An Analysis of the Performance of Machine Learning Algorithms for Prediction of Lung Cancer

Dr. Snehal Rathi
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
snehal.rathi@viit.ac.in

Mrs. Pranali Kshirsagar
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
pranali.kshirsagar@viit.ac.in

Ankita Mandhare
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
ankita.22110173@viit.ac.in

Prasad Jagadale
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
prasad.22110189@viit.ac.in

Komal Patil
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
komal.22110255@viit.ac.in

Sachin Nakate
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
sachin.22110381@viit.ac.in

***Abstract*—Cancer accounted for nearly 10 million deaths in 2020. Prostate, lung, breast, colon, colorectal, stomach, skin, etc are the most common cancers. Lung cancer is the leading cause of deaths related to cancer accounting for estimated 1.8 million deaths in 2020 according to the International Agency for Research on Cancer (IARC). Factors like smoking, occupational hazards, air pollution, chronic lung diseases of the past, and hereditary cancer syndromes can be causes of lung cancer disease. It is often diagnosed at later stages and in the last few decades, Lung Cancer (stage 4) is one of the critical and incurable diseases. Hence beforehand detection and diagnosis is necessary. To detect the cancerous condition various machine learning algorithms have been used in this process due to its accuracy. Logistic regression, naive Bayes, MLP, XGBoost, Decision Tree, random forest, ridge model, SVC, gradient boosting, etc. are some of the algorithms. Their ability to detect lung cancer is well-known. The paper factors that cause lung cancer and application of ML algorithms will be discussed. Two datasets from kaggle have been taken and compared with different algorithms. The data has been preprocessed to improve its reliability. To analyze the performance of an algorithm, accuracy,precision,recall value is used according to different datasets . The data is divided into training and testing in a 7:3 ratio for fair results. People will be able to see pros and cons of different algorithms through this research paper.**

Keywords—cancer,machine-learning,algorithms,application, logistic regression, MLP(Multi-Layer Perceptron Learning), Xgboost, Decision tree, ridge model, random forest, naive bayes,gradient boosting.

I. Introduction

Today, lung cancer, also called lung carcinoma, is the leading cause of worldwide cancer-related deaths. There are various types of cancers: stomach, skin, prostate, lung colorectal, etc. According to a survey among cancer patients 22% are diagnosed with lung cancer [fig.1]. Smoking, air pollution, and workplace hazards like chemicals and asbestos are the risk factors for lung cancer disease. It starts when abnormal cells grow in an unrestrained way in the lungs. Early identification and treatment are very important for the patient to have a higher chance of surviving. Thus nowadays this is a great challenge for doctors to detect and cure lung cancer at an early stage. Lung cancer treatments include immunotherapy, radiotherapy (radiation), surgery, chemotherapy etc. [18]

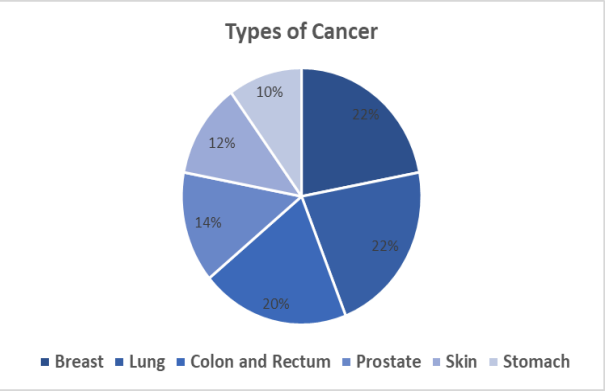

Fig.1 *Types of Cancers* [1]

In 1980, Computer-Aided Diagnosis (CAD) started in which with the help of images doctors interpreted about diseases. The algorithms like decision tree, SVM, K-nearest method, etc. used in machine learning showed their impact in the healthcare sector. The research will convince and show this impact on people.

