IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

Lung Cancer Prediction Using Machine Learning

Vasupalli Saranya[1], Maddila Suresh Kumar[2]

[1]P. G. Student, Dept of Computer Science, GITAM School of Science, GITAM (Deemed to be University),

[2]Assistant Professor, Dept of Computer Science, GITAM School of Science, GITAM (Deemed to be University)

Abstract: Lung cancer has become the most prevalent and threatening cancer worldwide. The main reason for this lung cancer is due to cigarette smoking. Lung cancer is the cancer that originates in the tissues of the lungs. It occurs when cells in the lung start to grow rapidly in an uncontrolled manner. Lung cancer can start anywhere in the lungs and affect any part of the respiratory system. So, lung cancer needs to be detected at early stages and this can be predicted by using various machine learning algorithms. In this paper, KNN algorithm is used for predicting lung cancer and also accuracy score has been calculated along with confusion matrix to identify the correct and incorrect predicted features. In this paper, lung cancer prediction is done by importing the dataset and performing machine learning algorithms on the dataset to find out whether the users are affected with cancer.

Keywords: lung cancer, prediction, KNN, accuracy, performance

I. INTRODUCTION

Lung cancer is one of the main cause of death and health issues in many countries with a 5-year survival rate of only 10-16%. Lung cancer is considered as the deadliest cancer in the world. This lung cancer would actually damage the cells of the respiratory system there by damaging the whole tissues which can leads to breathing problems and difficulty in intake of fresh air anymore and these cases might also lead to the death of a person.

It occurs when cells in the lung start to grow rapidly in an uncontrolled manner. Lung cancer can start anywhere in the lungs and affect any part of the respiratory system. So, this lung cancer needs to be detected at early stages and this can be predicted by using various machine learning algorithms. In this paper, it can be predicted by using some of the popular machine learning algorithms.

1.1 About the Project

The lung cancer prediction can be analysed using classification algorithm called KNN (K Nearest Neighbours). The key objective is the early diagnosis of lung cancer by examining the performance of this algorithm and predicting the stage of cancer for a patient. The implementation of the proposed approach on a lung cancer database reveals 95 % accuracy which gives accurate prediction results to the users.

1.2 Advantages of the System

1. The lung cancer detection using the machine learning algorithm can be used in early diagnosis of cancer in individuals. This can be very helpful for doctors, radiologists for giving a better result for the patients who consult them.

2. This technique can be used for curing the individuals and can also control the occurrence of lung cancer and can save millions of life. Machine learning algorithm can be used for more such improvements in health care and also other sectors.

2. MACHINE LEARNING ALGORITHM

K Nearest Neighbor Algorithm

The k-nearest neighbor algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.
For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used. While this is technically considered "plurality voting", the term, "majority vote" is more commonly used in literature. The distinction between these terminologies is that "majority voting"

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

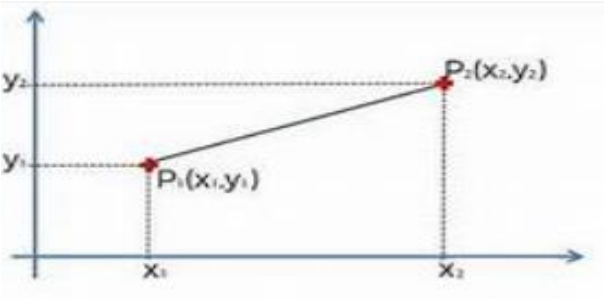
Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

technically requires a majority of greater than 50%, which primarily works when there are only two categories. When you have multiple classes—example four categories, you don't necessarily need 50% of the vote to make a conclusion about a class; you could assign a class label with a vote of greater than 25%.

2.1 Distance Metrics

Euclidean distance (p=2): This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

2.2 Advantages and Disadvantages

Just like any machine learning algorithm, k-NN has its strengths and weaknesses. Depending on the project and application, it may or may not be the right choice.

Advantages

- **Easy to implement**: Given the algorithm's simplicity and accuracy, it is one of the first classifiers that a new data scientist will learn.
- **Adapts easily**: As new training samples are added, the algorithm adjusts to account for any new data since all training data is stored into memory.
- **Few hyperparameters**: KNN only requires a k value and a distance metric, which is low when compared to other machine learning algorithms.

