International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Lung Cancer Detection

P. Uma Devi, T. Vamsi Krishna Sai, K. Sivani, N. Vijay
UG Students, Dept of CSE, Kallam Haranadhareddy Institute Of Technology, Andhra Pradesh, India

ABSTRACT
One of the cancers that claims the most lives globally is lung cancer. In this endeavor, we'll use a machine learning algorithm to detect cancer early and begin treatment. KNN and Decision Tree algorithms are used to forecast cancer accurately. In this study, we predict and categorize a dataset of lung cancer patients using the sklearn and pandas libraries. The dataset is trained using feature scaling and dataset slicing methods. After that, accuracy score and confusion matrix are used to forecast the precision of the outcome.

Keywords: KNN, Decision Tree, Machine Learning (ML).

1. INTRODUCTION

Lung cancer is one of the leading causes of death and a major public health problem in many nations. Although those who smoke are at a higher chance than those who do not can also develop lung cancer. As you smoke more and for longer periods of time, your chance of developing lung cancer rises**.**

Robotics, autonomous driving, medical research, and other areas all use machine learning. Supervised learning employs pre-existing input and output values. The results are predicted judgements for unknown feature vectors. This demonstrates that prediction precision is the most crucial feature for machine learning (ML). Cancer is one application for machine learning.

Every year, 14 million individuals receive cancer diagnoses worldwide, 266 thousand of which are for breast cancer. A 1% rise in forecast accuracy would have a negative impact on at least 140,000 people. The objective was to develop a machine learning algorithm that, compared to existing algorithms, was quantitatively better. The University of California-Irvine Machine Learning Repository has access to each of these databases. Similar to prediction accuracy, learning curves are diagrams that demonstrate how accuracy rises as more training data is made available. Faster learning requires less training data to achieve the same level of prediction precision.

One of the deadliest illnesses on the globe is lung cancer. More individuals die from lung cancer each year than from breast, brain, or prostate cancer combined.

The leading cause of cancer-related death in individuals between the ages of 45 and 70 is lung cancer. More people die from lung cancer each year than from breast, colon, and prostate cancer put together, making up more than 25% of all cancer-related fatalities. Lung cancer detection utilizes numerous current technologies.

Air travels down the trachea, also known as the windpipe, when it reaches the nose or mouth. It then reaches a location known as the carina. The windpipe splits in half at the carina, creating two main stem bronchi. The left lung is stimulated by one, and the right lung by the other. The channel-like bronchi continued to divide into smaller bronchi and then even smaller bronchioles from that point, just like tree limbs do.

The alveoli are the final stop for this pipework that is constantly diminishing. Lung malignant growth expands when cells sever or come from a tumor and move through the lymphatic system and circulatory system to remote areas of the body. For medical professionals, the disclosure of these cells is a fundamental concern.

2. LITERATURE SURVEY

Arnaud A. van Riel, and Mathilde explained the presented engineering, which consists of various channels of 2-D Convolutional networks, for which the production is consolidated using a deeply

committed fusion method to induce the greatest classification. However, the morphological variety of knobs is continuously more pronounced. The article "Unsupervised Deep Implanting for Clustering Analysis" goes into great detail about clustering. Many data-driven application areas depend critically on clustering, which has historically been viewed as a collection of distinct abilities and computations. Centered learning models have been applied to clustering in a relatively small number of studies. Misclassification of any picture, however, does not have the desired effect.

In a paper on computed tomography imaging, Mario Buty1, Zi Yue Xu1, and Ming Chen Gao released their findings. It is a typical technique for locating and assessing lung carcinoma. Although these criteria are primarily subjective and arbitrarily defined, expert qualitative evaluations on various parameters characterizing a nodule's appearance and form are frequently used in clinical practice to assess the malignancy of lung nodules. Without human assistance, our method obtains a Dice-Sorensen Coefficient (DSC) of 63:44 percent, which is higher than attained without close monitoring. In this process, it offers less precision. This research suggested an autonomous lung cancer detection system that shortens diagnosis time while increasing accuracy and yield. For human interpretation and analysis, the amount of information in MR images is simply too large. The diagnosing method entails four stages. Using a probabilistic neural network, the Normal and Irregular were classified.