The symptoms of cancer are categorized according to the size of tumor and its location. There are several types of lung cancer like Adenocarcinoma, Small Cell Lung Cancer (SCLC), Non-small cell carcinoma(NSCLC),Squamous cell carcinoma, etc. SCLC is rare but often its growth is rapid while on other hand, NSCLC is common but it grows slowly.

Table 1 Causes for Lung Cancer[1]

Causes for Lung Cancer	Cases(%)
Smoking	85%
Air pollution, hereditary cancer syndrome, previous chronic lung cancer disease	15%

It is difficult to identify cancer at an inaugural stage as in some cases there are no symptoms like chest pain in the chest, coughing, loss syndrome (shoulder pain), etc. Among lung cancer patients, 85% are affected due to smoking. Impregnation of tobacco smoke is known as passive smoking. There are some cases in which heredity is also a reason for lung cancer. Doctors and Physicians can identify the existence of cancer by tests like CT scans, X-rays, etc.

Table 2 Global lung cancer incidence and mortality rate,2000 [1]

World	Cases per 100000 Population	Deaths per 100000 Population
Male	34.92	31.43
Female	11.05	9.53

According to Survey of WHO (2020), the death rate according to case ratio is the same in males and females. Now after 20 years, the death rate has drastically increased. Lung cancer caused 1.8 million deaths among nearly 10 million deaths due to cancer according to GLOBOCAN (Global Cancer Observatory) 2020 produced by International Agency for Research on Cancer (IARC)**.**[18]

II. Machine Learning In Healthcare

Machine learning has made significant inroads in the last few years and has the potential to revolutionize how medical professionals diagnose, treat, and manage the care of patients in the future. As a result of machine learning in healthcare, Several key areas have been explored some of them are as follows:

i. *Disease Diagnosis and Detection:* An algorithm based on machine learning can analyze medical images like CT scans, X-rays, and MRI's to help detect diseases like cancer, tuberculosis, and diabetic retinopathy in the early stages of development.

ii. *Drug Discovery and Development*: It has been found that machine learning models can help identify potential drug candidates by analyzing chemical properties and predicting the efficacy of the drug. It is also possible to use machine learning algorithms to optimize clinical trial designs, which may result in a faster process of drug development by using ML algorithms.

iii. *Patient Monitoring:* In the future, wearable devices and sensors will be able to collect real-time patient data, which will be able to be analyzed by machine learning models to provide early warnings of health conditions and help monitor them.

iv. *Image and Signal Processing:* ML is a technique used for improving the quality of medical images, such as denoising and enhancing images, by using machine learning. It is possible to diagnose cardiac and neurological conditions by analyzing ECG(electrocardiogram) and EEG signals.

v. *Predictive Analysis*: By identifying admissions of patients, readmission, and disease outbreaks, predictive models assist hospitals in allocating resources efficiently.

There is no doubt that machine learning has enormous potential in healthcare, but it also raises important ethical and privacy issues that must be addressed. The protection of patient data, the assurance of the reliability of algorithmic models, and the addressing of biases in AI models are all critical concerns that must be addressed as the field continues to develop. In addition, regulatory agencies such as the FDA(Food and Drug Administration) are actively working on guidelines for the use of artificial intelligence in healthcare to ensure that it is safe and effective.

III. Literature review

Ensemble machine learning models validate to determine eligibility for risk-based lung cancer screening. Machine learning model used data of around 2,16,714 ever-smokers & 26,616 high-risk smokers .Model was developed to predict the risk i.e diagnosis of lung cancer and death from lung cancer. Three variables are used mainly that are age, smoking duration, and pack-years.[21].

In this research data in the UCI repository consist of 32 tuples in the dataset, and each tuple has 57 features. The K-Nearest Neighbor algorithm(KNN) compared with the Decision Tree classifier, the KNN method came out with 68.9% Accuracy.SVM and smote together are used by optimizing the features of the dataset an accuracy of 98.8% is obtained.[3]

In the dataset of primary lung adenocarcinomas 86 features and ten non-neoplastic features are collected[4]. The whole dataset contains almost 7100 tuples. The dataset is divided into two parts, a training and testing model in the ratio of 7:3. The result is then calculated on both the training and testing model. The model used here is Multi-layer perceptron(MLP), random subspace, SMO. The accuracy of this model is Multi-layer perceptron is 86.667%, random subspace is 68.33%, SMO is 91.667%[4].

