International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Lung Cancer Detection using Deep Learning

Anmol Kalia[1], Komal Ahuja[2]

[1] *anmolkalia200@gmail.com*, M.Tech. Scholar, Department of Computer Science and Engineering Global Research Institute of Management & Technology, Radaur Kurukshetra University, Kurukshetra (Haryana)

[2] Assistant Professor, Department of Computer Science and Engineering Global Research Institute of Management & Technology, Radaur Kurukshetra University, Kurukshetra (Haryana)

ABSTRACT

To uncover hidden configurations from oversized records, data mining technique combines numerical analysis, machine learning, and databank knowledge. Additionally, the expansion of frequent submissions in the thriving healthcare industry makes the removal of medical records an important investigative component. Lung disease emerges as the determining factor after all global fatalities are taken into account. When determining whether a patient may be at risk for lung disease, medical professionals face an interesting debate because it calls for broad problem-solving and specialized knowledge. To construct models using Data Mining alone or in combination with computational approaches, numerous changes have been made. The data withdrawal method, which employs a combinational strategy from arithmetical research, device knowledge, and catalogue expertise to find uncovered configurations, is exposed to large datasets. The objective of the automated system is to examine all or a portion of the patient's data and offer a qualitative evaluation of the likelihood that the patient has lung disease. This probability is expressed using a natural ordinal number between 0 and 4, where 0 denotes that the patient is healthy and 4 denotes that there is a strong likelihood that he is suffering from disease. The results will be judged noteworthy if the classifier can accurately predict the patient's state of health in 85% of the cases or if its accuracy in absolute terms exceeds 85%.

Keywords: Random Forest, Area under Curve, Naive Bayes, Decision Tree, Convolutional Network

INTRODUCTION

This section will outline some patterns and facts concerning cancer, specifically lung cancer, to help the reader appreciate the significance of early and accurate identification of this disease. Cancer, as described in [9], is a condition characterized by uncontrolled development and division of malignant cells, which results in tumors masses, as seen in figure 1.1. Furthermore, some of these cells may move throughout the body via the blood and lymph systems, attacking various organs.

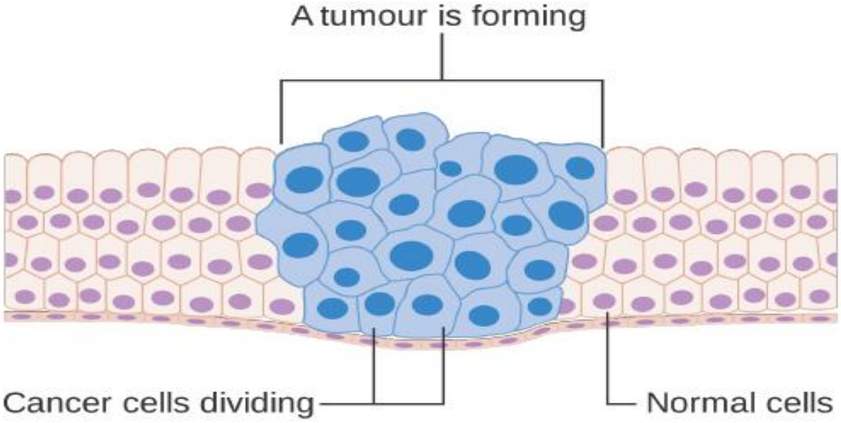


Figure 1.1: Representation of cancer cells spreading through healthy tissue*

**** Credits to Cancer Research UK / Wikimedia Commons [1]***

In a normal state, cells follow a certain life cycle that consists of various stages in which a newborn cell develops, replicates, and eventually dies. The final phase is known as apoptosis, and it was initially described by Kerr, Wyllie, and Currie in [2]. Furthermore, as noted in [2], it is part of the natural cell turnover process, in which timeworn cells are substituted by fresh cells. When cancer arises, the regular life cycle fails in favor of an aberrant behavior in which cells do not die when they should and multiply at a quicker rate than usual. In this scenario, cancer patients have solid lumps known as tumors [3]. Tumors are categorized as malignant or benign: benign tumors are lumps that can reach enormous dimensions but do not travel to new areas and do not grow again once removed. Malignant ones, on the other hand, can infect various tissues and could grow back after removed. There are two forms of lung cancer, rendering to the American Cancer Society (ACS): slight cell lung cancer (SCLC) and non-slight cell lung cancer (NSCLC). The former is the more frequent, while the latter is less communal, secretarial for about 10-15% of circumstances [4]. This investigation will focus on NSCLC, which may be further classified into subgroups such as adenocarcinoma, which begins in the glands, and squamous cell carcinoma, which begins in the skin that composes the airways [5]. Although lung cancer was not too frequent before the jump of the 20th era, it is now solitary of the foremost reasons of demise globally. Some argue that rising cigarette usage is one of the reasons why this disease is becoming more prevalent. Tobacco has been widely used across the world for centuries, as stated in [6], but cigarettes now include not just tobacco, but also chemical additives and other potentially dangerous elements. Throughout the twentieth century, the incidence of lung cancer has steadily climbed from the 1930s through the 1950s, when it became the leading cause of death among males [7]. In addition to smoking, other known factors of this rise in the prevalence of pathology include air pollution, nutrition, and genetic predisposition. Indeed, as stated in [8], more than 1,700,000 fresh cases and further than half a million cancer demises are predictable in the US in 2018. According to data collected by the ACS during the last period, the occurrence degree of cancer in women has remained steady while decreasing by about 2% per year in males. More specifically, concentrating on lung cancer, the predicted number of new cases in the US will be approximately 230,000, roughly evenly distributed between male and female, with a projected mortality rate of around 150,000. Some development has been completed in current years in lowering the death fatality rate. Prevention, therapies, and early diagnosis, which are the focus of this thesis, have enabled such remarkable accomplishments. Considering that the overall year (2007-2013) relative persistence degree for lung cancer at all stages in the United States was about 18%, [9] it is evident that early and correct detection of tumors lesions is critical. The prospect of an early diagnosis can allow the opportunity to select the most appropriate therapy, therefore boosting the chance of survival.