Rivansyah Suhendra uses a Support Vector Machine model. It is one of the traditional techniques for grouping 'n' features that is most effective. The identification of a hyperplane completes the categorization. To divide data into two groups, SVM runs a linear distinct hyperplane through a dataset. Any number is separated using the hyperplane. The most effective hyperplane intensifies the edge. The edge is the space between a few closely spaced objects and the hyperplane. The hyperplane is controlled by these neighboring points. The best edge predictor is this one. The hyperplane's edge can be expanded with the aid of maximal edge detectors. As it summarizes the error, this is the finest option.

For the purpose of comparing ML algorithms, Parmeshwar R. Hegde suggested a model about binary and multiclass classification. Binary categorization According to a classification rule, category is the process of dividing the components of a support into two categories (understanding which bundle each one belongs to). Examples of situations where it is necessary to determine whether a person has an abstract trait, a designated trademark, or a binary classification include the following: The proximity of the illness is the determining factor in medical testing to determine whether a patient has a particular disease or not. A "leave or fall behind short" test method, such as determining whether a goal has been attained or not during a go/no-go meeting.

3. PROBLEM STATEMENT:

To classify the dataset and provide the highest accuracy of the result, we use K Neighbors Classifier & Decision Tree Algorithm.

4. PROPOSED SYSTEM

KNN and Decision Tree Algorithm are used to identify lung cancer. Decision Trees are supervised learning techniques used for regression and categorization. The objective is to acquire representative decision rules inferred from the data features in order to build a model that predicts the value of a target variable.

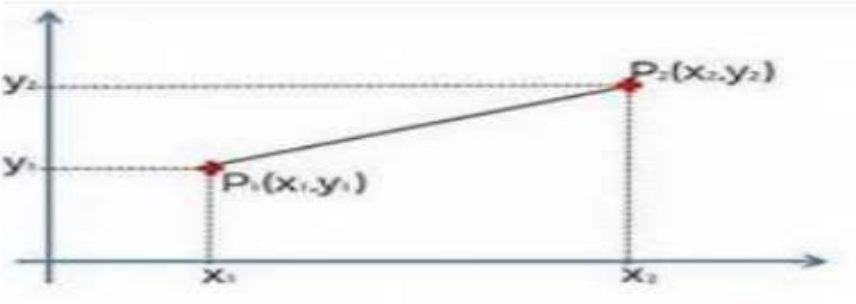
The k-nearest neighbor algorithm, also referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to make classifications or forecasts about the grouping of a single data point. Although it can be applied to classification or regression problems, it is usually used as a classification algorithm because it relies on the idea that similar points can be found close to one another.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

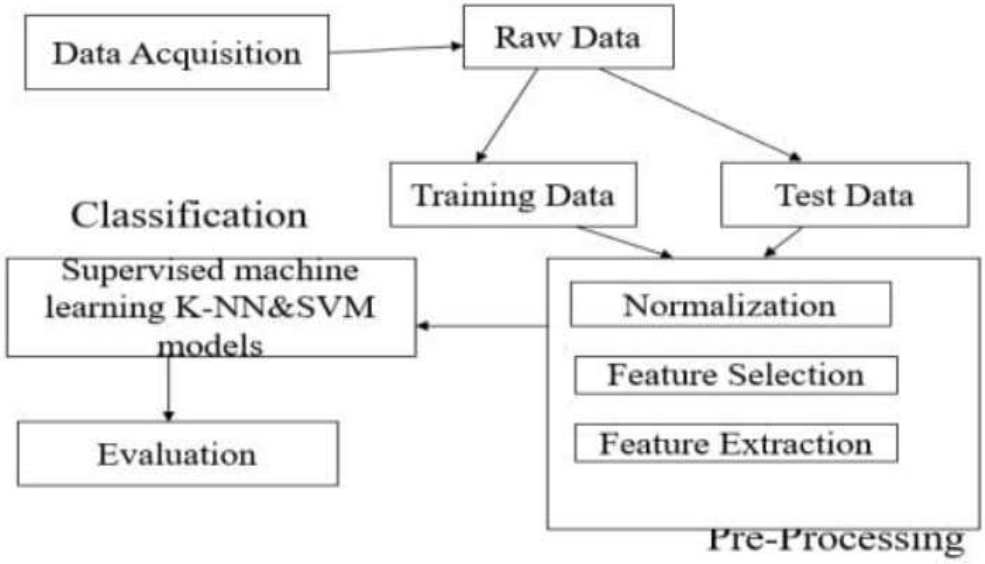
A class label is chosen for classification problems based on a majority vote, meaning that the label that is most frequently reflected around a particular data point is used. Although the word "plurality vote" is technically appropriate here, the term "majority vote" is more frequently used in literature.