Syed Saba Raoof and team describes the factors that cause lung cancer. Various ML algorithms and their strengths and weaknesses have been discussed. SVM is the most useful method for classification,prediction and regression. The findings indicate that Deep Learning algorithms like CNN (Convolution),ANN (Artificial),FCN (Fully Convolutional),RNN (Recurrent) should be used to enhance the accuracy of identifying as well as prediction of lung cancer.[6]

V.Krishnaiah and his colleagues present that Naive Bayes is the most effective model for predicting Lung Cancer disease followed by IF-THEN rule,Decision Trees and Neural Network. The drill through feature in Decision Trees make it easier to understand and interpret. Among all these, Naive Bayes is better because it could identify all significant medical predictors.[7]

IV. Process Methodology

The patients are categorized depending upon their symptoms . First, data collection takes place and after that the next step is to process the data, which is the most important step . The accuracy, efficiency, and quality of selected data can be predicted by using various algorithms. Classified data is trained and tested on the standard datasets. The data is analyzed to understand the most useful method to identify significant chemical properties associated with lung cancer. All the features in data are given as input. These preprocessing techniques convert the raw data into

organized format . Data for this process are categorized into training and testing sets where the ratio is 7:3[19].The model understands chemical properties that affect accurate prediction through rigorous features selection using methods like correlation analysis .This approach provides insights to chemical contributors by achieving high accuracy.

1) *Data Selection*:-

For this research, two datasets are used from Kaggle. First is a repository named Lung Cancer Patient Dataset and second is the Lung Cancer survey respectively. The first dataset contains 25 features, of around 1000+ patients and dataset 2 comprises a total of 17 features around 600+ tuples. The main aim of our proposed study is to analyze and compare the accuracies and performances of the best preferred model , among all given models.

2) *Preprocessing the data*:-

Preprocessing the data is the inaugural state in cancer detection and there are various techniques to preprocess. The data needs to be clean by filling in the missing values, smoothing noisy data, and removing outliers. Data reduction involves dimensional reduction (removing unimportant attributes), numerosity reduction (obtaining smaller volume while preserving originality), etc. Transforming data includes feature construction, aggregation, discretization (concept hierarchy), normalization (scaled to a smaller range), etc. So, imputing the dataset to the model increases the overall reliability of the entire dataset.

3) *Training and Testing Sets*:-

The datasets are divided by the training and testing sets into a 7:3 ratio. The dataset should be divided randomly for correct results. Firstly, the training dataset is used to train the algorithm and it is then evaluated on the testing dataset. All 9 different algorithms are used to compare performances.

4) *Model Efficiency*:-

There are 9 algorithms tested on the dataset. For each model, we have calculated Accuracy, precision, and recall value. Accuracy is used to measure the correctness. It is given by,

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

Precision is a positive prediction made by the model.

Precision = TP/ (TP + FP)

Recall is a measure to find all instances of positive class.

Recall = TP/ (TP + FN)

here,

TP = True Positive

TN= True Negative

FP= False Positive

FN=False Negative

V. DATASET

Two kaggle datasets have been used, the first dataset contains 25 attributes.The field “level” tells us about whether the patient has lung cancer disease. it contains an integer value ranging from 0 (no_cancer)to 2(cancer) . The integer value 1 implies that a person may or may not have the disease. Dataset 2 has 15 attributes. Lung Cancer disease can be predicted using these following attributes in graphs. Some Of the attributes like age,gender,smoking,alcohol,chest pain,etc are common in both the datasets.Mean is shown in red color while median is in black. Mean tells about typical values while it may be influenced by skewness whereas median represents middle value and is less affected by skewness. A KDE (Kernel Density Estimation) gives a smoothed outlook of the data. Fig.2.1 and fig.2.2 have more smooth graphs as compared to fig.3.1,fig.3.2 and the size of the dataset might be one of the reasons. Fig.3.1,fig.3.2 has two distinct peaks which may indicate bimodal distribution , different subgroups, or distinct patterns within the distribution.

