Lung Cancer

Alyssa Alexandra lee[1], Aren Deza[1], and Victor Sumbong[1]

University of The Philippines Visayas, Miag-ao, Iloilo City

1 Dataset Link

https://drive.google.com/file/d/1O27QpQU2-oN4NUHeOZOFmLFFMmp7cm3I

2 Possible Questions

1. Which factor has the strongest relationship to getting lung cancer?
2. Build a model that can predict lung cancer based on given data.
3. Which ML method is the best in predicting if you have a risk of lung cancer?
4. Build a model that can predict the severity of a lung cancer patient's symptoms based on given data.
5. How strong is the relationship between a history of smoking/passive smoking and the severity of a patient's lung cancer symptoms?
6. How is a patient with lung cancer best treated?
7. What are the three factors that people with high level of lung cancer have in common?

3 Description of The Project

Lung cancer is a disease characterized by a malignant growth or tumor formed from the uncontrollable division of abnormal cells, specifically within the lungs [1]. It is known to be the most common type of cancer and is also the main cause of cancer death globally. The chance to get and develop this disease increases with a myriad of factors such as tobacco smoking, exposure to chemicals in the workplace, long-term exposure to air pollution, exposure to radon gas, having a family history of lung cancer, etc.

In the Philippines, it is common to see a person smoking as you walk down the street. According to MacroTrends [5], the prevalence of Filipino smokers aged 15 and above on a daily or non-daily basis is 22.90% in the year 2020. The population density in the Philippines is also greatest in the National Capital Region with 21,765 persons per square kilometer [6]. This region also accounts for 13,484,462 persons of the 109,033,245 or 12.37% total population of the country. In NCR, the most densely populated city was Manila with 73,920 persons per square kilometer. According to IQAir [4], in 2019, Manila had an average of 18.2 $\mu g/m^3$ of PM2.5 or particulate matter under 2.5 micrometers which poses health concerns. This recording placed them under the "Moderate" category of the World Health Organization (WHO). However, with the lockdown being lifted and traffic becoming more of a problem, the air quality is forecasted to worsen. In a press release by the Department of Health in 2021 [2], Lung cancer is the 2nd leading type of cancer in the country and is the leading cause of mortality in all cancer types.

If it is identified and diagnosed in its earliest stages, lung cancer can fortunately be treated. However, a patient suffering from lung cancer can be asymptomatic while the disease is in its earlier stages of progression, which can make it difficult to detect. For this reason, having the ability to quickly and reliably predict whether a patient may have or is at risk of contracting lung cancer can be invaluable in reducing their chances of dying from the disease.

Therefore, the researchers aim to use machine learning technology to identify significant factors among lung cancer patients and construct a model that can be used to predict and evaluate an individual's risk of contracting the disease while choosing the best machine learning algorithm that the researchers have learned.

There have been studies such as "A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms" by Gupta et al. [3] which compared 3 machine learning algorithms in their ability to predict lung cancer, however, their dataset was biomedical images and they had an image processing step. Other similar studies also used a collection of labeled images for their dataset. A similar study that did not utilize an image dataset is "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms" by Radhika et al. [7] It used SVM, Logistic Regression, Naïve Bayes, and Decision Tree machine learning algorithms as well as a different dataset than the researchers will use. This paper differs due to its different dataset as well as the machine learning algorithms that will be utilized and compared.

4 Proposed Method

To fulfill our intended goal of predicting lung cancer, we will compare the performance of various machine learning algorithms such as Naïve Bayes and Logistic Regression.

Features such as age, gender, level of air pollution exposure, level of alcohol use, level of dust allergy, level of occupational hazards, level of genetic risk, level of chronic lung disease, level of balanced diet, level of obesity, level of smoking, level of passive smoker, level of chest pain, level of coughing of blood, level of fatigue, level of weight loss, level of shortness of breath, level of wheezing, level of swallowing difficulty, and level of clubbing of finger nails will serve as predictors to the model [8]. 75% of the dataset will be the training set while the other 25% will be the test set.

5 Dataset

The dataset contains information on patients with lung cancer. This includes background information such as their age, gender, exposure to air pollution, alcohol use, occupational hazards, genetic risk, balanced diet, obesity, history of smoking or passive smoking, and presence of dust allergies and chronic lung disease. It also provides data of their lung cancer symptoms based on severity, which includes chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of finger nails, snoring, and the level of lung cancer. This dataset was retrieved from the website, Kaggle (https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link).

6 Metrics of Evaluation

Since this mini-project utilizes the classification technique, the performance of the created model will be measured by its accuracy, precision, recall, f1-score, and confusion matrix.

7 Tools and Packages

In order to conduct this study, the researchers will make use of the following tools: (1) Jupyter Notebook, (2) Google Docs, and (3) TeXworks. Jupyter Notebook is where the model will be built. Google Docs and TeXworks are utilized for the creation of the documents.

The LaTeX packages used are the following: (1) color, (2) graphicx, (3) amsmath, (4) listings, (5) pdflscape, (6) geometry, and (7) natbib. In order to follow a LaTeX template, these packages should be installed. The Python packages that will be employed are: (1) Pandas, (2) NumPy, (3) scikit-learn, (4) Matplotlib, and (5) StatsModels. These packages are necessary in Naïve Bayes and Logistic Regression.

Bibliography

[1] Cancer Research UK (2019). Lung cancer.

[2] Department of Health (2021). Doh leads national lung cancer awareness month, free medicines available nationwide.

[3] Gupta, A., Zuha, Z., Ahmad, I., and Ansari, Z. (2022). A study on prediction of lung cancer using machine learning algorithms.

[4] IQAir (2022). Manila air quality index (aqi) and philippines air pollution.

[5] MacroTrends LLC (2022). Philippines smoking rate 2000-2022.

[6] Philippine Statistics Authority (2021). Highlights of the population density of the philippines 2020 census of population and housing (2020 cph).

[7] Radhika, P., Nair, R. A., and Veena, G. (2019). A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–4. IEEE.

[8] The Devastator (2022). Lung cancer prediction.