Measurements for Distance in Euclid (p=2): This distance metric, which can only be applied to real-valued vectors, is the most widely used one.

The straight line between the query location and the other point being measured is calculated.



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



4.1 Block Diagram Of Proposed System

5. METHADOLOGY

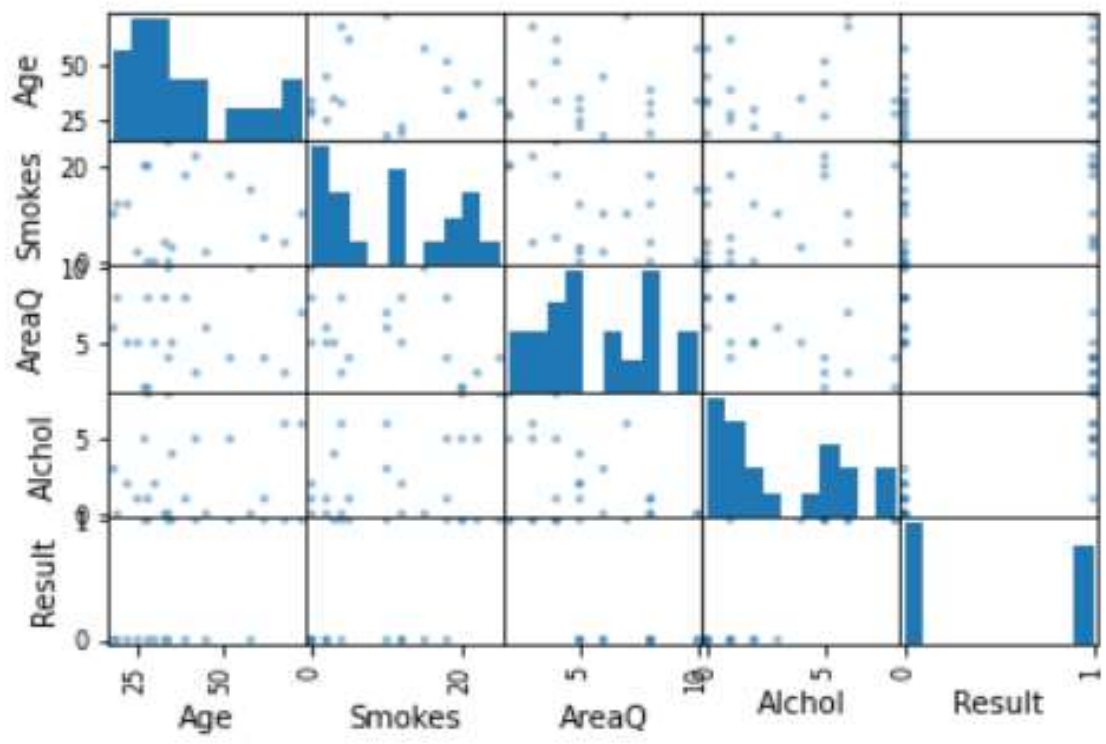
1. Slicing the dataset
2. Feature Scaling
3. Train the dataset
4. Use of Confusion matrix, f1 Score and accuracy score
5. Prediction