Disadvantages

- **Curse of dimensionality**: The KNN algorithm tends to fall victim to the curse of dimensionality, which means that it doesn't perform well with high-dimensional data inputs. This is sometimes also referred to as the peaking phenomenon, where after the algorithm attains the optimal number of features, additional features increase the amount of classification errors, especially when the sample size is smaller.

- **Prone to overfitting**: Due to the "curse of dimensionality", KNN is also more prone to overfitting. While feature selection and dimensionality reduction techniques are leveraged to prevent this from occurring, the value of k can also impact the model's behavior. Lower values of k can overfit the data, whereas higher values of k tend to "smooth out" the prediction values since it is averaging the values over a greater area, or neighborhood. However, if the value of k is too high, then it can underfit the data.

3. DATASET AND ATTRIBUTES

The below is the sample image of the dataset that is imported from Kaggle to predict the lung cancer of few individuals by using KNN. This dataset includes attributes such as Age, Gender, Air Pollution, Dust Allergy, Shortness of breath, Obesity, Alcohol Consumption, Wheezing, Swallowing Difficulty, Dry Cough, Lung Cancer.

These attributes are numerical attributes which can be easily used for prediction than categorical ones. The last Attribute

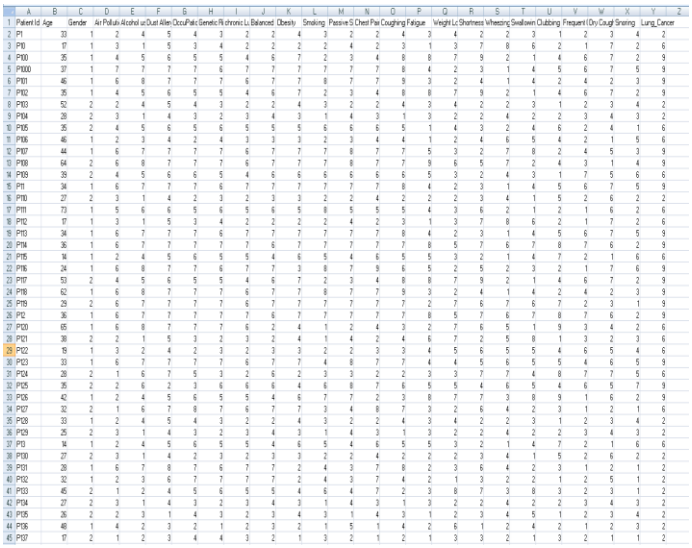
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

if this dataset is Lung Cancer which predicts the lung cancer result with three numerical values 2, 6, 9.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Lung Cancer	
2	P1		33	1	2	1	4	5	4	3	2	2	4	3	2	2	4	3	2	3	1	2	3	1	2	
3	P10		17	1	3	1	5	3	4	2	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	6	
4	P100		35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	9	
5	P1000		37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	9	
6	P101		46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	9	
7	P102		35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	9	
8	P103		52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	2	
9	P104		28	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	2	
10	P105		35	2	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2	4	6	2	4	1	6	
11	P106		46	1	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	6	5	4	2	1	5	6	
12	P107		44	1	6	7	7	7	7	6	7	7	7	8	7	7	5	3	2	7	8	2	4	5	3	9	
13	P108		64	2	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7	2	4	3	1	4	9	
14	P109		39	2	4	5	6	6	5	4	6	5	6	6	6	6	5	3	2	4	3	1	7	5	6	6	
15	P11		34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	9	
16	P110		27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	2	
17	P111		73	1	5	6	6	5	6	5	6	5	8	5	5	5	4	3	6	2	1	2	1	6	2	6	
18	P112		17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	6	
19	P113		34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	9	
20	P114		36	1	6	7	7	7	7	6	7	7	7	7	7	7	9	5	7	6	7	8	7	6	2	9	
21	P115		16	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	6	
22	P116		24	1	6	8	7	7	6	7	7	3	8	7	9	6	5	2	5	2	3	2	1	7	6	9	
23	P117		53	2	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	9	
24	P118		62	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	9	
25	P119		29	2	6	7	7	7	7	6	7	7	7	7	7	7	2	7	6	7	6	7	2	3	1	9	
26	P12		36	1	6	7	7	7	7	6	7	7	7	7	7	7	8	5	7	6	7	8	7	6	2	9	
27	P120		65	1	6	8	7	7	7	6	2	4	1	2	4	3	2	7	6	5	1	9	3	4	2	6	
28	P121		38	2	2	1	5	3	3	3	2	4	1	4	2	4	6	7	2	5	8	1	3	2	3	6	
29	P122		19	1	3	2	4	2	3	2	3	3	2	2	3	3	4	5	6	5	5	4	6	5	4	6	
30	P123		33	1	6	7	7	7	6	7	7	4	8	7	7	4	4	5	6	5	5	4	6	5	9		
31	P124		28	2	1	6	7	5	3	2	6	2	3	3	2	2	3	3	7	7	4	8	7	7	5	6	
32	P125		35	2	2	6	2	3	6	6	6	4	6	8	7	6	5	5	4	6	5	4	6	5	7	9	
33	P126		42	1	2	4	5	6	5	5	4	6	7	7	2	3	8	7	7	3	8	9	1	6	2	9	
34	P127		32	2	1	6	7	8	7	6	7	7	3	4	8	7	3	2	6	4	2	3	1	2	1	6	
35	P128		33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	2	
36	P129		25	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	2	
37	P13		16	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	6	
38	P130		27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	2	
39	P131		28	1	6	7	8	7	6	7	7	2	4	3	7	8	2	3	6	4	2	3	1	2	1	2	
40	P132		32	1	2	3	6	7	7	7	7	2	4	3	7	4	2	1	3	2	2	1	2	5	1	2	
41	P133		45	2	1	2	4	5	6	5	5	4	6	4	7	2	3	8	7	3	8	3	2	3	1	2	
42	P134		27	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	2	
43	P135		26	2	2	3	1	4	3	2	3	4	3	1	4	3	1	2	3	4	5	1	2	3	4	2	
44	P136		48	1	4	2	3	2	1	2	3	4	3	1	5	1	4	2	6	1	2	4	2	1	2	3	2	
45	P137		17	2	1	2	3	4	4	3	2	1	3	2	1	2	1	3	3	2	1	3	2	1	1	2	

