A META ANALYSIS OF PEOPLE LIFESTYLE DATA TO PREDICT LUNG CANCER USING DEEP LEARNING APPROACHES

[1]SHABIR AHMAD, [2]TAEG KEUN WHANGBO

[1,2]IT Convergence Engineering, Gachon University, Korea Republic of
E-mail: [1]shabir@gachon.ac.kr, [2]tkwhangbo@gachon.ac.kr

Abstract - Lung cancer has costed over 500,000 deaths in the past few years in United States. Early screening of Lung cancer using Computer Tomography can reduce the risk of mortality significantly. However, it is wiser to correlate the cause of Lung cancer with the people lifestyle to proactively prevent it rather than diagnosing it. There are volumes of literature studies which uses Computer Tomography images for predicting the lung cancer using Deep learning approaches. However, such approaches can help only in diagnosing whether the patient has lung cancer or not. In this paper, we investigate the strongest correlated features based on the biomarkers dataset of people and predict the early symptoms of lung cancers. This work can aid people in predicting the risk of lung cancer with low-cost machine learning model and in the case of positive patients take the appropriate decisions.

Keywords - Lung cancer, Ensemble, Healthcare 5.0, Deep learning.

I. INTRODUCTION

Healthcare 5.0[1][2], the 5th revolution in Healthcare is undergoing through an extensive research phase. The vision of Healthcare 5.0 is to proactively manage the health of human being with non-conventional techniques such as exergames, digital mental health interventions, serious games[3][4], to name a few. It is a well-known medical fact that every disease can be prevented by adopting healthy lifestyle. Thus, investigating it with the people of biomarkers or lifestyle is one of the vital research topics keeping in mind the vision of Healthcare 5.0. Lung cancer has taken many lives across the globe over the years and it was reported 180,000 fatalities in 2018 in the United states only [5]. Early detection of Lung cancer is paramount to be able to save patients from death.

In modern era of Digital of transformations, early detection of disease is made possible with the combination of Computer tomography (CT) [6] images with sophisticated machine learning algorithms. The feature extracted from these images using convolution neural networks and its derived networks such as ResNET, GoogleNet aid in diagnosing the patients with a sufficiently high accuracy. These approaches traditionally do amazing job of diagnosing fatal diseases, however, at times the provided CT images [7] or the machine learning algorithms generate true positive or false negative and also at times suffer from human biasness.

Nonetheless, with a high accuracy and satisfactory result it can successfully diagnose the disease and play no part in the prevention of diseases well before its occurrence. The proactive management of healthcare as a pivotal aspect of healthcare 5.0 vision can be addressed with combining early diagnosis with disease prevention by adopting healthy lifestyle. This work is not to fill the gap in the current state-of-the-art in the field of medical imaging to diagnose lung cancer, but to aid the prevention of lung cancer by providing of meta analysis of people of biomarkers and lifestyle data. We use a recent dataset taken from online system for cancer risk status survey system which survey people for different questions such as smoking, anxiety, asthma, to mention a few.As part of the this work, we preprocess the raw data and perform machine learning algorithms namely regression, XGBoost and Random Forest to predict the risk index of the subject under his lifestyle history. The XGboost being the better model among the alternative can play a pivotal role to identify people with high cancer risk and as a result, can help in the proactive management of cancer lung as a step forward toward the realization of Healthcare 5.0. The rest of the paper is organized as follows: Section 2 discusses the experimental setup and model. Section 3 presents results and implementation environment. Section 4 provides analysis and discussion and finally section 5 concludes the paper and identifies future directions.

II. DETAILS EXPERIMENTAL

In this section, a details steps of the proposed work is exhibited. The proposed system model is elucidated in Figure 1. The data is gathered from different sources through surveys and questionnaires using an online webform. The data collected is in raw form, so the next step clean the data using various approaches such as preprocessing, pattern analysis and possibility of data extension and expansion. Afterwards, selection of appropriate algorithm is selected and other candidates' state-of-the-art methods are considered for comparisons. Finally, the trained model is tested on some new data and the risk prediction is made based on the model.

