SIMPLIFIED LUNG CANCER PREDICTING SYSTEM

MACHINE LEARNING 1 - REPORT
FINAL TEAM #9 PROJECT
AHN YEBIN - KIM DOHOON - WONG JULIE

1. Introduction

The lungs are the organs of respiration. The purpose of inspiration is to bring oxygen to all the body's cells, necessary for their proper functioning. Air breathed in through the nose and mouth travels through the trachea, bronchi and bronchioles to the alveoli in the lungs. The inspired oxygen then passes through the walls of the alveoli into the bloodstream, attaching itself to red blood cells, which transport it to the body's various cells. Exhalation, in the opposite direction, allows the carbon dioxide released by all the body's cells to be excreted into the bloodstream, where it passes through the alveoli into the air. Lung cancer is characterized by the uncontrolled multiplication of abnormal cells in lung tissue, particularly in the bronchi. The terms bronchial cancer or bronchopulmonary cancer are also used to designate lung cancer. There are non-small-cell lung cancer and small-cell lung cancer. Non-small-cell lung cancer usually originates in the glandular cells located in the outer part of the lung. This type of cancer is called adenocarcinoma. Non-small-cell lung cancer can also originate in the thin, flat cells known as squamous cells. This is known as squamous cell carcinoma of the lung. Large-cell carcinoma is another type of non-small-cell lung cancer, but is less common. There are also several rare types of non-small-cell lung cancer, including sarcoma and sarcomatoid carcinoma. Small-cell lung cancer usually originates in the cells lining the bronchial tubes in the center of the lungs. The main types of small-cell lung cancer are small-cell carcinoma and mixed small-cell carcinoma (a mixed tumor with squamous or glandular cells).

According to the National Cancer Information Center, lung cancer is one of the most common cancers worldwide. In terms of both incidence and mortality rates, it ranks among the top cancers. Annually, there are 2.09 million new cases (estimated by the World Health Organization). The incidence of lung cancer can vary across the world because of factors such as smoking prevalence, exposure to environmental pollutants, and genetic predisposition. That amount makes it the most commonly diagnosed cancer globally. According to SNU Bundang Hospital, the five-year survival rate of all lung cancer patients has not exceeded 15%. On the other hand, early-diagnosed lung cancer has a cure rate of more than 70%. Consequently, there is a critical need for preventive measures, effective treatment, and early detection of lung cancer. Indeed, the significant mortality rate can be due to a late-stage diagnosis, limited treatment options, and the aggressive nature of certain lung cancer subtypes.

Therefore, we suggest a simplified system to predict lung cancer, so that patients can recognize their disease and get in-depth tests at an early stage. Using machine learning to predict it would empower patients by significantly improving their chances of successful treatment.

2. Literature Reviews

- Global epidemiology of lung cancer [1]

The leading cause of cancer-related deaths globally is lung cancer. In most industrialized countries, the incidence of lung cancer peaked in the 1980s and has since declined. When it is about emerging ones, they have varying rates of lung cancer mortality. This can be due to a disparity among men and women or the place of living because for example in rural areas, it is more difficult to access care. Furthermore, the most common subtype of lung cancer is adenocarcinoma, then it is squamous cell and small cell lung cancers. The most common genetic markers found in adenocarcinoma are EGFR

and KRAS mutations. The ALK gene rearrangements predict response to specific treatments. Moreover, the major risk factor is smoking and the exposure to smoke, biomass fuels, pollution and genetic factors can increase the risk too. There is ongoing research to establish relationships between diet, nutrition and genetic factors and cancer risks. Currently, the most recommended for people with high-risk of developing it is screening methods like low-dose chest tomography. Finally, even with advancements in tumor biology, mortality rates are still worldwidely high.

- Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening [2]

This paper is about the National Lung Screening Trial (NLST) which determines if low-dose computed tomography (CT) screening could reduce mortality from lung cancer. There were 53,454 high-risk individuals who underwent three rounds of tests. 26,722 participants went for low-dose CT and 26,732 for single-view posteroanterior chest radiography. The results showed that the adherence rate to screening was high. However, a significant number of positives were false positives. Lung cancer is a deadly disease and long-term survival rates are low. The screening has small adverse events. After three rounds of screening, it resulted in a 20.0% reduction in lung-cancer mortality and a reduction in overall mortality compared to radiography. However, false positive results and complications from invasive diagnostic evaluations were concerning. Finally, several key points were emphasized: adherence to screening protocols and accurate identification of lung cancers and deaths were crucial. And there is the necessity to do further analysis regarding potential harmful effects and cost-effectiveness of screening. Then, molecular markers were identified as potential ways for gaining additional insights into low-dose CT screening.