1. Air pollution
2. Alcohol Use
3. Genetic risk
4. Chronic Lung disease
5. Obesity
6. Smoking
7. Chest Pain
8. Weight Loss
9. Shortness of breath
10. Frequent cold
11. Coughing
12. Yellow fingers
13. Anxiety

Data Preprocessing is the most important and time-consuming part of the process. Various algorithms and methods are used to understand different trends and patterns. Visualizations consisting of bar charts, histograms, scatter plots, box plots, sunbursts, etc make it quite easier to extract hidden information which helps in making decisions. Correlation and relationships between variables can be found using scatter plots and sunbursts graphs. In the dataset we used for lung cancer prediction 70% data is used to train the model and 30% is used for testing.

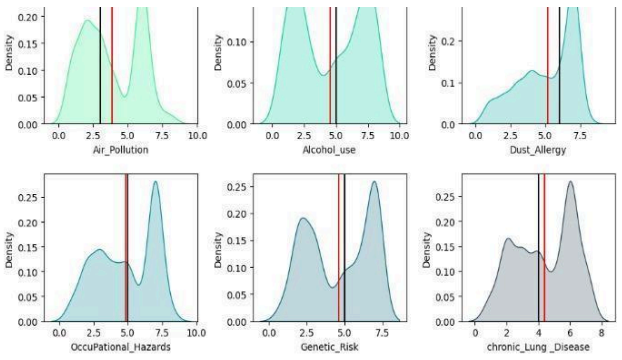

fig.2.1 Visualization of the first dataset

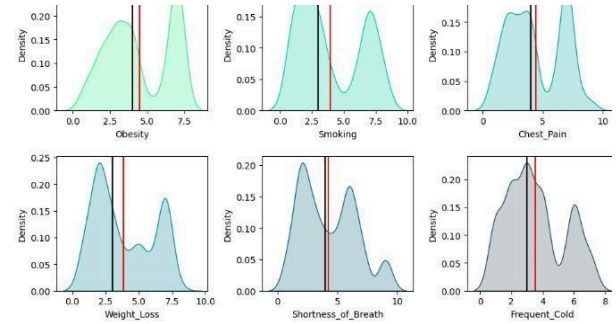

fig.2.2 Visualization of the first dataset

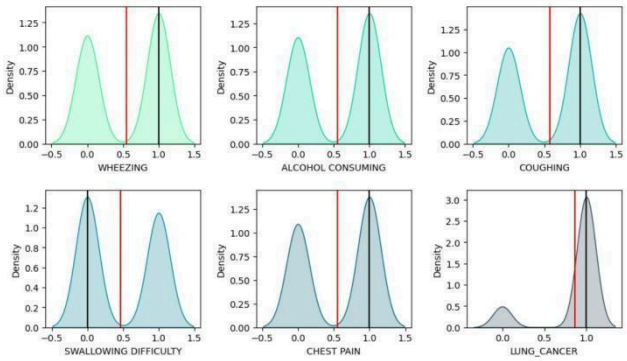

fig.3.1 Visualization of the second dataset

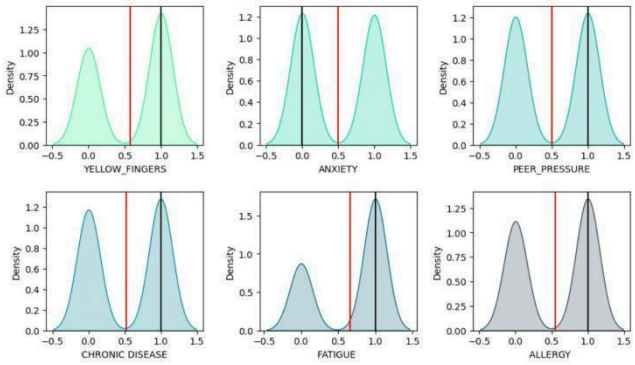

fig.3.2 Visualization of the second dataset

VI. Machine learning Algorithms

The machine learning algorithms which are studied and that exist for lung cancer prediction are as follows: Python libraries like numpy and pandas are used for data manipulation and data analysis, matplotlib -seaborn for data visualization and scikit-learn provides machine learning algorithms.

1. **Logistic Regression:** It is a type of statistical model that can be used for predictive analysis and classification. It estimates the probability of an event occurring. This algorithm is used to analyze the correlation between a set of dependent binary variables and independent variables. It is represented by the following formulas:(1)
   Logit(π) = 1/(1+exp(-π))

   ln(π/(1-π)) = Beta_0 + Bets_1*X_1 +... + B_k*K_k[2]

   In the above expression logit(π) is a dependent binary and x is the independent variable. For developing this model, the sigmoidal function is used on input features and takes a decision bound on the class problem to estimate the probability.[2]

   Results obtained from the first Dataset :
   Logistic Regression: 86%
   Results obtained from the second Dataset :
   Logistic Regression: 97%

2. **Random Forest:** Random forest algorithm is a reliable method for predicting and evaluating the risk of various diseases such as lung cancer. It is a combination of all classifiers and three decision trees. Random Forest is used for noisy and overfitting problems for better outcome prediction. The final result is determined by averaging independent tree predictions because each tree generates its prediction. It can be used for large data without any data reduction with features such as smoking history, age, and genetics.

   Results obtained from the first Dataset :
   Random Forest:100%
   Results obtained from the second Dataset :
   Random Forest:98%