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

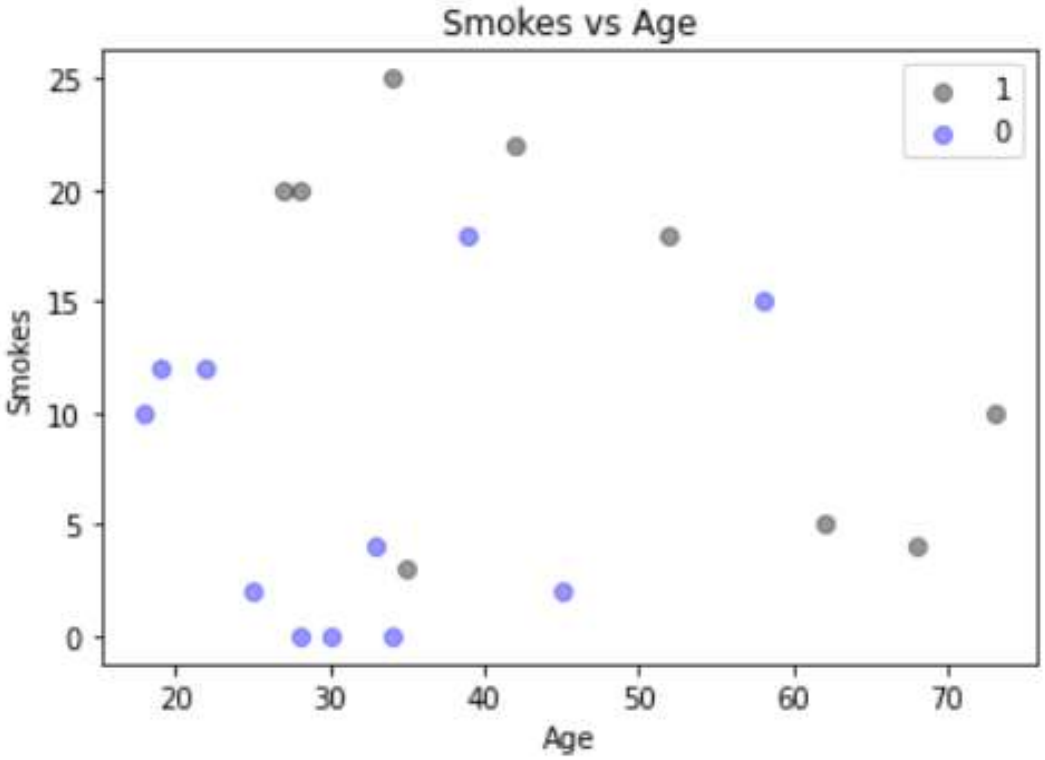
6. RESULTS AND DISCUSSION

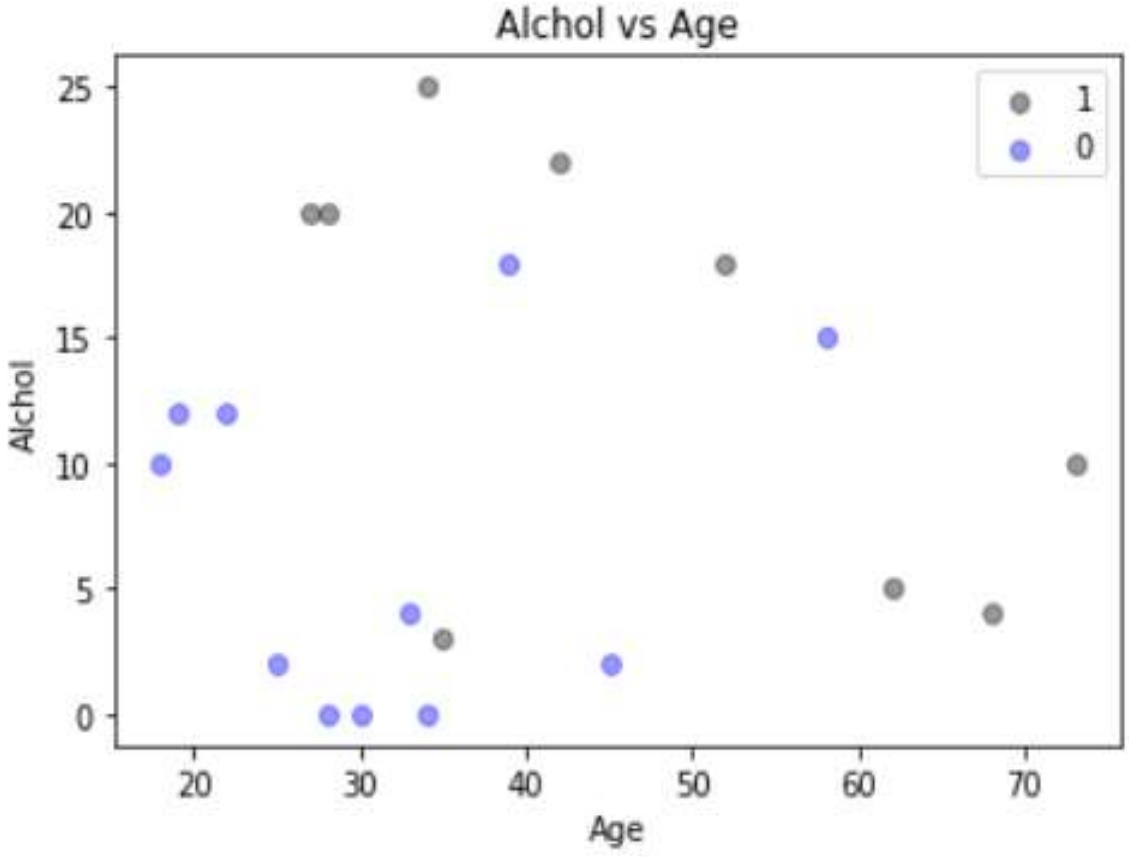
Dataset:
20

	Name	Surname	Age	Smokes	AreaQ	Alchol	Result
0	john	wick	35	3	5	4	1
1	john	constanty	27	20	2	5	1
2	camela	andreson	30	0	5	2	0
3	alex	tellers	28	0	8	1	0
4	diego	macadona	68	4	3	6	1