- Focus on lung cancer [3]

The long term key to control lung cancer is the prevention of smoking initiation and helping people to stop it. The treatment to block the progression of this cancer has been fraught with difficulty. There may be a way thanks to vitamin E and selenium. Moreover, there are some methods to prevent smoking initiation and to help people with the development of new imaging, molecular genetics, and genetic epidemiological methods for early detection and the chemoprevention of lung cancer.

- Small-cell lung cancer [4]

Small-cell lung cancer (SCLC) comes from epithelial cells, and it can be diagnosed in elderly heavy smokers. This type of lung cancer has an aggressive nature, an early dissemination, high initial response rates to chemotherapy, and paraneoplastic syndromes can be perceived. First of all, these patients need to stop smoking. The primary treatment for extensive-stage SCLC is usually a combination of chemotherapy, with etoposide and a platinum salt being the standard first-line regimen. And people with limited-stage SCLC needs surgery, adjuvant chemotherapy, and chest irradiation. Prophylactic cranial irradiation is also recommended to reduce the risk of brain metastases. Other therapies and biological targets are being investigated such as growth factor inhibitors, hedgehog signaling pathway inhibitors, and mitochondrial apoptosis pathway targeting. Moreover, there is also another treatment which has potential, it is the individualization of therapy using novel agents. However, further research is needed to improve outcomes, particularly for patients with relapsing or refractory SCLC because treatments have limited efficacy.

- Early lung cancer diagnostic biomarker discovery by machine learning methods [5]

Several machine learning models were applied and compared, and the results showed that Naïve Bayes, Random Forest, Neural Network, and SVM models with high sensitivity have dependable and stable potential for early lung tumor prediction.The top 5 relative importance metabolic biomarkers (taurine, Palmitoyl-l -carnitine, proline, PE (36:4) and 2-DG) were developed by FCBF algorithm and could be potential candidates for preclinical screening of lung cancer.

- A Review of most Recent Lung Cancer Detection Techniques using Machine Learning [6]

In this paper, there are different techniques to detect lung cancer thanks to machine learning. CT scan images are the most accurate results, with marker-controlled watershed segmentation rather than other segmentation techniques. Small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC) are the two main kinds of lung cancer. What causes cancer is a nodule tumor from the cells in the airways of the respiratory system. Machine learning can help for an early diagnosis and evaluation of lung nodules. Indeed, artificial intelligence can build analysis of CT images. Researchers used machine learning techniques (SVM multi-classifiers and deep learning) and sensor arrays based tests for the detection of lung cancer. High accuracy was achieved thanks to models such as artificial neural networks and convolutional neural networks. Moreover, the complete fusion method (with deep-seated neural networks) assists radiologists in decision-making for chest x-ray radiographs. In the pre-processing phase, there are de-noising, thresholding, binarization, normalization and zero centering, followed by segmentation and the features can be extracted for classification.

- Detecting Lung Cancer Using Machine Learning Techniques. [7]

Lung Cancer is one of the dangerous diseases that increases mortality, and researchers have found that early detection can increase patient survival. The process of identifying mutations or cancer hotspots is complex and difficult, but ML technology (random forest and CNN) can effectively predict the outcome of cancer types by recognizing and detecting patterns in complex datasets.

