A RELATIVE DATA ANALYSIS ON LUNGE CANCER PREDICTION IN BANGLADESH (August, 2023)

Anik Paul Shuvo[1], Md. Mizanur Rahman[2], and Fazla Rabbi[3], Khan Raqib Mahmud[4], IEEE
[1-3]University of Liberal Arts Bangladesh, Dhaka-1207, Bangladesh
[4] Shah Jalal University of Science and Technology, Sylhet-3114, Bangladesh

ABSTRACT The most common cancer-related cause of death worldwide is lung cancer. Recent studies, primarily from the developed world, have linked demographic differences with treatment outcomes and survivathe coml. It is important to examine the relationship in the low- and middle-income population as well, given the variation in demographic characteristics across economic strata. The current study aims to evaluate the relationship between demographic traits and lung cancer survival in Bangladeshi patients. Early cancer detection can aid in the complete eradication of the disease. So, the requirement of techniques to detect the occurrence of cancer nodules in the early stage is increasing. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal endings. Data mining is an effective method for assisting people in their scientific and engineering endeavors and overall wellness. Those techniques are extracting the hidden information from the large databases which helps to find the relationships and patterns from the data. The data set can be used within the data mining and exploratory data analysis process.

Keywords Bangladesh, Cancer, developing country development, Cancer Detection, Data Analysis, Lungs Cancer, Prediction, Prevention, Result, Symptoms

I. INTRODUCTION

Bangladesh is a developing nation with numerous problems, particularly in the health sector. Cancer is the most dangerous disease over the years. Cure to it is most likely calling for more suffering. Cancer management is a priority due to the current trend of increased incidence in this region.Lung Cancer is one of the cancers in which many people are affected by the time. Every year lung cancer kills more people than any other cancer. A country like Bangladesh where about 35% of the total population is smokers is not still aware of lung cancer fully. In addition, it is important for people to know more about lung cancer. Not only that the capital of Bangladesh, Dhaka city is so populated and might be one of the reasons for lung cancer in Bangladesh. It is sad that the government has not taken the initiative to educate people about lung cancer. About 17.8% of people affected by lung cancer die every year. In an underdeveloped country like Bangladesh people are not fully aware of the symptoms of lung cancer and they are not aware of the possible reasons for lung cancer. The reason behind choosing this topic is, Bangladesh is a small country with a large population and many problems. The problems are not being taken care of due to ignorance and lung cancer is being ignored too. Many lives could have been saved if people were well aware of this cancer. Many people would not suffer from this cancer if they knew the precautions. The goal of this research is to make people aware of lung cancer, make people aware of the symptoms and precautions of lung cancer. In the future this research may help people to have a good overall understanding of lung cancer. Not only that, they can be careful and take precautions. A small step at a time can make big changes.

In one of the articles, the current scenario of cancer in Bangladesh and its management with brief history is outlined. The gradual improvement in cancer management highlights the joint effort of the public and commercial sectors. Recent introduction of the state-of-the-art facilities and the training facilities for human resource development are also outlined. The existing challenges and cooperation from local NGOs and other overseas sources are also highlighted to provide an insight regarding possible ways to tackle these challenges to ensure a better future. We are directing an expectation of having cellular breakdown in the lungs thinking about the accompanying elements: age, orientation, smoking admission, liquor utilization, and furthermore in the event that an individual has side effects connected with cellular breakdown in the lungs. In addition, different elements like how long the side effects of cellular breakdown in the lungs persevering in an individual will likewise be thought about. The viability of this task is that it can assist individuals with realizing their disease risk for a minimal price and furthermore it empowers them to take the suitable choice in light of their malignant growth risk status. The information is gathered from the site on the web. The consequence of the blood tests from smokers is gathered from a web-based site which is utilized as a screening device to foresee the event of cellular breakdown in the lungs in smokers; it very well may be analyzed at an early age. Cellular breakdown in the lungs is the one of the main sources of disease passings. A new study suggests that cases of lung cancer have been on the rise in Bangladesh, with the number of smokers and air pollution levels rising. The burden of lung cancer in the nation has reportedly increased by about 200% in just three years, according to the most recent Hospital Cancer Registry Study. The report also claimed that from January 2015 to December 2017, a

