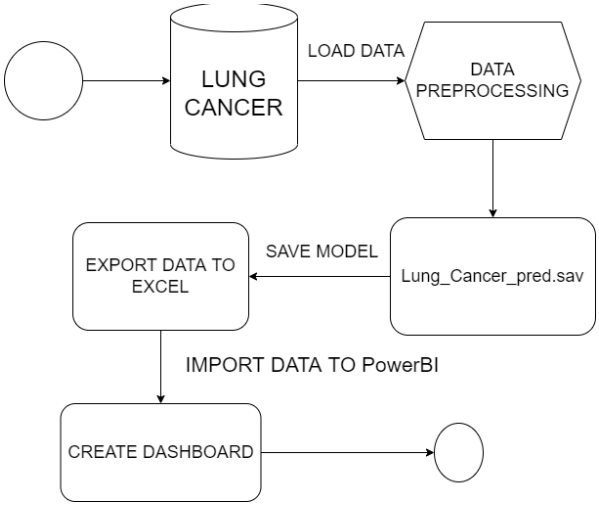

Fig. 1. Proposed workflow.

II. Materials

The dataset [15] used for this study is downloaded from Kaggle repository to predict lung cancer. The data series consists of 309 samples with 14 features in CSV format. There were 39 cases who do not have lung cancer, while 270 cases were diagnosed with lung cancer. The dataset includes various features, namely; gender, age, smoking, yellow finger disease, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, and chest pain.

In this study, we used all available features in all experiments. The dataset is divided into training (67%) and

testing (33%) sections. To minimize the biasness of proposed system, features are selected randomly to repeat the experiments ten times. Finally, it determines the accuracy as the average accuracy of all the experiments.

Fifty-five judgments will be generated as to whether the patient has lung cancer as a result of this method, which provides a separate or concurrent disease diagnosis of one or more patients as input to the model trained with the data beforehand.

Microsoft's PowerBI software is externally included in this model that works weekly with each new patient to display graphical results. As a result, this model, which is updated weekly by doctors and professors, can be used to determine the probability of lung cancer in the country and around the world. With new patient data added each week, this model will evolve into a self-taught model that improves itself and provides higher accuracy as it evolves. Users will benefit from this procedure as it will result in a more reliable decision-making system.

III. Methods

A. Linear Regression

A linear strategy to modeling the connection between a scalar answer and one or more explanatory factors is known as linear regression in statistics (also known as dependent and independent variables). Simple linear regression is used when there is only one explanatory variable, whereas multiple linear regression is used when there are more than one. Multivariate linear regression, on the other hand, predicts numerous correlated dependent variables rather than a single scalar variable.

The associations are represented using linear predictor functions, whose unknown model parameters are derived from the data via linear regression. Linear models are a type of model that fits such description. The conditional mean of the answer given the values of the explanatory variables (or predictors) is most usually considered to be an affine function of those values; the conditional median or some other quantile is less commonly assumed to be an affine function of those values.

Linear regression, like all other types of regression analysis, is concerned with the conditional probability distribution of the answer given the values of the predictors, rather than the joint probability distribution of all these variables, which is the domain of multivariate analysis. In general terms, the formula for linear regression is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \beta_i, \quad i = 1, \ldots, n, \tag{1}$$

B. Decision Tree

Decision tree [3] is a type of supervised machine learning that can classify or predict outcomes based on the responses to a series of previous questions. Decision tree isn't always required to provide a straightforward response or option. Instead, it can give the data scientist choices for making their own knowledgeable conclusions. A decision tree is like a tree in appearance. The root node is where the tree begins. A sequence of decision nodes flow from the root node, depicting the decisions that must be taken. The leaf nodes that sprout from the decision nodes depict the decisions' implications. Each decision node represents a query or a split point, and the leaf nodes that sprout from it reflect the various responses. Like how a leaf grows on a tree limb, leaf nodes develop from decision nodes. Therefore, each part of a Decision tree is referred to as a "branch." In general terms, the formula for decision tree is as follows:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}, \tag{2}$$

where ni_j, w_j, C_j, $left(j)$, and $right(j)$, refer to importance of node j, weighted number of samples reaching node j, impurity value of the node j, child node from left split on node j, and child node from right split on node j, respectively.

C. Random Forest

Random forest is a flexible and easily to use machine learning approach that in the most circumstances produces great results, without hyper parameter tweaking. It is a supervised learning approach that produces a forest out of an ensemble of decision trees, which are frequently trained using the bagging technique. The bagging method combines the numerous learning of the models to enhance the results.

Random forest has an advantage of being able to address classification and regression challenges, which makes up the bulk of modern machine learning systems. While growing the trees, the random forest adds additional unpredictability to the model. When dividing a node, it seeks for the best feature from a random set of qualities rather than the most critical trait.