3. Data description

There are many limitations to using clinical data in a personal environment. As clinical data contains patients' sensitive information, accessing data is restricted. Because of the limitations, we used clinical synthesized data from the National Cancer Data Center[8]. It is synthetic data generated by an AI-based model algorithm based on clinical data. It contains basic information, health information, family history, body measurement, diagnostic information, diagnostic test, imaging test, bronchoscopy, lung function test, biopsy, immune pathology, molecular pathology, surgical information, anticancer drug treatment, radiation therapy, and death information. Cancer stage is represented by the TNM staging system. Here, each prediction was aimed at these; whether they have cancer. If it is, whether they have small or non-small cell carcinoma. Because it is hard to classify their stages only with our data, we considered that each data has non-small cell lung cancer if they have one among adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Likewise, each data has small cell lung cancer if they do not have non-small cell carcinoma and they have T-staged cancer. 1 for true and 0 for false.
To predict whether they have cancer or not, we used 4 models; random forest, gradient boosting, ada boost, and MLP. Because of the imbalance between the cancer and non-cancer data, we used the oversampling method SMOTE. Without oversampling, the ratio of data and the performance were almost the same, while AUC was around 0.5. 5-fold cross validation was used to evaluate models. As criterions, accuracy, AUC, and f1 score were used.
Originally, it had 15000 rows and 34 columns.

[Description of all columns]

	Description	Type	Value	Null
No	Number of index	int64	max: 14999 / min: 0	15000 non-null

Age	age at diagnosis	int64	max: 88 / min: 31	15000 non-null
Adenocarcinoma	Histological Diagnostic Name	int64	0: false / 1: true	15000 non-null
Large cell carcinoma	Histological Diagnostic Name	int64	0: false / 1: true	15000 non-null
Squamous cell carcinoma	Histological Diagnostic Name	int64	0: false / 1: true	15000 non-null
TX, T0, T1, T1a, T1b, T1c, T2, T2a, T2b, T3, T4, N1, N2, N3, M1a, M1b, M1c	Stage	int64	0: false / 1: true	15000 non-null
Type of Drink	Type of Drinking	int64	1: beer / 2: soju / 3: liquor / 99: etc	15000 non-null
Smoke	Smoking Status	int64	0: not / 1: current / 2: past	15000 non-null
Height	Height	float64	max: 188.3 / min: 142.0	15000 non-null
Weight	Weight	float64	max: 105.1 / min: 34.3	15000 non-null
FEV1_FVC_P	FEV test result	int64	max: 99 / min: 31	15000 non-null
DLCO_VA_P	DLCO test result	int64	max: 161 / min: 27	15000 non-null
EGFR mutation Detection	Whether EGFR mutation is found	int64	max: 99 / min: 0	15000 non-null
Operation	Surgery or not	int64	0: false / 1: true	15000 non-null
Chemotherapy	Chemotherapy	int64	0: false / 1: true	15000 non-null
Radiation Therapy	Radiation Therapy	int64	0: false / 1: true	15000 non-null
Death	Death	int64	0: false / 1: true	15000 non-null
Survival period	Survival period	int64	max: 730 / min: 16	15000 non-null

Before proceeding with EDA, unnecessary columns (‘No’, 'TX', 'T0', 'N1', 'N2', 'N3', 'M1a', 'M1b', 'M1c') were removed in the preprocessing step. TX means unable to evaluate, and T0 means no evidence. N and M mean metastasis, but it is hard to classify only with lung function tests. So, to predict whether they have lung cancer or not, we removed them[9]. There was no null value and no value to be judged as an outlier in the data, so any operations were not performed other than checking the information (data type, max, mean, etc.). However, several columns that were deemed necessary in the EDA stage were added.

[Description of the added column]

	Description	Type	Value	Null
Cancer	Cancer type 0 (false) when 0 for all diagnostic criteria*, 1 (true) when any one is 1	int64	0: false(non-cancer) / 1: true(cancer)	15000 non-null
Drink_mapped	Specify what kind of alcohol it is by mapping it into an object type	object	Beer, Soju, Liquors, Others	15000 non-null
Smoke_mapped	Specify which smoking type corresponds to by mapping it to the object type.	object	Not, Current, Past	15000 non-null
Smoke_re_mapped	Not divided into not/current/past, but only whether it is smoked or not.	object	Not, Smoke	15000 non-null
BMI	Calculate BMI using Height and Weight	float64	max: 51.32 / min: 10.70	15000 non-null
BMI_mapped	Categorized according to the calculation method**	int64	0: Underweight / 1 : Normal / 2: Overweight + Obese	15000 non-null
Smoke_re_int_mapped	Replace the 'Smoke_re_mapped' column with int form and mark 1 (true) for smoke and 0 (false) for Not	int64	0: false(Not) / 1: true(Smoke)	15000 non-null
FEV_filter	0 (false) and 1 (true) indicate whether the value of 'FEV1_FVC_P' is less than 70	int64	0: false / 1: true(below 70)	15000 non-null
DLCO_filter	0 (false) and 1 (true) indicate whether the value of DLCO_VA_P is less than 80	int64	0: false / 1: true(below 80)	15000 non-null