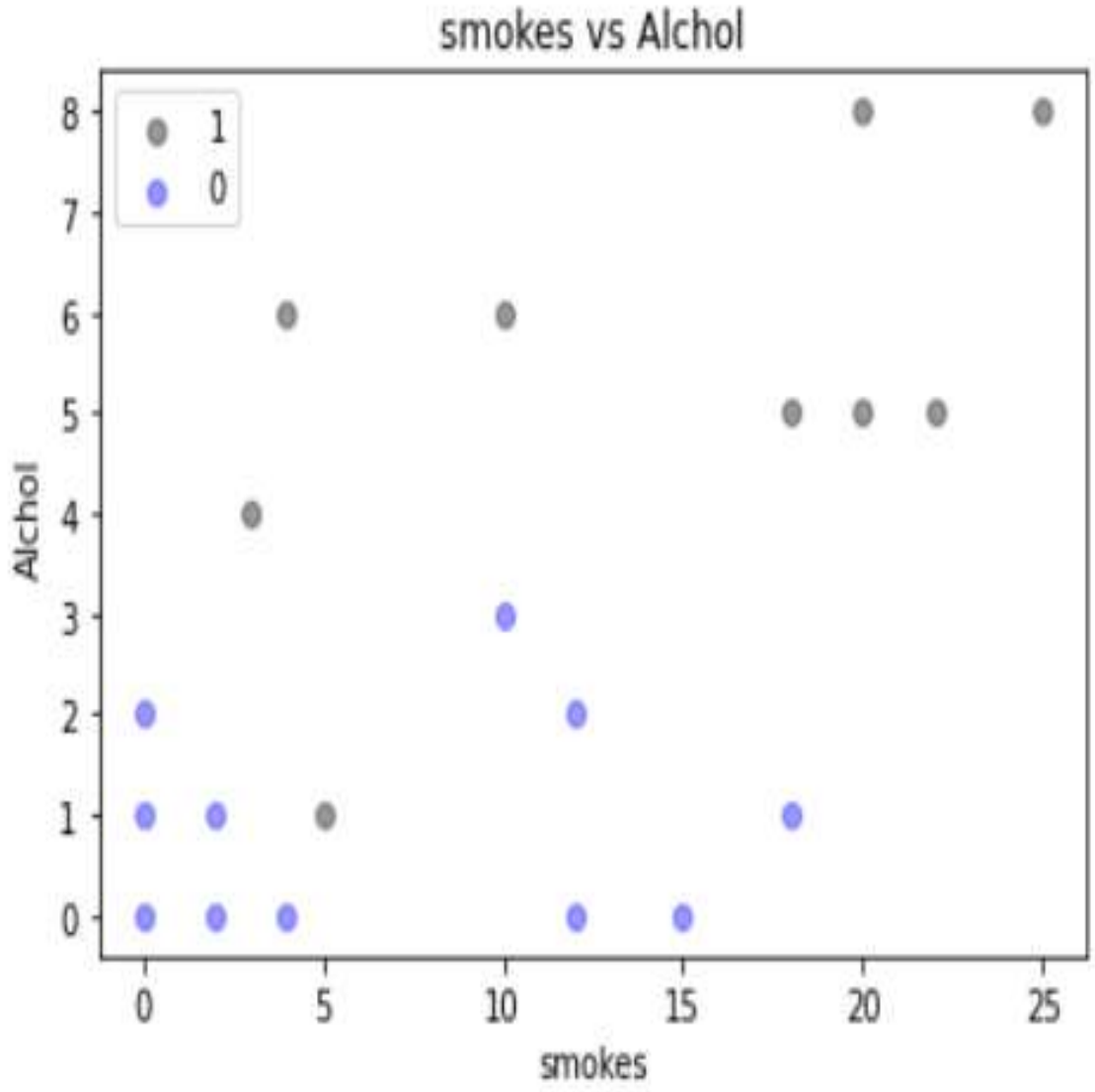


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362



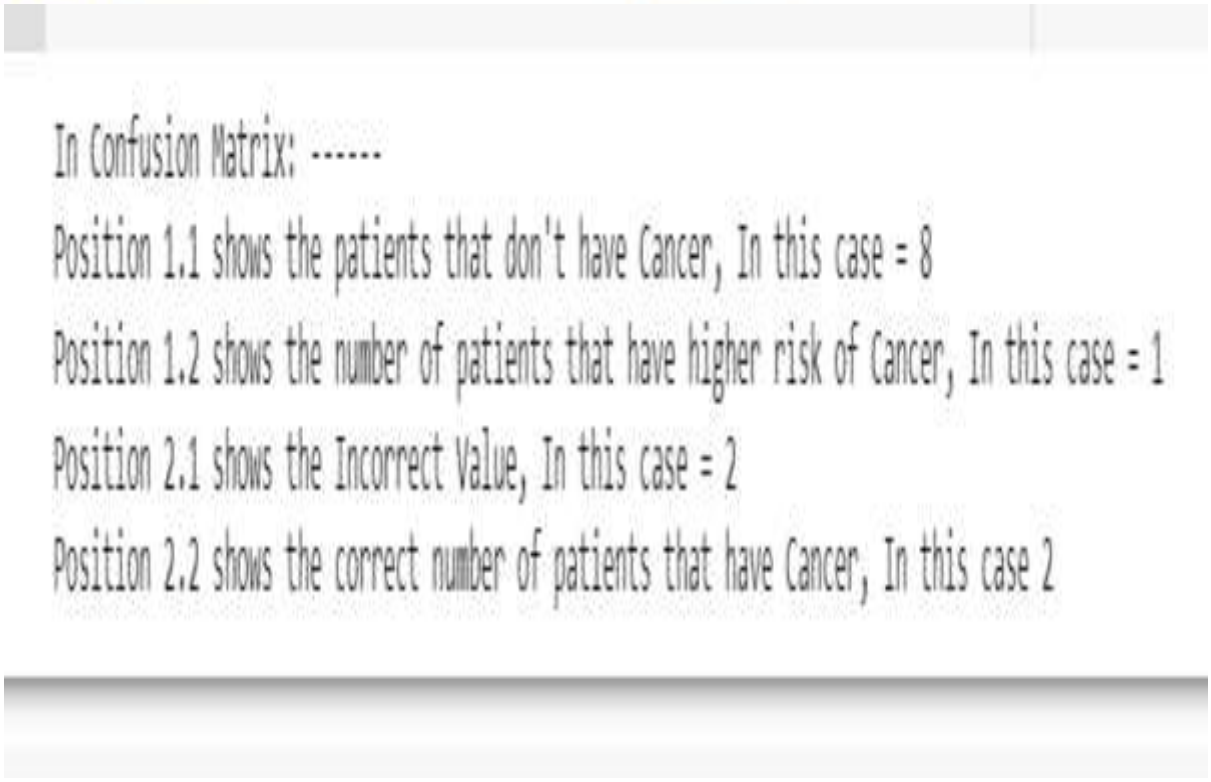


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362




```
Confusion Matrix:
[[1 0]
 [2 1]]
```

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



7. CONCLUSION

We processed the dataset to differentiate the effected patient and its level of the growth of the cancer by the machine learning system. Here it presented to an approach to find best accuracy of the cancer result to assist the radiologist and for the future enhancement.

REFERENCES

1. Rendon-Gonzalez, E., & Ponomaryov, V. (2016, June). Automatic Lung nodule segmentation and classification in CT images based on SVM. In 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW) (pp. 1-4).
2. K. Kuan, M. Ravaut, G. Manek, H. Chen, J. Lin, B. Nazir, C. Chen, T. C. Howe, Z. Zeng, and V. Chandrasekhar, ‘‘Deep learning for lung cancer detection: Tackling the kaggle data science bowl 2017 challenge,’’ 2017,
3. F. Ciompi, K. Chung,S. J .vanRiel,A .A .A .Setio, P . K. Gerke, C. Jacobs, E. T. Scholten, C. SchaeferProkop, M. M. W. Wille, A. Marchianò, U. Pastorino, M. Prokop, and B. van Ginneken, ‘‘Towards automatic pul- monary nodule management in lung cancer screening with deep learning,’’ Sci. Rep., vol. 7, no. 1, Jun. 2017, Art. no. 46479.