3. **Naive Bayes:** Naive Bayes theorem is an easy algorithm for learning that uses the Bayes theorem which is a classification based on the probabilistic method. It has features used for categorization and classification that state that the existence or absence of any feature does not affect the other features. It constantly improves categorization accuracy. It evaluates conditional probability conditioned on input cancer features observation value of p(y|x), each of the classes of y, with an object x. It uses all attributes resulting in sophisticated deterioration in performance that provides robustness in absent value with increasing computational efficiency. In the dataset Bayesian Theorem is applied as follows: p(y|x)=p(x|y)*p(y)/p(x) (3)

   Results obtained from the first Dataset :
   Gaussian Naive Bayes: 86%
   Results obtained from the second Dataset :
   Gaussian Naive Bayes: 92%

4. **Multi-layer Perceptron:** MLP is a key feature of deep learning feed-forward artificial neural network model. MLP correlates the input data of attributes onto a set of relevant outputs data. It is suitable for complex computational tasks.

   Results obtained from the first Dataset:
   MLP: 90%
   Results obtained from the second Dataset:
   MLP: 98%

5. **XGBoost:** The XGBoost tree is highly effective for large datasets with high prediction capability. To minimize overfitting, it employs a variety of regularization techniques, it begins with a single leaf and generates the tree by using an average of two

observations as thresholds .xg boosts enhanced various different model through parallel processing and it handle noisy and missing data effectively by using iterative calculation making it efficiently running speed accurate training and higher scalability.

Results obtained from the first Dataset:

XGBoost: 100%

Results obtained from the second Dataset:

XGBoost: 97%

6. **Decision Tree:** The training data is used to make appropriate decisions on available data for lung cancer prediction to learn and get ready and train the unknown data for making decisions,unravels complex relationships. It appears like a flowchart where each internal node indicates a "test"(whether the patient has cancer or not based on dependent data) and the outcome of this test represents each branch and leaf node in the tree denoting a class label (by evaluating all attributes choice is decided).[9]

Results obtained from the first Dataset:

Decision Tree: 100%

Results obtained from the second Dataset:

Decision Tree: 94%

7. **Ridge Model:** It is an improved version of linear regression that, by applying regularization techniques of ridge expression in the lung cancer dataset, aims to provide the objective of the chance of having lung cancer and other relevant outcomes which include L2 regulation, to avoid overfitting problems addresses multicollinearity. It interprets the labels based on statistical-based fundamental relationships as compared to the value of variance obtained from the least square method that gives a lower value [3]. In the loss function, choose the appropriate regression parameter (alpha), as the higher value of alpha increases the overfitting problem and the lower value shows the underfitting.