total of 76,543 new patients attended the outpatient department of the National Institute of Cancer Research and Hospital (NICRH). Of them, 35,369 patients were confirmed or had provisional diagnosis of cancer to be included in the final analysis. A total of 5,887 people with lung cancer were admitted to the hospital in these three years. The figure was 1983 in 2014, as per the report, indicating a nearly 200% rise in cases in just three years. Dr Md Habibullah Talukder, head of Cancer Epidemiology Department at NICRH, said: "Smoke from factories and exhaust from vehicles are the two leading causes of cancer in Bangladesh. Immediate steps are needed to reduce pollution" [1]. Symptoms that may suggest lung cancer include Chronic coughing or change in regular coughing pattern, Dyspnea (shortness of breath with activity),

- Wheezing,
- Hemoptysis (coughing up blood),
- Chest or abdomen pain,
- Cachexia
- Clubbing of the fingernails,
- Dysphasia (difficulty swallowing),
- Pain in shoulder, chest, arm,
- Dysphonia (hoarse voice),
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

II. LITERATURE REVIEW

Mortality and morbidity due to tobacco use is very high. The paper "Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes" is about evaluation of symptoms, signs of lung cancer. The authors of the paper discuss the symptoms of lung cancer in evaluating time to time. They also discuss some important factors to identify lung cancer. One of the most important factors is Chest Radiography. According to the authors, Chest Radiography is the first thing to consider if someone is suspected to have lung cancer. "More than 90% of patients with lung cancer will be symptomatic at presentation" (Michael et al., 2003)[2].

The paper "Symptoms in Adults with Lung Cancer: A Systematic Research Review" written by Mary E Cooley discusses the symptoms of lung cancer, mostly the symptoms of lung cancer in adults. In addition, this study was conducted in the USA and Canada. Health is something which needs to be maintained. Whenever a healthy lifestyle is disrupted for some reason. Health does not cooperate like before as a result many symptoms are seen. It is important for all you consider those symptoms. Otherwise it will be too late to recover from that disease or health issues. In most of the countries symptoms or certain health problems are ignored by people. The reason behind it is people consider it as "normal". The author, Mary E Cooley clearly states that for a particular group of people symptoms management is important. It is so shocking that in the United States lung cancer is the second most common cancer and it is in both men and women. "Twenty-five percent of all cancer deaths are due to lung cancer" (Cooley, 2000). Many results about lung cancer show that patients with advanced level lung cancer show the highest number of symptoms. Lung cancer not only affects physical health but also mental health. It causes mental distress. The

a. Fever
b. Hoarseness of voice

xii. Loss of appetite
xiii. Puffiness of face
xiv. Nausea and vomiting

C. Lung cancer risk factors:

1. Smoking:
   i. Beedi
   ii. Cigaratte
   iii. Hukka
   iv. Marijuana
2. Second-hand smoke
3. Radon exposure
4. High dose of ionizing radiation
5. Occupational exposure to mustard gas, chloro methyl ether, inorganic arsenic, chromium, nickel, radon asbestos
6. Air pollution.

symptoms of lung cancer patients vary from month to month. According to Mary E Cooley the most common symptoms in newly diagnosed patients of Lung Cancer are fatigue, pain, loss of appetite, coughing, and insomnia. There is debate about the statement "Men and Women have different symptoms". However, it is not proven. This research was conducted almost 23 years ago and it is not conducted in all countries of the world. All the results may vary[3].

The paper "Symptoms of lung cancer" gives information about symptoms of lung cancer. It is one of the most common cancers. Most of the data of this paper is collected by observing 100 lung cancer patients. "The most common and severe symptoms were pain (86), dyspnoea (70) and anorexia (68)" (Kretch et al., 1992). There are also other symptoms described in both man and woman. This paper suggests there is effective care for lung cancer. This paper also suggests changes in symptoms according to different ages of the patients[4]. The paper "A cluster of symptoms over time in patients with lung cancer" states if the symptoms of lung cancer changes over time. Data of this paper was taken by observing 112 lung cancer patients. "Death 6 to 19 months after diagnosis was predicted by age, stage of cancer at diagnosis, and symptom severity at 6 months" (Gift, 2003)[5].