Figure 1: Representation of Dataset

3.1 Libraries

Python's standard library is very extensive. The library contains built-in modules (written in C) that provide access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. The Python installers for the Windows platform usually include the entire standard library and often also include many additional components. For Unix-like operating systems Python is normally provided as a collection of packages, so it may be necessary to use the packaging tools provided with the operating system to obtain some or all of the optional components. Some of the libraries used are:

1.Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

2.Numpy: The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

3.Scikit-learn: It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with Numpy and SciPy.

4.Matplotlib: This library is responsible for plotting numerical data. And that's why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

3.2 Figures

Here, representation of architecture of KNN algorithm and use case diagram of its working in the project is shown

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

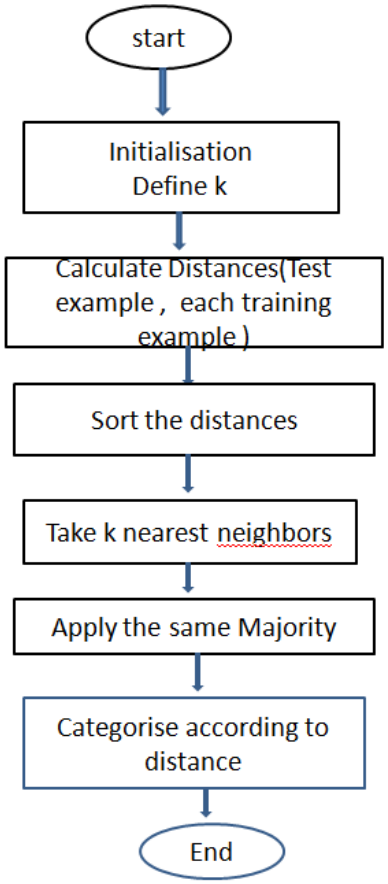


Figure1. Architecture of working of KNN algorithm with dataset imported.

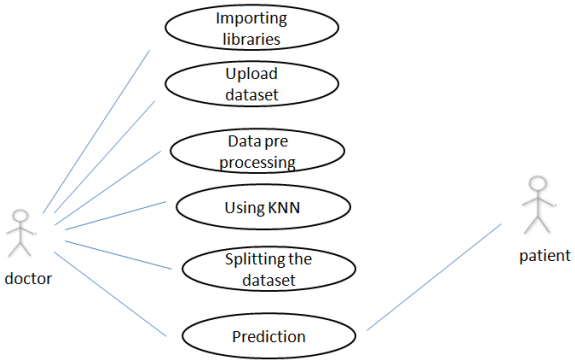


Figure2 : Use Case Diagram representing the working of the project.