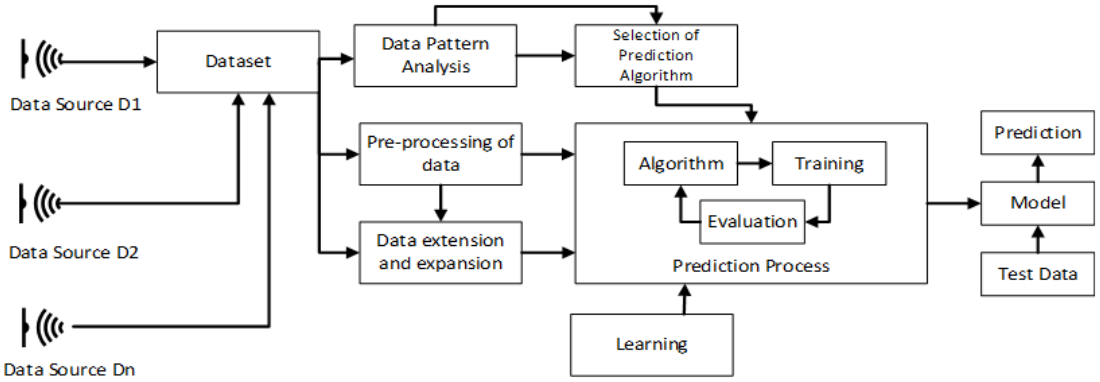

Figure 1. Proposed system model and flow of operations for lung cancer risk prediction.

2.1. Datasets

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system as shown in Figure 2. There are 16 total number of attributes for which 284 instances are recorded, which include gender, age, smoking habits, yellow fingers, anxiety, to mention a few. Figure 1 shows the snapshot of dataset along with the meta information about each individual attribute. The data in most of the cases shows 1 and 2 values, indicating false in former and true in later case. Thus part of the preprocessing includes normalize it to 0 and 1 for false and true values respectively. For instance, one of the perception of the lung cancer diagnosis is that the potential patient fingers turned yellowish. Thus, indicating a value of 2 indicates a person with yellow fingers and 1 otherwise.

#	Column	Non-Null Count	Dtype
0	GENDER	309 non-null	object
1	AGE	309 non-null	int64
2	SMOKING	309 non-null	int64
3	YELLOW_FINGERS	309 non-null	int64
4	ANXIETY	309 non-null	int64
5	PEER_PRESSURE	309 non-null	int64
6	CHRONIC DISEASE	309 non-null	int64
7	FATIGUE	309 non-null	int64
8	ALLERGY	309 non-null	int64
9	WHEEZING	309 non-null	int64
10	ALCOHOL CONSUMING	309 non-null	int64
11	COUGHING	309 non-null	int64
12	SHORTNESS OF BREATH	309 non-null	int64
13	SWALLOWING DIFFICULTY	309 non-null	int64
14	CHEST PAIN	309 non-null	int64
15	LUNG_CANCER	309 non-null	object

Figure 2. Snapshot of Dataset

2.2 Data Analysis and Preparation

First a list of steps are performed on the raw data. The preprocessing steps include replacing some of the values for efficient processing. For example, 1 is processed as 0 and 2 is processed 1. Similarly, Male is represented as 0 and Female is represented as 1. As part of the data analysis, a number of processes are performed.

The histogram of the important features are generated to know the distribution of the values. This can help in the general trend of the people. Figure 3, for instance, shows histogram of all the features. It is obvious that the data is recorded for people with a slightly more fraction of males than females. Similarly, the number of people with age 60 to 70 years is maximum. In order to investigate the impact of harmful lifestyle the dataset has more records corresponding to people who have yellow fingers, smoking habits, fatigue and some sort of allergy and shortness of breath.