The paper "Symptoms and the early diagnosis of lung cancer" discusses symptoms of lung cancer as well as its influence in the Western world. The paper lacks the information of the Eastern World lung cancer report. "Detection of the tumor at an earlier stage leads to an improved prognosis, patients presenting with stage IA non-small cell lung cancer and undergoing surgical resection having a 5 year survival of around 60%" (Birring et al., 2005). According to the authors of the paper, there are many symptoms of lung cancer. This paper also reviewed other papers where relevant information is given. This

paper found smoking might be one of the most common reasons behind lung cancer (Birring,2005)[6]. The paper "Lung cancer and exposure to arsenic in rural Bangladesh" not only talks about lung cancer in Bangladesh but also talks about arsenic. According to the authors of this paper, most lung cancer patients in Bangladesh have a bad habit of smoking. "Data were obtained from 7286 subjects who underwent lung biopsy in 2003–2006 at a diagnostic center taking referrals from throughout Bangladesh" (Mostafa et al., 2008). Not only male smokers are suffering from lung cancer but also female smokers are suffering. The problem is this research was conducted in rural areas of Bangladesh somewhat lacking the information about city people affected from lung cancer (Mostafa, 2008)[7].

In SEER-NPCR 2010 to 2017, there were over 1.5 million new lung cancer cases; of these, approximately 1.28 million were NSCLC (51% male; 70% aged ≥65 years). In SEER-18 over the same period, demographic distributions were similar in the NSCLC population (53% male; 68% aged ≥65 years; median [Q1-Q3] age at diagnosis was 70 [63-77] years) (Ganti AK, Klein AB, Cotarla I, Seal B, Chou E, 2021)[8]. Young nonsmoking patients with lung cancer are 2-fold more likely to be female, advocating for broader sex-based screening criteria. Potential roles of genetic mutations, estrogen signaling, and infectious elements in sex-based differences in presentation, histology, prognosis, and treatment response are explored (Lillian L. Tsai, 2022)[9]. Globally, the ASR of lung cancer prevalence, incidence and YLDs in 2019 were 38.84/100,000 persons, 27.66/100,000 persons, and 6.62/100,000 persons, respectively. Over the past 30 years, the ASR of incidence (EAPC = -0.09) decreased, although that of prevalence (EAPC = 0.51) and YLDs (EAPC = 0.03) increased. The global prevalence counts were greater in males than females at all age groups and increased with age, peaking in the 65–69 age group for both sexes. The increase in incidence was mainly attributed to population aging. For YLDs, EAPC was negatively correlated with the human development index ($p = 0.0008$) and ASR ($p < 0.0001$) in 1990 across nation-level units (Chen, X., Mo, S. & Yi, B, 2022) [10]. Although the terms "sex" and "gender" have been historically interchangeable in medical research, their uses are distinct as sex is conventionally based on anatomy and physiology, whereas gender typically refers to identity, behavior, or socially constructed roles. As such, research in potential lung cancer disparities has not disentangled sex versus gender. Nonetheless, the established differences in lung cancer incidence and mortality rates between males and females are attributed to historic patterns in tobacco smoking as noted above (Matthew B. Schabath, 2019) [11]. The use of tobacco cigarettes is the single greatest risk factor in the development of lung cancer, with up to 90% of lung cancers attributed to smoking. An understanding of this causal relationship developed only slowly and gradually, not least because of the decades-long latency period between smoking initiation and lung cancer occurrence. Prior to the 20th century, tobacco had been used for centuries without significant disease burden. In the pre-Columbian Americas, tobacco was used primarily for medicinal and ritual purposes. Tobacco was brought to Europe at the end of the 15th century and utilized in various forms including snuff, pipes and cigars. Cigarettes were, until the late 19th century, expensive, hand-rolled, and not considered acceptable in polite society or around women (de Groot PM, Wu CC, Carter BW, Munden RF, 2018) [12].