Consequently, there is a lot of variation, which leads to a better model. In random forest, the approach for splitting a node only analyzes a random subset of the attributes. It is a form of classifier that improves from choice trees in terms of accuracy. It comprises of a vast number of selection trees. In order to analyze the instance, every optional tree gives an organizing for adding information; irregular random forest collects the characterization and determines the outcome based on the majority votes.

The contribution of each tree is determined by analyzing the info from the original dataset. To create the tree at each node, a subset of discretionary highlights is chosen at random from the discretion highlights. Each tree is allowed to grow organically without being trimmed. Essentially, irregular random forest allows a huge number of weakly coupled classifiers to merge and create a powerful classifier. In general terms, the formula for random forest is given as;

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \tag{3}$$

where fi_i and ni_j refer to importance of feature i and node j, respectively. These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \tag{4}$$

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each tree is calculated and divided by the total number of trees as given in (5):

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{Total\ trees}, \tag{5}$$

where $RFfi_i$, and $normfi_{ij}$ refer to importance of feature i calculated from all trees, and normalized feature importance for i in tree j, respectively.

D. k-Nearest Neighbor

The k-Nearest Neighbor is amongst the basic machine learning technique that follows supervised learning method. It considers the similarity between the newly formed state and data, and the previous states and assigns the new state to another category that that seems closer to the existing categories.

The k-Nearest Neighbor algorithm retains all previous data and uses similarities to classify new data points. This means that when new data is generated, it can be quickly classified into an appropriate category. It can be used for both regression and classification, however it is generally exploited for classification tasks. As it is a nonparametric approach, it makes no assumptions about the underlying data.

It records the information only during the training phase and classifies it in a category very similar to the new data when new data is received. It is also known as lazy learner algorithm as it cannot learn from the training set immediately. Instead, it keeps the dataset and takes an action on it when it's time to categorize it. In nonparametric statistics, the k nearest neighbor technique is used for data collection and relapse. It's worth noting that the data in both situations comprises the segment space's k-nearest development procedure models. The outcome is determined by whether k nearest neighbor is utilized for the demand or the backslide by denoting the set of the k nearest neighbors of $\boldsymbol{x}$ (test point) as $\boldsymbol{S_x}$, Formally, $\boldsymbol{S_x}$ is defined as $\boldsymbol{S_x} \subseteq \boldsymbol{s.t.}\ |S_x| = k$ and $\forall(x', y') \in D \backslash S_{x,}$.

$$dist(x, x') \geq \max_{(x",y")\in S_x} dist(x, x") \qquad (6)$$

(i.e every point in D but not in $\boldsymbol{S_x}$ is at least as far away from $\boldsymbol{x}$ as the furthest point $\boldsymbol{S_x}$). We can then define the classifier $h(\boldsymbol{x})$ as a function returning the mos common label in $\boldsymbol{S_x}$:

$$h(x) = model\ (\{y'' : (x'', y'') \in\ S_x\}), \qquad (7)$$

where mode $(\cdot)$ means to select the label of the highest occurrence.

IV. Experimental Results

For this study, the dataset contains 309 samples of different patients, where each sample has 16 attributes. This study utilizes 67% (208 samples) of the data for training and validation, and 33 percent (101 samples) for testing in all trials, as shown in Table I. It repeated experiments ten times to prevent model bias, and determined accuracy as the average of all experiments. Furthermore, because the initial data only included 14 characteristics (other than gender), therefore, no further analysis for feature selection or calculation of feature significance is done. Thus, it considered all the attributes equally essential.

TABLE I. Dataset Descriptione

Class Label [a]	No. of Instances in		
	Training set	*Test set*	*Total*
Cancer	182	88	**270**
Not cancer	26	13	**39**
Total	**208**	**101**	**309**

[a.] patient has lung cancer or not

Extensive experiments are carried out and each model performance is analyzed using various performance matrices that includes precision, f1-score, recall and accuracy. Tables II – V depict the performances obtained for linear regression, decision tree, random forest, and k-nearest neighbors, respectively.

TABLE II. Performance Analysis for Linear Regression Model

Class Label [a]	Performance Metrics		
	Precision	*Recall*	*F1-score*
Cancer	1.0	0.16	0.58
Not cancer	0.88	1.0	0.94
Accuracy	**-**	**-**	**0.60**
Macro average	**0.94**	**0.54**	**0.71**
Weighted average	**0.94**	**0.54**	**0.71**

TABLE III. Performance Analysis for Decision Tree Model

Class Label [a]	Performance Metrics		
	Precision	*Recall*	*F1-score*
Cancer	1.0	0.84	0.91
Not cancer	0.87	1.0	0.93
Accuracy	**-**	**-**	**0.92**
Macro average	**0.93**	**0.92**	**0.92**
Weighted average	**0.93**	**0.92**	**0.92**