4. Sun W, Tseng T-LB, Qian W, Zhang J, Saltzstein EC, Zheng B, et al. Using multiscale texture and density features for near-term breast cancer risk analysis. Med Phys 2015;42(6Part1):2853–62.
5. Hossain MS, Muhammad G. Cloud-Based Collaborative Media Service Framework for HealthCare. Int J Distrib Sens Netw 2014;10(3):858712.
6. Amin SU, Alsulaiman M, Muhammad G, Mekhtiche MA, Shamim Hossain M. Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. Future Generat Comput Syst 2019;101:542–54.

7. H. Jiang, He Ma, W. Qian, M. Gao and Y. Li, "An Automatic Detection System of Lung Nodule Based on Multi-Group Patch-Based Deep Learning Network,"2018. [online].
8. B. A. Skourt, A. El Hassani, and A. Majda, ''Lung CT image segmentation using deep neural networks,'' Procedia Comput. Sci., vol. 127, pp. 109–113, Jan. 2018
9. Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques. "Interational journal of computer science and Information Technologists 4.1 (2013): 39-45.
10. P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks,"2019.
11. H. Yu; Z. Zhou and Q. Wang, "Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model, "2020.
12. S. Shadroo, A. M. Rahmani, "Systematic survey of big and data mining in internet of things," 2018.