III. METHODOLOGY

Quantitative data has been collected from online website. We used Linear Regression, Logistic Regression, Naïve Baye's and KNN. All these methods were very helpful in deducting our expected outcome. Moreover, we applied Decision tree, Bar plot, Scatter plot, Box plot to have a better display our data analysis.

IV. DATA ANALYSIS

This project is about lung cancer prediction. To predict the risk of having lung cancer, the factors that are considered are: age, gender, air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease etc. Since in patientid column there were categorical values, these values are changed to numerical values. Then, the count of records in each column of the dataset are checked. Since the count of records for all the columns is 1000, this indicates that there are no blank columns in the dataset. Since all the predictor columns are not continuous in nature, there are no chance to have 0's in any column, which indicates that there are no missing data. After that, to detect outliers, boxplots are used. Since there is an outlier in 'age' section, to remove it, 5th and 95th percentiles are imputed.

V. DATA VISUALIZATION:

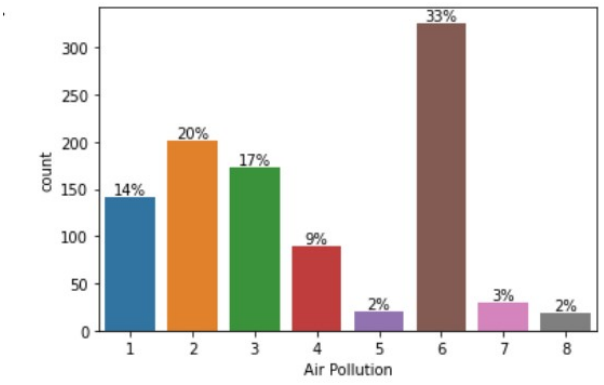


Fig:1 Air pollution

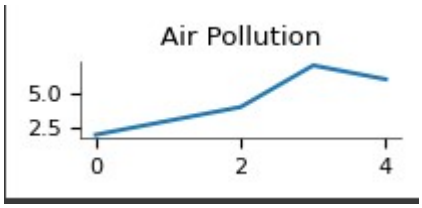


Fig:2

A counterplot of the factor "Air Pollution" is drawn in fig:1 and fig:2. It has been found out that most of the patients (i.e 33%) have intensity level '6' of air pollution and these patients have high risk of having lung cancer which means that the more the person is exposed to air pollution, that person has greater risk of having lung cancer.

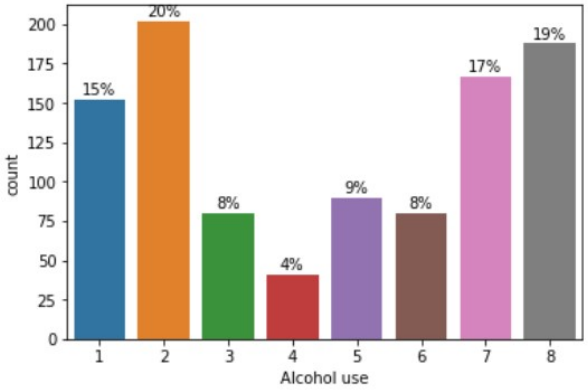


Fig:3 Alcohol

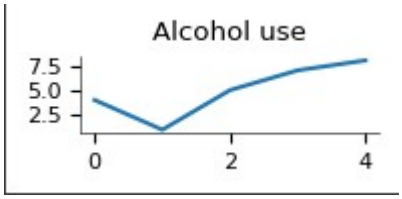


Fig:4

A counterplot of the factor "Alcohol use" is drawn Fig:3 and Fig:4. It has been found out that intake of alcohol does not cause lung cancer.