* 'Adenocarcinoma', 'Large cell carcinoma', 'Squamous cell carcinoma', 'T1', 'T1a', 'T1b', 'T1c', 'T2', 'T2a', 'T2b', 'T3', 'T4'
** provided by the Korea Centers for Disease Control and Prevention

Also, for the second EDA, one column that was not included for the first EDA was added. To be precise, the 'Cancer' column changed because the cancer classification was different. Previously, when all diagnostic criteria were 0, it was determined as 0 (false, non-cancer). However, this time, through "Adenocarcinoma," "Large cell carcinoma," and "Squamous cell carcinoma," cancer was also divided into cancers of small cells and cases that were not. The three diagnoses above are not small-sized cancers. Therefore, if one of the three is 1, it is classified as 2 (non-small-sized cancel), if all three are 0, and if there is even one of the remaining diagnostic criteria*, it is classified as 1 (small-sized cancel), and if it is 0 in all cases, it is classified as non-cancer. And other columns were the same for the first EDA.

	Description	Type	Value	Null
Cancer	Cancer type	int64	0: non-cancer / 1: small-sized cancer / 2: non-small sized cancer	15000 non-null

* 'T1', 'T1a', 'T1b', 'T1c', 'T2', 'T2a', 'T2b', 'T3', 'T4'

4. Analysis

The first EDA focused on cancer and non-cancerous cases and created EDAs. The hypotheses are 1) the older the person, the higher the probability of lung cancer, 2) the more smoking, the higher the probability of lung cancer, 3) the higher the BMI, the higher the probability of lung cancer, and 4) treatment will increase the survival rate.

As a result of drawing the age distribution of cancer patients as a histogram, the proportion of the older group was high(figure 1). Therefore, for hypothesis 1, it can be judged that the higher the age, the higher the incidence of cancer. However, further verification and analysis are needed to confirm the hypothesis.

The correlation between drink and cancer was tried to be examined, but it could not be confirmed because the drink data was represented by the type of drink, not by the presence or intake of the drink. So, only what type of alcohol people with cancer enjoyed a lot could be found out, and as a result, soju and others were high (figure 2).

To find out the relationship between smoking and the onset of cancer, a bar plot of accumulated values was drawn (figure 3). People with cancer were found to be past smoking > non-smoking > current smoking. The difference in data size between those with cancer and those without cancer made it difficult to determine the results of those without cancer, so more bar plots were created with only data from those without cancer (figure 4). As a result, like people with cancer, past smoking > non-smoking > current smoking resulted in results. What this shows is that smoking and the onset of cancer are not related. Even if someone quit smoking six months ago, it can be recorded as smoking in the past. Also, according to an Oxford academic journal article, smoking in the past can also affect his or her lungs, it is classified as not/current/past, but not/smoke. So even if someone has smoked in the past, it is included as smoking. Even when the classification was changed, a cumulative form of bar plot was drawn (figure 5), and another bar plot was created to determine the distribution of people who did not have cancer (figure 6). As a result, the proportion of smoke was high in both people with and without cancer. In conclusion, in Hypothesis 2, it can be said that smoking is not related to the occurrence of cancer.

To find out the relationship between BMI(weight(kg)/height(m)^2) and the onset of cancer, BMI histograms and boxplots were drawn according to the presence or absence of cancer. However, since it was difficult to confirm whether it was related or not only with histograms and boxplots, a qui-square test was conducted(figure 7,8,9). The qui-square test is a method of testing relevance (independence) through the difference between expected and observed values. In this test, the null hypothesis is 'no link between BMI and the onset of cancer'. As a result of the test, since p is < .05, the null hypothesis can be rejected and it can be judged that there is a relationship between BMI and lung cancer(figure 10). In addition, since the result of the independent t-test is $p < .05$, it can be said that there is a significant difference between the BMI of the cancer group and the BMI of the non-cancer group(figure 11). Therefore, there is a significant relationship between BMI and lung cancer in Hypothesis 3, but further analysis is needed to confirm that the hypothesis is true.