2.2 Feature Prioritization and Reduction

While cleaning the dataset, one of the vital processes is the feature prioritization. Feature prioritization sometimes led to another process also known as feature reduction. In this subsection, we outline the important features using SNS heat map and identify the different features and their correlation among them as shown in Figure 4. From the graph, it is evident that swallowing difficulty, alcohol consumption, allergy has strong positive impact followed by chest pain, coughing, yellow finger. On the other hand, shortness of breath, chronic disease and smoking have weak positive correlation with lung cancer.

Alternatively, gender has negative correlation with lung cancer. There are some correlation of other attributes which are not directly relevant to lung cancer, but can indirectly have impact on predicting the risk of cancer. For instance, there are very strong correlation instances among features such as chest pain and gender, swallowing difficulty and coughing, shortness of breath and fatigue which can impact the overall outcome of the lung cancer risk.

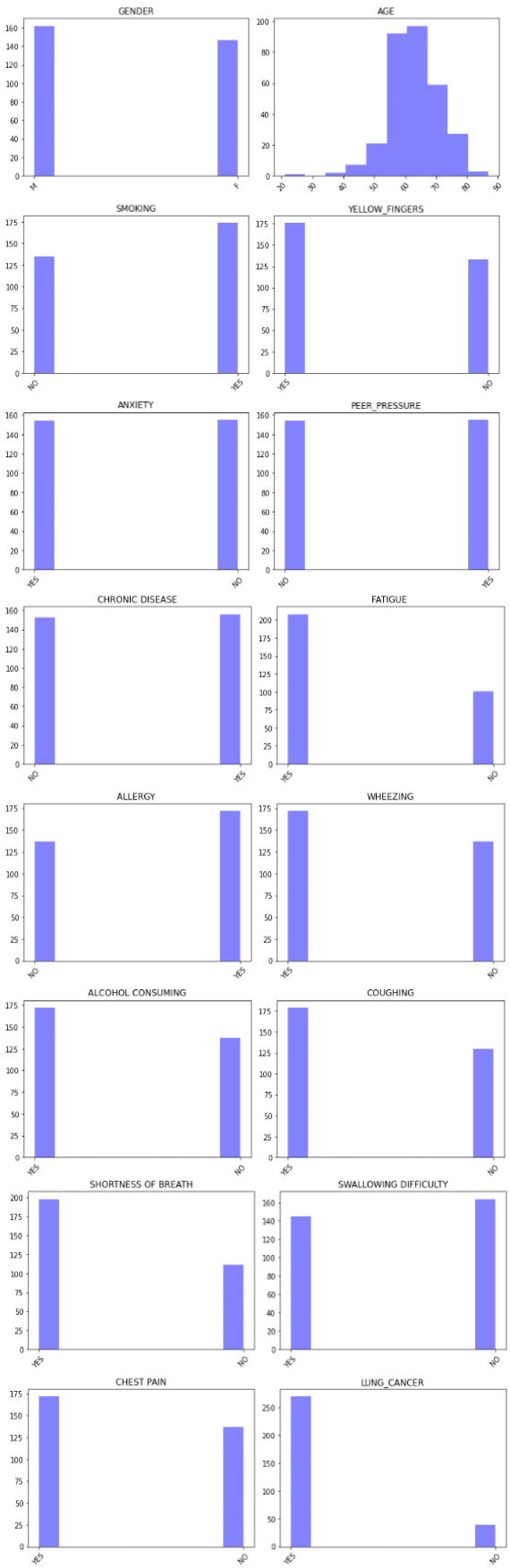

Figure 3. Features histogram analysis.