Loss Function = Ordinary least squares + Alpha Summation (Squared Coefficient Values)

Results obtained from the first Dataset:

Ridge Model: 92%I

Results obtained from the second Dataset:

Ridge Model: 90%

8. **SVC:** SVC is a supervised learning algorithm suitable for predicting patients' lung cancer outcomes (presence or absence) that has the objective of maximizing the margin in classes. By using different kernel functions, has to undergo sort of categorization, and the designation, hyper-flat is the decision limit as the basic framework that the SVC model provides

Results obtained from the first Dataset:

SVC: 39%

Results obtained from the second Dataset:

SVC: 98%

9. **Gradient Boosting:** This model used different kernel function and regularization parameters that include adjusting both the number of trees and their specific parameters. To minimize loss, gradient descent was utilized, encompassing mean squared error and cross entropy. Each algorithm iteration calculated the gradient of the loss function in relation to prediction. To meet the necessary conditions, the modified model's predictions were continually integrated into the ensemble. This unique model allowed for improved performance through customization, robustness, interpretability, and flexibility.[12]

Results obtained from the first Dataset:

Gradient Boosting: 100%

Results obtained from the second Dataset:

Gradient Boosting: 98%

Selection of this algorithm is based on predictive performance, computational efficiency and interpretability in diverse datasets, This nine machine learning algorithm provides comprehensive exploration for lung cancer prediction,providing a foundational and robust approach.

VII. RESULTS

The accuracy comparison graph shows the accuracy comparison of nine models namely Logistic Regression (LR), Random Forest Algorithm, Naive Bayes, MLP, XGBoost, Decision Tree, Ridge Model, SVC, Gradient Boosting.

This shows the effectiveness of models to predict lung cancer. The results are summarized in a table. These results differ according to the dataset chosen, that's been compared to two datasets results for the research. Models like Ridge Model, Decision tree, random forest, Xgboost, gradient boosting shows more accuracy on dataset 1 than dataset 2. And on the other hand, SVC, MLP, Naive bayes shows more accuracy in dataset 2.

Dataset 1 has 25 attributes and dataset 2 has 15 attributes. The number of tuples in dataset 1 is more than dataset 2 and because of that, in most of the models the accuracy is more for dataset 1. In the case of SVC, it struggles with high dimensional data (Curse of Dimensionality) that's why it shows more accuracy in dataset 2 than dataset 1. In general, for more accuracy more data is necessary.

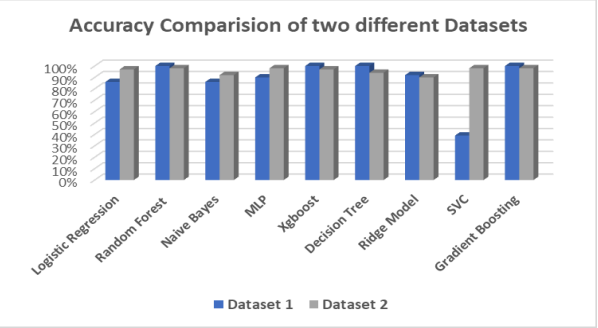

fig.4 *Accuracy Comparison Graph*

VIII. CONCLUSION

In conclusion, the study halts the importance and significance of lung cancer disease prediction using a Machine learning algorithm.[3] the study says that lung cancer has made serious strides in recent years. The

advancements improve the understanding, treatment, and diagnosis of disease. Prediction of lung cancer is a major concern to address but algorithms help in early detection, personalized Treatment, drug Discovery, and in various factors.