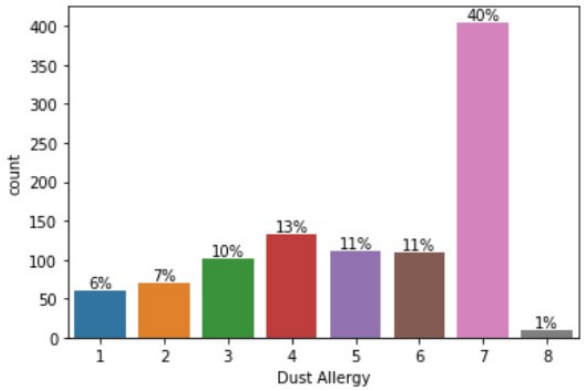


Fig:5 Dust Allergy

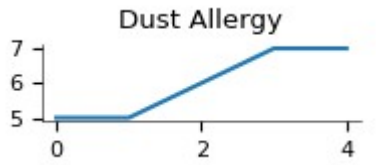


Fig:6

A counterplot of the factor "Dust Allergy" is drawn in Fig:5 and Fig:6 to find the percentage of the patients having a particular intensity level. It has been found out that most of the patients (i.e 40%) have intensity level "7" and they all have high risk of having lung cancer which that this factor plays a significant role in having lung cancer.

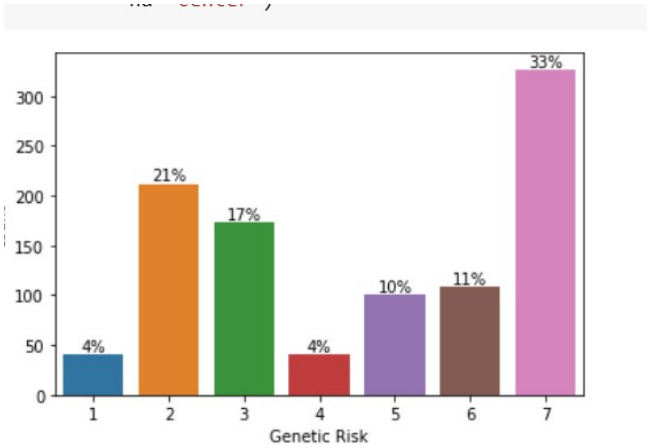


Fig:4 Genetic Risk

A countplot of the factor "Genetic risk" is drawn. It has been found out that if any family member of a person has lung cancer, then that person has greater genetic risk of having lung cancer.

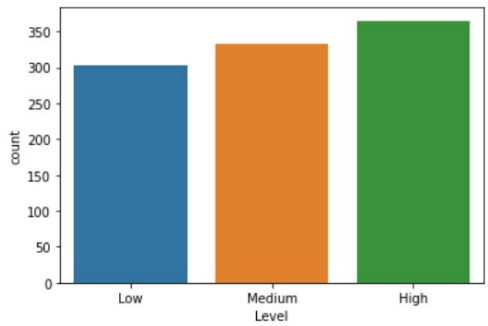


Fig:5

In the above plot is done to understand the distribution of patients having high risk, medium risk and low-level risk of having lung cancer in the data set. The plot infers that majority of the data consists of high-level risk patients.

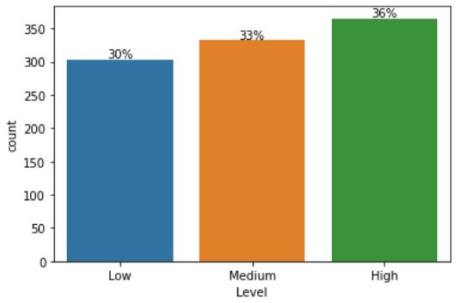
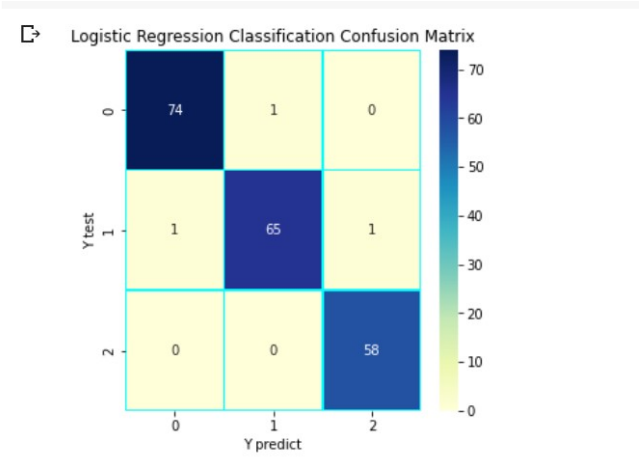