The purpose of this analysis was to investigate the relationship between lung function tests and cancer incidence. The first is the FEV test. In this data, the FEV test has a value representing the FEV1/FVC ratio, and the FEV1/FVC is an indicator of bronchial obstruction and is in the normal range of 70% to 80%. If it is less than 70%, it is judged that there is a problem with breathing. When the box plot and KDE plot were drawn using the entire data, rather, the Q2 (median) and the densest point in non-cancer patients were rather lower than in cancer patients(figure 12). The figure is over 70%, but there is confusion in determining the relevance. Therefore, it was decided to check separately only for less than 70%. In this case, the test results of people with cancer were distributed lower in the box plot. Q1 (lower quartile), Q2 (median), and Q3 (upper quartile) all had low values in cancer patients(figure 13). In conclusion, if the FEV is less than 70, the lower the value from 70, the higher the likelihood of cancer

patients, but it is difficult to determine whether there is a relationship between cancer and FEV testing. The following is the DLCO test. DLCO is an indicator of how efficiently gas exchange is performed in the lungs, and 80 to 120 is the normal range. If the DLCO is less than 80, lung function is judged to be impaired. When drawing box plots and KDE plots using the entire data, Q2 (medium) and the highest density showed similar values in the two groups(figure 14). In addition, only the case of less than 80, such as FEV, was extracted separately and a box plot was drawn(figure 15). In this case, data from people with cancer were distributed low. In conclusion, if it is less than 80 in DLCO, the lower the value, the more likely it is a cancer patient, but it is difficult to determine whether there is a relationship between DLCO and the onset of cancer.

The association between death and each treatment was investigated. There are three treatments: Operation, Chemotherapy, and Radiation Therapy. Since the treatment and Death data are represented by 0 (false) or 1 (true), a heat map was drawn. And then conditional probabilities were used to determine whether each treatment increased the survival rate. First is the case of operation(figure 16). The probability of survival after operation was 79.2%, and the probability of survival without operation was 78.6%, which was slightly higher in the case of surgery. Next is the case of chemotherapy. The probability of survival after receiving chemotherapy was 78.5%, and the probability of surviving without chemotherapy was 79.5%, which was slightly higher when not receiving chemotherapy. Finally, it is the case of radiation therapy(figure 18). The probability of survival after receiving radiation therapy was 88.1%, and the probability of survival without radiation therapy was 79.1%, and the survival rate was higher when receiving radiation therapy. In addition, the survival probability curve was drawn using the survival time of the deceased(figure 19). Due to the narrow range of opaque areas, uncertainty is low, and over time, it is possible to see what the survival rate is.

The second EDA distinguished cancer patients in more detail, indicating whether they have small-sized cancer or not. First of all, I will summarize the features that have similar conclusions as the first EDA. First, age and cancer may be related because both non-small-sized and small-sized cancers had more people with cancer as they were older. Second, all three groups (non-cancer, small-sized cancer, and non-small-sized cancer) were high in the order of past smoking > non-smoking > current smoking. In addition, as in the first EDA, both past and current smoking were classified as smoking and divided into smoking or non-smoking. All three groups had a high rate of smoking, making it difficult to confirm the significance of smoking and cancer. Third, both the FEV test and the DLCO test were difficult to confirm the test results and the significance of the cancer when considering the entire data. However, considering the data below 70 and 80, which are the criteria for lung function disorders, respectively, the distribution of cancer cases was lower than that of non-cancer cases. However, in both tests, the distribution when it was a non-small-sized cancer was not lower than when it was a small-sized cancer.

It explains the part that appears differently from the first EDA. First, it's BMI. The first EDA confirmed the significance between BMI and cancer incidence. Also in this EDA, it was difficult to confirm the relationship only with the box plot and the violin plot, so chi-square test and independent t-test were conducted (figure 20, 21). Since the result of the chi-square test is $p > .05$, it cannot be ignored that there is no relationship between cancer types(non-cancer, small-sized cancer, non-small-sized cancer) and BMI (figure 22). On the other hand, the independent t-test shows that there is a statistically significant difference since p is lower than .05 (figure 23). In conclusion, it means that there are differences between groups, but not related. Therefore, when combined with the first EDA, it can be seen that BMI is associated with cancer and non-cancer classification, but not with cancer size.