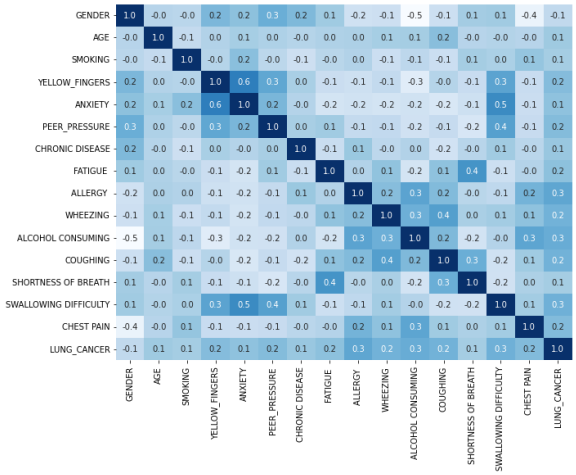

Figure 4. Features correlation map.

III. RESULTS AND DISCUSSION

3.1. Implementation Environment

We have implemented the proposed work using python programming language. For data manipulation and analysis we used pandas and numpy. Similarly, some helping python libraries are utilized such as matplotlib, sns, to name a few. Similarly, we used pytorch for training and testing machine learning models. The model persistence is provided using Python pickle package and for boosting up the training process we equipped our machine with Cuda 10.1.

Technology	Description
Core programming language	Python 3.10
Operating system	Window 10
Cuda	10.1
Helping libraries	Pandas, Numpy, Sklearn, matplotlib
Deep Learning Library	Pytorch
RAM	64 GB
Model Persistence	Pickle

Table I. Implementation tools and technologies

3.2. Model Analysis and Results

In order to analyze the models, we have trained an ensemble classification model using different models such as logistic regression, support vector machine, naïve Bayesian, decision tree, random forest and XGBoost. The models are automatically tested for accuracy, precision, recall and other performance evaluation metrics. The final model is selected as the best model, which is turned out to be the XGBoost model. The confusion matrices of each individual model along with the final ensemble model is depicted in Figure 5. While KNN classification generates 0 true positive and 5 false positive, yet the accuracy of these models are above 90. The final selected model has an overall accuracy of 93.85%.

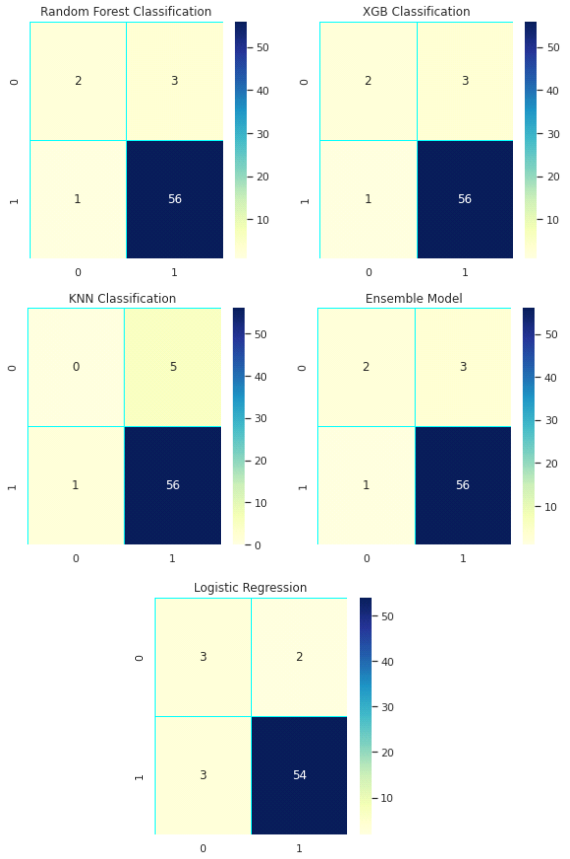

Figure 5. Confusion matrices comparisons.

Moreover, based on the best model, the most important features are investigated, and which is shown in figure 6. The age has an adverse effect on lung cancer and people with high age is more likely to have lung cancer. Other crucial features are shortness of breath, yellow finger, and alcohol consumption. Therefore, a person with more age and any of the other top 3 features can have a very high risk of lung cancer. In contrast, some features such as chest pain, smoking and allergy, are although regarded commonly as potential risk for lung cancer, but our results show they are the least important features.