Fig:6

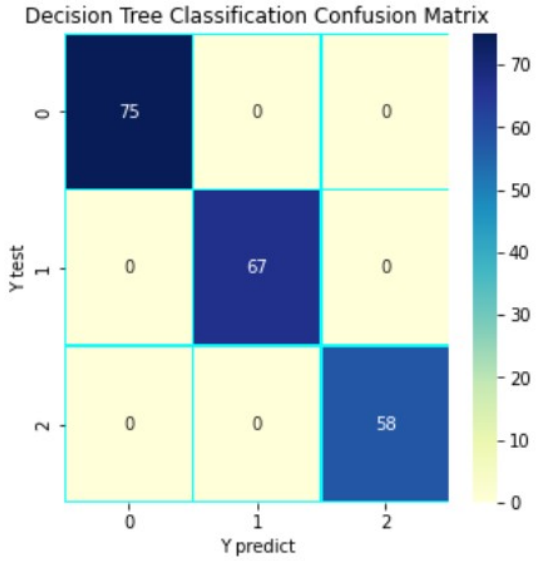
The above plot shows the percentage distribution of high-risk vs medium risk vs low risk lung cancer patients. About 36% of the data contains records belonging to those who have high risk of having lung cancer. 33% of those who have medium level risk and 30% of those who have low level risk of having lung cancer.

VI. PREDICTION MODEL:

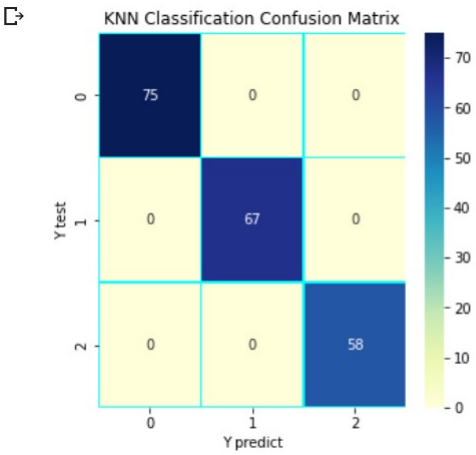


Logistic regression is a technique that can be applied to binary classification problems. This technique uses the logistic function or sigmoid function, which is an S-shaped curve that can assume any real value number and assign it to a value between 0 and 1, but never exactly in those limits. Thus, logistic regression models the probability of the default class (the probability that an input (X)(X) belongs to the default class (Y=1)(Y=1)) (P(X)=P(Y=1|X))(P(X)=P(Y=1|X)). In order to make the prediction of the probability, the logistic function is used, which allows us to obtain the log-odds or the probit. Thus, the model is a linear combination of the inputs, but that this linear combination relates to the log-odds of the default class.

Started from make an instance of the model setting the default values. Specify the inverse of the regularization strength in 10. Trained the logistic regression model with the training data, and then applied such model to the test data. The accuracy rate is 98%.



A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome. The decision tree analyzes a set of data to construct a set of rules or questions, which are used to predict a class, i.e., the goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In this sense the decision tree selects the best attribute using to divide the records, converting that attribute into a decision node and dividing the data set into smaller subsets, to finally start the construction of the tree repeating this process recursively. The accuracy rate is 100%



K-Nearest neighbors is a technique that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). This technique is non-parametric since there are no assumptions for the distribution of underlying data and it is lazy since it does not need any training data point model generation. All the

training data used in the test phase. This makes the training faster and the test phase slower and costlier. In this technique, the number of neighbor's k is usually an odd number if the number of classes is 2. For finding closest similar points, find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance. The accuracy rate is 100%.

The naive Bayesian classifier is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. Thus, using Bayes theorem (P(X|Y)=P(Y|X)P(X)P(Y))(P(X|Y)=P(Y|X)P(X)P(Y)), we can find the probability of XX happening, given that YY has occurred. Here, YY is the evidence and XX is the hypothesis. The assumption made here is that the presence of one particular feature does not affect the other (the predictors/features are independent). Hence it is called naive. In this case we will assume that we assume the values are sampled from a Gaussian distribution and therefore we consider a Gaussian Naive Bayes. The accuracy rate is 87.5%.