TABLE IV. Performance Analysis for Random Forest Model

Class Label [a]	Performance Metrics		
	Precision	*Recall*	*F1-score*
Cancer	1.0	0.92	0.96
Not cancer	0.93	1.0	0.96
Accuracy	**-**	**-**	**0.96**
Macro average	**0.96**	**0.96**	**0.96**
Weighted average	**0.96**	**0.96**	**0.96**

TABLE V. Performance Analysis for K-Nearest Neighbor Model

Class Label [a]	Performance Metrics		
	Precision	*Recall*	*F1-score*
Cancer	1.0	0.76	0.87
Not cancer	0.81	1.0	0.90
Accuracy	**-**	**-**	**0.88**
Macro average	**0.91**	**0.88**	**0.88**
Weighted average	**0.90**	**0.88**	**088**

As mentioned before, although the classifiers and data mining methods that will be used when human health is at the forefront cannot work flawlessly, algorithms and classifiers with the margin of error and probability of being confused

should be preferred. Although the final decision is left to the doctors, in some cases or human-induced errors, the machine's ability to see and detect unpredictable results will have a great impact on people's lives. Each classification method used has yielded results with different accuracy. Among them, random forest classifier attains the highest result with an accuracy of 96.08%, followed by decision tree classifier with 92.17%, whereas k-nearest neighbor accomplished an overall accuracy of 88.26%, and finally linear regression with the lowest accuracy and the highest margin of error, with an accuracy of 60.30%, as shown in Fig. 2. This mean accuracy suggests that classification, not regression, is the best method if machine learning based models are desired to predict lung cancer.

Fig. 2. Accuracy comparison of exploited machine learning algorithm.

V. Conclusion

This study compared several machine learning-based algorithms for lung cancer prediction. Across all trial rounds, the random forest classifier outperformed others algorithms by attaining an overall accuracy of 96.08% for lung cancer prediction. The dataset used for this study contains 15 attributes including gender of the patients. For experiments, the dataset is split into 67/30 as train/test sets. It is noted that decision tree also performed better whereas linear regression performed worst. In the future, the study can be expanded by testing advanced deep learning models including pre-trained networks, so that patients can have right diagnoses at the right time for early detection in order to provide better and early treatment.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA. Cancer J. Clin.*, vol. 70, no. 1, pp. 7–30, Jan. 2020, doi: 10.3322/caac.21590.

[2] 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 Biol.*, vol. 7, no. 9, p. 170070, Sep. 2017, doi: 10.1098/rsob.170070.

[3] J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. Al-Turjman, "A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images," *Interdiscip. Sci. Comput. Life Sci.*, vol. 13, no. 1, pp. 103–117, Mar. 2021, doi: 10.1007/s12539-020-00403-6.

[4] A. Alnemer and J. Rasheed, "An Efficient Transfer Learning-based Model for Classification of Brain Tumor," in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Oct. 2021, pp. 478–482. doi: 10.1109/ISMSIT52890.2021.9604677.

[5] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Nov. 2019, pp. 1–4. doi: 10.1109/UBMYK48245.2019.8965556.

[6] Y.-W. Wang *et al.*, "Dual energy CT image prediction on primary tumor of lung cancer for nodal metastasis using deep learning," *Comput. Med. Imaging Graph.*, vol. 91, p. 101935, Jul. 2021, doi: 10.1016/j.compmedimag.2021.101935.

[7] Y. Yang, L. Xu, L. Sun, P. Zhang, and S. S. Farid, "Machine learning application in personalised lung cancer recurrence and survivability prediction," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1811–1820, 2022, doi: 10.1016/j.csbj.2022.03.035.

[8] K. S. Pokkuluri, N. S. S. S. N. Usha Devi, and S. Mangalampalli, "DLCP: A Robust Deep Learning with Non-linear CA Mechanism for Lung Cancer Prediction," 2022, pp. 299–305. doi: 10.1007/978-981-16-8987-1_31.

[9] S. Shanthi and N. Rajkumar, "Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods," *Neural Process. Lett.*, vol. 53, no. 4, pp. 2617–2630, Aug. 2021, doi: 10.1007/s11063-020-10192-0.

[10] J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, and P. Boffetta, "Risk factors for lung cancer worldwide," *Eur. Respir. J.*, vol. 48, no. 3, pp. 889–902, Sep. 2016, doi: 10.1183/13993003.00359-2016.

[11] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man. Mach. Stud.*, vol. 27, no. 3, pp. 221–234, Sep. 1987, doi: 10.1016/S0020-7373(87)80053-6.

[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.

[14] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012. doi: 10.1002/9781118391686.

[15] M. A. Bhat, "Lung Cancer," *Kaggle*, 2021. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer (accessed Jan. 10, 2022).