Below represents some of the plots that have been used for better visualization of the predictions results of the dataset.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

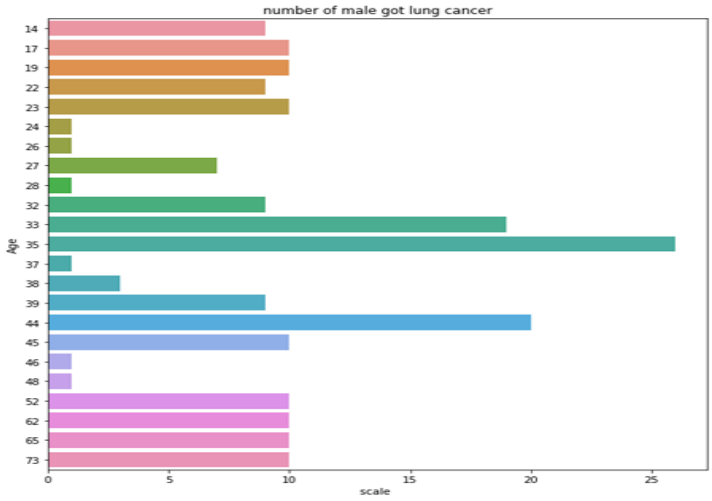


Figure 4 : Representation using Count plot

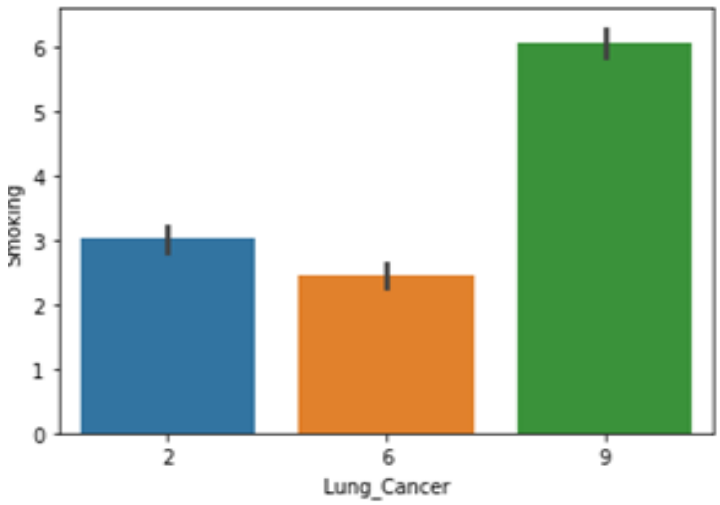


Figure 5: Representation using Bar Plot

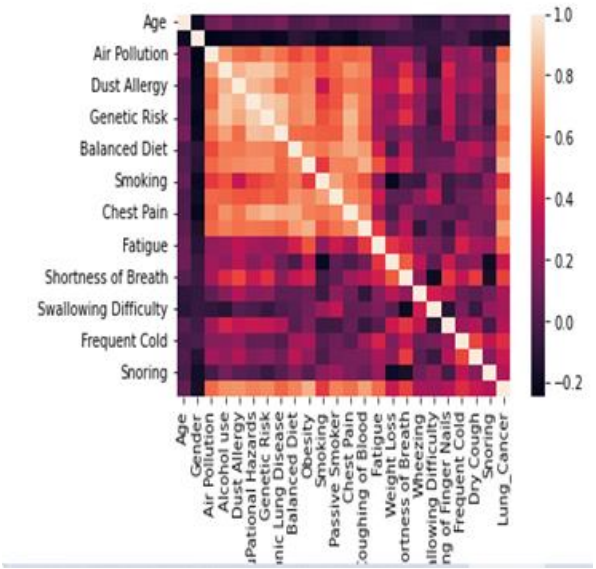


Figure 6: Representation using Heat Map

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

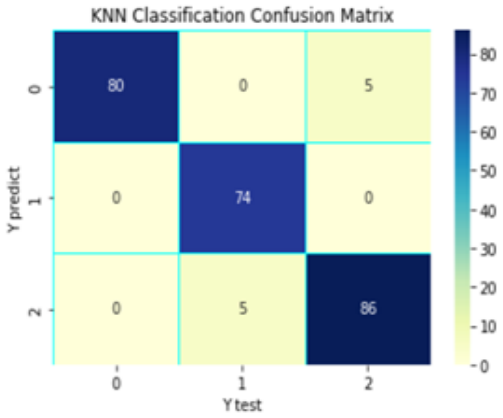
DOI: 10.17148/IARJSET.2022.94116

4 RESULTS

4.1 Confusion Matrix

A confusion matrix **is** a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

The confusion matrix obtained for this system based on the dataset imported is shown below:



4.1 Classification Report

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model. If you have never used it before to evaluate the performance of your model then this article is for you. In this article, I will take you through an introduction to the classification report in machine learning and its implementation using Python.

CLASSIFICATION REPORT

	precision	recall	f1-score	support
2	1.00	0.94	0.97	85
6	0.94	1.00	0.97	74
9	0.95	0.95	0.95	91
accuracy			0.96	250
macro avg	0.96	0.96	0.96	250
weighted avg	0.96	0.96	0.96	250

5 CONCLUSION

Lung cancer prediction using machine learning process helps us to predict the lung cancer of the patient and there are many algorithms to predict this. Some of the algorithms are SVM, Linear Regression, Logistic Regression and KNN. But even though we have many algorithms correct prediction is necessary to detect the lung cancer of the patient so that necessary precautions can be taken by the individual and incorrect prediction might leads to harmful consequences for the individual.

So, in order to predict correctly the particular algorithm needs to be accurate and hence can have a better performance on predicting the result. KNN is one of the algorithms with good accuracy of 96% and can be a good algorithm that would

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94116

be able to predict lung cancer more effectively than other algorithms. This can be better shown by importing the dataset and calculating its accuracy score and even plots are drawn for better visualization of the data.

6. REFERENCES

[1] Musa Olha, Alang "Analisis Penyakit PARU-PARU MENGGUNAKAN ALGORITMA," vol. 9, pp. 348– 352, 2017.
[2] J. Park, J. Jung, S. H. Yoon, J. M. Goo, H. Hong, and J. Yoon, "Inspiratory Lung Expansion in Patients with Interstitial Lung Disease: CT Histogram Analyses," Sci. Rep., no. October, pp. 1–13, 2018.
[3] H. Nagano, T. Kinjo, Y. Nei, S. Yamashiro, J. Fujita, and T. Kishaba, "Causative species of nontuberculous mycobacterial lung disease and comparative investigation on clinical features of Mycobacterium abscessus complex disease: A retrospective analysis for two major hospitals in a subtropical region of Japan," pp. 1–12, 2017.
[4] F. Feng, Y. Wu, Y. Wu, and G. Nie, "The Effect of Artificial Neural Network Model Combined with Six Tumor Markers in Auxiliary Diagnosis of Lung Cancer," pp. 2973–2980, 2012.
[5] J. Ramos-González, D. López-Sánchez, J. A. Castellanos-Garzón, J. F. de Paz, and J. M. Corchado, "A CBR framework with gradient boosting based feature selection for lung cancer subtype classification," Comput. Biol. Med., vol. 86, pp. 98–106, 2017.
[6] H. Cheng, C. Jin, J. Wu, S. Zhu, Y. J. Liu, and J. Chen, "Erratum to: Guards at the gate: physiological and pathological roles of tissue-resident innate lymphoid cells in the lung (Protein & Cell, (2017), 8, 12, (878- 895), 10.1007/s13238-017-0379-5)," Protein Cell, vol. 8, no. 12, p. 932, 2017.
[7] A. M. Kwon, "A rank weighted classification for plasma proteomic profiles based on case-based reasoning," BMC Med. Inform. Decis. Mak., vol. 18, no. 1, pp. 1–10, 2018.
[8] H. R. Shahraki, S. Pourahmad, and N. Zare,? "Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data," vol. 2017, 2017.
[9] Novita Mariana and dkk, "PENERAPAN ALGORITMA K-NN (nearest Neighbor) UNTUK DETEKSI PENYAKIT (KANKER SERVIKS) Novita Mariana, Rara Sriartati Redjeki, Jeffri Alfa Razaq Abstrak," vol. 7, no. 1, pp. 26–34, 2015.
[10] O. Musa, "Sistem Informasi Pemetaan Pendidikan Menggunakan Algoritma Data Mining," vol. 01, pp. 26–32, 2015.