VII. CONCLUSION

Several lung cancer risk prediction models have been developed, but very few have assessed the predictive ability of lung function, it's risk status and accuracy rate. A prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database. The most effective model to predict patients with Lung cancer disease appears to be SOM algorithm. The self-organizing map (SOM) is an excellent tool in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties -of the data. When the number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped. In this paper, we analyze different approaches to clustering of the SOM are considered. The two-stage procedure-first using SOM to produce the prototypes that are then clustered in the second stage-is found to perform well when compared with direct clustering of the data and to reduce the computation time. IN some cases even in the advanced level Lung cancer patients does not show the symptoms associated with the Lung cancer Prevalence of Lung cancer disease is high in India, especially in rural India, did not get noticed at the early stage, because of the lack of awareness. Also it is not possible for the voluntary agencies to carry out the screening for all the people. The emphasis of this work is to find the target group of people who needs further screening for Lung cancer disease, so that the prevalence and mortality rate could be brought down. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining. We sought to develop and internally validate a model incorporating lung function using data from the dataworld. Our prediction model will not only can help people to know lung function at a low cost but also help them take necessary steps based on their cancer risk level. This study has limitations in the size of the 'Big Data' used in Machine Learning. Although the data provided by Kaggle, which was used in this study, is most suitable, it is assumed that there is a limit to learn the data fully due to the lack of absolute figures. In addition to Kaggle, we requested the big data on lung cancer to the Health Insurance Review & Assessment Service Institute, the National Cancer Center, and the Korean Association for Lung Cancer, but it was impossible to obtain big data on lung cancer that was suitable for domestic conditions without the cooperation of the agencies, since the data was not disclosed to the public during the project. Therefore, it is judged that if there would be agency's cooperation in the subsequent research for comparing and analyzing various kinds of algorithms other than those used in this study, a more accurate screening machine for the lung cancer could be created.

VIII. REFERENCES

1. https://archive.dhakatribune.com/health/2021/01/25/report-lung-cancer-on-the-rise-in-bangladesh
2. Beckles, M. A., Spiro, S. G., Colice, G. L., & Rudd, R. M. 2003. Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes. Chest, 123(1), 97S-104S.
3. Cooley, M. E. (2000). Symptoms in adults with lung cancer: a systematic research review. Journal of pain and symptom management, 19(2), 137-153.
4. Krech, R. L., Davis, J., Walsh, D., & Curtis, E. B. (1992). Symptoms of lung cancer. Palliative Medicine, 6(4), 309-315.
5. Gift, A. G., Stommel, M., Jablonski, A., & Given, W. (2003). A cluster of symptoms over time in patients with lung cancer. Nursing research, 52(6), 393-400.
6. Birring, S. S., & Peake, M. D. (2005). Symptoms and the early diagnosis of lung cancer. Thorax, 60(4), 268-269.
7. Mostafa, M. G., McDonald, J. C., & Cherry, N. M. (2008). Lung cancer and exposure to arsenic in rural Bangladesh. Occupational and Environmental Medicine, 65(11), 765-768.
8. Ganti AK, Klein AB, Cotarla I, Seal B, Chou E. "Update of Incidence, Prevalence, Survival, and

Initial Treatment in Patients With Non–Small Cell Lung Cancer in the US". JAMA Oncol. 2021;7(12):1824–1832. doi:10.1001/jamaoncol.2021.4932
9. Lillian L. Tsai. (November 2022). "Lung Cancer in Women" (Vol. 114). The Annals of Thoracic Surgery.
10. Chen, X., Mo, S. & Yi, B. The spatiotemporal dynamics of lung cancer: 30-year trends of epidemiology across 204 countries and territories. BMC Public Health 22, 987 (2022).
11. Matthew B. Schabath. (1 October 2019). Cancer Progress and Priorities: Lung Cancer (Vol. 28). American Association for Cancer Research.
12. de Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018 Jun;7(3):220-233. doi: 10.21037/tlcr.2018.05.06. PMID: 30050761; PMCID: PMC6037963.