It was also examined whether the treatment method increased the survival rate by dividing it into a small-sized cancer and a non-small-sized cancer. After drawing the heatmap, the survival rate when treatment was received and when treatment was not received was calculated for each case using the conditional probability. In the case of small-sized cancers, the probability of survival when undergoing operation was 79.7%, and the probability of survival without operation was 79.4% (figure 23). When

chemotherapy was received, the probability of survival was 77.3%, and the probability of survival without chemotherapy was 81.5% (figure 24). In addition, when receiving radiation therapy, the probability of survival was 77.9%, and the probability of survival without radiation therapy was 80.1% (figure 25). Therefore, even though there was only a small difference in the small-sized cancer, it cannot be said that it helped increase the survival rate other than surgery. In the case of non-small-sized cancers, the probability of survival when undergoing operation was 79.0%, and the probability of survival without operation was 78.2% (figure 26). When chemotherapy was received, the probability of survival was 78.9%, and the probability of survival without chemotherapy was 78.6%. (figure 27). In addition, when receiving radiation therapy, the probability of survival was 78.9%, and the probability of survival without radiation therapy was 78.6% (figure 28). For non-small-sized cancer, all treatment methods were effective in increasing the survival rate even a little.

Finally, the survival curves of the non-small-sized canceler and the small-sized canceler were drawn. Previously, it was hypothesized that the probability of survival would be higher in the case of a small-sized cancer than in the case of a non-small-sized cancer. However, in Kaplan-Meier survival analysis, the two curves almost overlapped (figure 29). Since this means that the survival rate is similar without a significant difference, the difference in survival rate was statistically evaluated through the log-rank test. The log-rank test is a statistical test method to see if there is a difference in survival rates between groups by comparing survival time data. Since the test result was $p > 0.05$, the null hypothesis that there is no significant difference in survival rates between the two groups cannot be rejected (figure 30). Therefore, it can be concluded that the size of the cancer is not related to the survival rate. We separated our problem into two steps; whether they have cancer or not, and whether their cancer is non-small cell or small cell lung cancer.

Among them, random forest got the best performance (figure 31). accuracy 0.95, AUC 0.98, f1 score 0.95. ‘lbfgs’ solver in MLP didn’t converge in 1000 iterations, but it took too much time. Because we cannot use GPU at scikit-learn, we cannot update the performance of MLPClassifier if other data is collected. Because of that, we selected a random forest classifier. It is okay to use them, because data is big enough, and we used cross-validation.

Secondly, in terms of size, also random forest got the best performance(figure 32). But unlike before, model performance decreased a little. This is because the changes from lung cancer are not revealed well, only in terms of lung function. Because it is hard to distinguish and even cannot distinguish their level, it is recommended to do more detailed tests like biopsy.

Also it is revealed that size of cancer itself and treatments themselves are not important features in terms of death(figure 33). It means that only doing operation, chemotherapy, or radiation therapy does not imply survival. Rather than themselves, we need to check the changes of patients caused by them.

5. Results

Regardless of the stage, we only considered whether they have or not. After that, we considered whether it is non-small cell lung cancer or small cell lung cancer. It is okay to use our model, because our model is aimed at daily-life steps. It only tells us how dangerous we are, and does not tell us the details. And after collecting more clinical data, our model can be more well-performing.

Firstly we considered predicting whether non-small cell carcinoma and small cell carcinoma or not. Gradient boosting performed 0.97347 +/- 0.00 accuracy and 0.519 +/- 0.06 AUC, and adaboost performed 0.97493 +/- 0.00 accuracy and 0.528 +/- 0.091 AUC at cross validation. In both models, BMI, DLCO_VA_P, age, and FEV1_FVC_P were the most important features.

Second we considered predicting whether non-small cell carcinoma or not. Gradient boosting performed 0.825 +/- 0.00 accuracy and 0.479 +/- 0.02 AUC, and adaboost performed 0.8256 +/- 0.00 accuracy and 0.48 +/- 0.022 AUC at cross validation.

Third we considered predicting whether small cell carcinoma or not. Gradient boosting performed 0.857 +/- 0.00 accuracy and 0.567 +/- 0.118 AUC, and adaboost performed 0.8578 +/- 0.00 accuracy and 0.574 +/- 0.1451 AUC at cross validation.