Two datasets are used for this research, and nine different algorithms are compared to analyze the result. The Excellent performance of the decision tree and gradient boosting it show the potential as the most effective tool for predicting the accuracy of identifying the risk of lung Cancer. It also has capability to capture complex things, by using these two specific models can be used by healthcare professionals for decision making purposes.

So, the results of research suggest that it is very important to predict the result when it comes to lung diseases. By discovering lung disease early, medical interventions can be implemented promptly, lifestyle modifications can be made, and targeted treatments can be offered. The future scope of Machine learning in lung cancer is AI-Driven precision medicine, integration of multi-omics data, predictive analytics, telemedicine and remote monitoring, early intervention and prevention etc.

References

[1] WHO (World Health Organization) Cancer survey 2020.

[2] IBM, "What is Logistic regression?," *www.ibm.com*, 2022. https://www.ibm.com/topics/logistic-regression

[3] Hindawi, BioMed Research International, "Lung Cancer Prediction from Text Datasets Using Machine Learning", Volume 2023, Article ID 9790635

[4] Jayadeep Pati1,Gene Expression Analysis for Early Lung Cancer Prediction using Machine Learning Techniques: An Eco-Genomics Approach, DOI 10.1109/ACCESS.2018.2886604, IEEE Access

[5] Li et al. (2002). Kent ridge bio-medical data set repository. Institute for Infocomm Research. http://sdmc. lit. org. sg/GEDatasets/Datasets. html.

[6] Syed Saba Raoof, M A. Jabbar, Syed Aley Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach"

[7] V.Krishnaiah, Dr.G.Narasimha, Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, ISSN:0975-9646

[8] Sang Min Park, Min Kyung Lim, Soon Ae Shin & Young Ho Yun 2006. Impact of pre diagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study. Journal of clinical Oncology, Vol 24 Number 31 November 2006.

[9] de Ville, Barry (2013). Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448–455. doi:10.1002/wics.1278

[10] Sperandei, Sandro (2014). Understanding logistic regression analysis. Biochemia Medica, (), 12–18. doi:10.11613/bm.2014.003

[11] Jerome H. Friedman (2002). Stochastic gradient boosting. , 38(4), 367–378. doi:10.1016/s0167-9473(01)00065-2

[12] https://www.geeksforgeeks.org/ml-gradient-boosting/

[13] Timor Kadir1 , Fergus Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques", http://dx.doi.org/10.21037/tlcr.2018.05.15

[14] Wang, Xiaosheng, and Osamu Gotoh. Microarray-based cancer prediction using a soft computing approach. Cancer informatics 7 (2009): 123.

[15] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A hybrid algorithm for lung cancer classification using SVM and neural networks," ICT Express, vol. 7, no. 3, pp. 335–341, 2021.

[16] A New Decision Tree Method For Data Mining In Medicine Kasra Madadi Ouya1 Department of Computing and Science, Asia Pacific University of Technology & Innovation.

[17] Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the twenty-first International Conference on Machine Learning. New York: ACM Digital Library; 2004. p. 78.

[18] S. Rathi, P. Mehta, V. Mishra, A. Karale, A. Choudhari and V. Shingare, "Comparative Study of Heart Disease Prediction Algorithm," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, 2023, pp. 1-6, doi: 10.1109/ASIANCON58793.2023.10270208.

[19] S. Patil, S. Rathi and V. Mankar, "A Novel Approach to Chest Disease Detection from Chest X-Ray Images," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-8, doi: 10.1109/ASIANCON55314.2022.9909156.

[20] Thomas Callender, Fergus Imrie, Bogdan Cebere, Nora Pashayan,Neal Navani, Mihaela van der Schaar, Sam M Janes,"Assessing eligibility for lung cancer screening: Parsimonious multi-country ensemble machine learning models for lung cancer prediction".