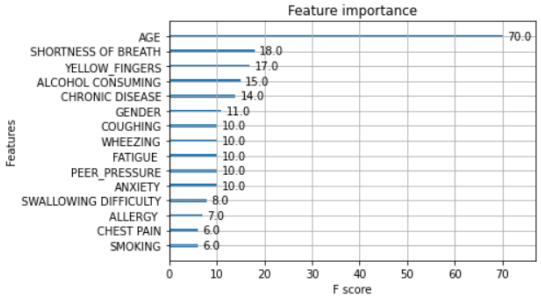

Figure 6. Features importance

IV. DISCUSSION

For proactive lung cancer management, it is vital to investigate the most crucial parameters. In this paper, we have investigated different state-of-the-art machine learning models to derive the features which are of massive importance for lung cancer. It has been previously highlighted that age is a very huge factor and people with age more than 70 are highly susceptible to lung cancer. According to our dataset, the average age of people on the high risk of lung cancer is around 60. However, it does not necessarily rule out the younger people. Similarly, in another article, allergy and lung cancer are highly correlated. Asthma patients are highly exposed to risk cancer and even more dangerous when a smoking person has asthma. From our results, it has been verified that the patients with asthma has no direct association with lung cancer and similarly smoking may not necessarily worsen the risk of lung cancer for people with allergy. Finally, it has been investigated in different articles that yellow fingers are one of the pre-symptoms of lung cancer [6]. However, from our results it is proved as inconclusive. The fingers can be turned yellow due to a variety of reasons in which smoking is one of the notables. Nevertheless, associating it with lung cancer is not conclusive in this study.

V. CONCLUSIONS

In this paper a potential risk of lung cancer based on the lifestyle of the people and their medical history has been investigated and it has been classified as the person is at high risk of getting lung cancer and thus has to plan accordingly by changing their lifestyle or changing other bad habits. We have used a dataset based on the online survey and questionnaires. After preprocessing and cleaning the data, a number of well-known training algorithms are applied to find the risk being high or low. The models are automatically compared and the best model among them is ensembled for final predictions. The accuracy of the final model shows 93% which is sufficient to infer the potential risk of lung cancer for a given test case. Such low-cost solutions can alarm users before being fall into such fatal disease and thus is a step-forward for the realization of Healthcare 5.0.

REFERENCES

[1] Bhavin, Makwana, et al. "Blockchain and quantum blind signature-based hybrid scheme for healthcare 5.0 applications." Journal of Information Security and Applications 56 (2021): 102673.

[2] Mohanta, Bhagyashree, Priti Das, and Srikanta Patnaik. "Healthcare 5.0: A paradigm shift in digital healthcare system using Artificial Intelligence, IOT and 5G Communication." 2019 International Conference on Applied Machine Learning (ICAML). IEEE, 2019.

[3] Ahmad, Shabir, et al. "Architecting Intelligent Smart Serious Games for Healthcare Applications: A Technical Perspective." Sensors 22.3 (2022): 810.

[4] Ahmad, Shabir, Faheem Khan, and Taeg Keun Whangbo. "Performance Evaluation of Topological Infrastructure in Internet-of-Things-Enabled Serious Games." (2022).

[5] Qu, Yan-Liang, et al. "Asthma and the risk of lung cancer: a meta-analysis." Oncotarget 8.7 (2017): 11614.

[6] Lakshmanaprabu, S. K., et al. "Optimal deep learning model for classification of lung cancer on CT images." Future Generation Computer Systems 92 (2019): 374-382.
[7] Avanzo, Michele, et al. "Radiomics and deep learning in lung cancer." Strahlentherapie und Onkologie 196.10 (2020): 879-887.
[8] Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954-961.

★★★