While BMI, FEV1_FVC_P, DLCO_VA_P, and age were the most important features in determining lung cancer, smoking did not play a significant role. This means that smoking does not directly affect lung cancer, but that a decrease in lung function caused by smoking may increase the incidence of lung cancer. It is not possible to accurately predict whether or not to smoke, and it is more reasonable to consider the period and environment that may affect lung function decline.

6. Conclusion

Our machine learning project will have significant implications for lung cancer diagnosis and treatment. Indeed, it provides an accurate prediction of lung cancer risk and our simplified system will enable healthcare professionals to identify high-risk individuals at an early stage.One of the best solutions to reduce mortality rate of lung cancer is to aware patients from an early stage of cancer for following treatments in order to cure it and so improve patient outcomes. Moreover, our project also contributes to the machine learning field and paves the way for future advancements in predictive analytics for cancer detection. After this team project, we believe that machine learning has the potential to make an important impact on lung cancer detection and prognosis. To build our system, we saw that the random forest model has the best accuracy, AUC and F1-score. Moreover, the more important features selected for random forest are almost the same as in EDA. The effects of smoking damage the lungs. However, smoking itself does not really have an impact on the development of lung cancer. There is a correlation between the decrease in lung function and smoking.

7. Future works

In the future, it will be better to have more accurate data because ours were made from AI generators based on some clinical data. However that means, we are proud of ourselves that we work on this subject ethically thanks to the anonymity of the data. We will need to find the best balance between accuracy and personal information. Then, we could suggest the best combination of treatments to cure lung cancer when it is possible (early stage). Moreover, we need to improve the features selected such as a frequency of smoking or drink because we only had the type of drink and the smoking feature without it frequency was a bit useless. Finally, thanks to our system, we could help people to get precision tests earlier if there are methods to test FEV, DLCO simply like corona kits. Indeed, these systems are only available in hospitals and so not available for everybody..

[Appendix]

https://github.com/kdh-yu/ML1-Project
This is our github to let you see all our codes.

[Reference]

1. Barta, J.A.; Powell, C.A.; Wisnivesky, J.P. Global epidemiology of lung cancer. *Ann. Glob. Health* 2019, *85*, 8. [Google Scholar]
2. National Lung Screening Trial Research Team , Aberle DR, Adams AM, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* 2011;365:395-409. 10.1056/NEJMoa1102873 [Google Scholar]
3. JD Minna, JA Roth, AF Gazdar. *Focus on lung cancer* - Cancer cell, 2002 [cell.com]
4. JP Van Meerbeeck, DA Fennell, DKM De Ruysscher. Small-cell lung cancer - *The Lancet,* 2011 - Elsevier [PDF]
5. Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, Qi-Biao Wu, Pei-Yu Yan, Liang Liu, Yi-Jun Tang, Xiao-Jun Yao, Mei-Fang Wang, Elaine Lai-Han Leung. Early lung cancer diagnostic biomarker discovery by machine learning methods *Translational oncology 14 (1), 100907, 2021* [ScienceDirect]
6. Dakhaz Mustafa Abdullah, Nawzat Sadiq Ahmed. A Review of most Recent Lung Cancer Detection Techniques using Machine Learning *International Journal of Science and Business 5 (3), 159-173, 2021* [Ideas]
7. Kumar Dutta, Ashit. 2022. “Detecting Lung Cancer Using Machine Learning Techniques.” *Intelligent Automation & Soft Computing* 31 (2): 1007–23. https://doi.org/10.32604/iasc.2022.019778.
8. Korea Disease Control and Prevention Agency National Health Information Report (no date) *질병관리청-국가건강정보포털*, *질병관리청 국가건강정보포털-비만*. Available at: https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=5292 (Accessed: 20 June 2023).
9. Williamson, T.J. *et al.* (2020) ‘Lung cancer stigma: Does smoking history matter?’, *Annals of Behavioral Medicine*, 54(7), pp. 535–540. doi:10.1093/abm/kaz063.
10. National Cancer Institute. Staging. Retrieved October 14, 2022, from https://www.cancer.gov/about-cancer/diagnosis-staging/staging