DATA MINING

This strategy mines undiscovered configurations from large records by combining numerical study, machine understanding, and databank knowledge. Furthermore, the rise in frequent submissions in the thriving healthcare industry makes medical records withdrawal a key investigation component. After accounting for all of the deaths that occur across the world, disease becomes the deciding factor. When determining if a patient has lung illness, medical experts have an unusual challenge since it necessitates broad testing and the presence of knowledge [10] [11].

Several changes have been made to construct models using Data Mining alone or in conjunction with computational approaches. Large datasets are subjected to the data pullout technique, which employs a combinational strategy from arithmetic research, device knowledge, and catalogue experience to discover previously undiscovered configurations [12].

Data mining is one of the most advanced analytical techniques accessible today; it also involves mathematical mockups, measurable methods, and machine knowledge practices (such as neural networks or decision trees). Data quarrying is a prominent subject of research that may be utilized across many disciplines. Data withdrawal is commonly discarded in areas such as communications, banking, health, education, and commercial marketplaces in order to recruit new customers, combat disease, cut expenditures, and increase benefits.

Many terms, such as information removal, facts extraction from statistics, statistics/arrangement research, and statistics analyzing, may all signify the same thing as data mining. The purpose of statistics removal is to find patterns in data that may be utilised to explain and predict behavior. Data mining models are composed of a series of instructions, computations, or complex "transmission purposes." They may be divided into two basic sets established on their goals, which are as follows:

- Supervised / Analytical Prototypes.
- Unsupervised Prototypes.

Supervised / Analytical Prototypes

The goal of supervised, predictive, directed, or focused modelling is to estimate or forecast the standards of a continuous numeric feature. These models include both a production arena or objective arena and contribution arenas or characteristics. Input fields are also known as predictors since they assist the model in selecting a forecast utility for the production arena [13].

2.2 Unsupervised Prototypes

In unverified or objectiveless mockups, there is merely an input area; there is no production arena. Undirected arrangement recognition denotes that it is not impacted by a certain external aspect. These models attempt to detect data patterns in the input fields [14].
Unsupervised prototypes comprise:
1. Cluster mockups.
2. Relationship and series mockups.

RESULTS AND DISCUSSION

3.1 Lung Nodules Segmentation

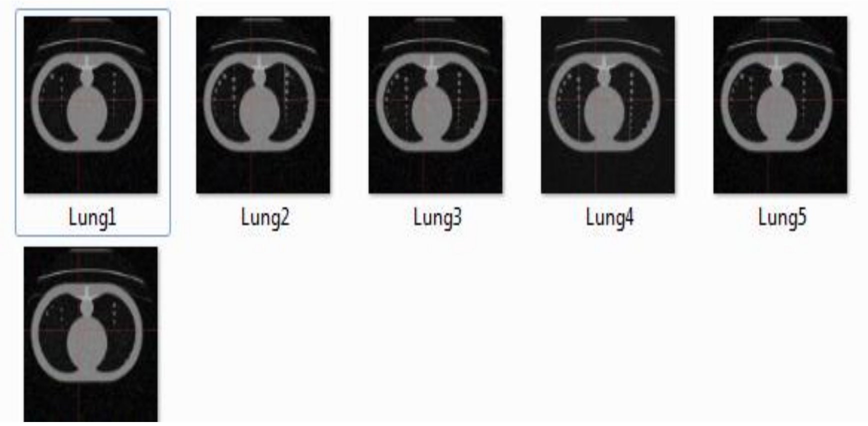


Figure 3.1: Sample Dataset

More than 1,000 patients' CT images are available in the LIDC-IDRI database, and each nodule has been reviewed by four qualified radiologists. The dataset's photos are displayed in part in figure 3.1.

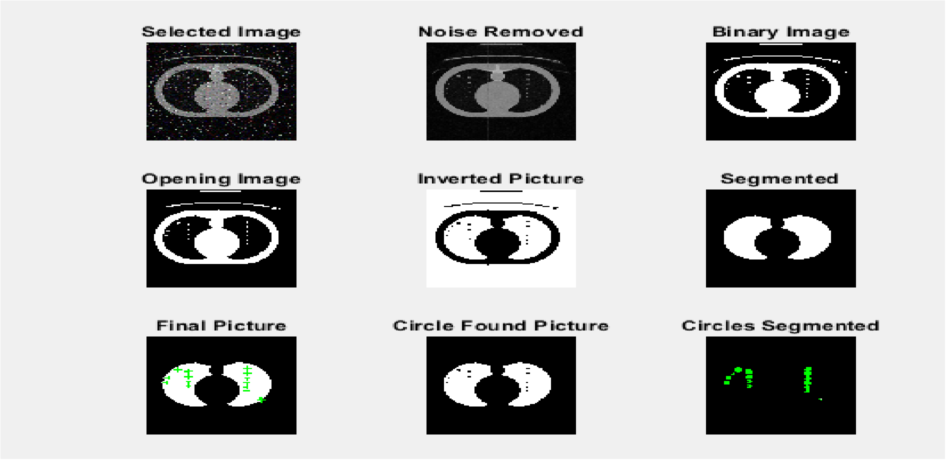


Figure 3.2: Results of Lung figure 1

Numbers of Circles = 14

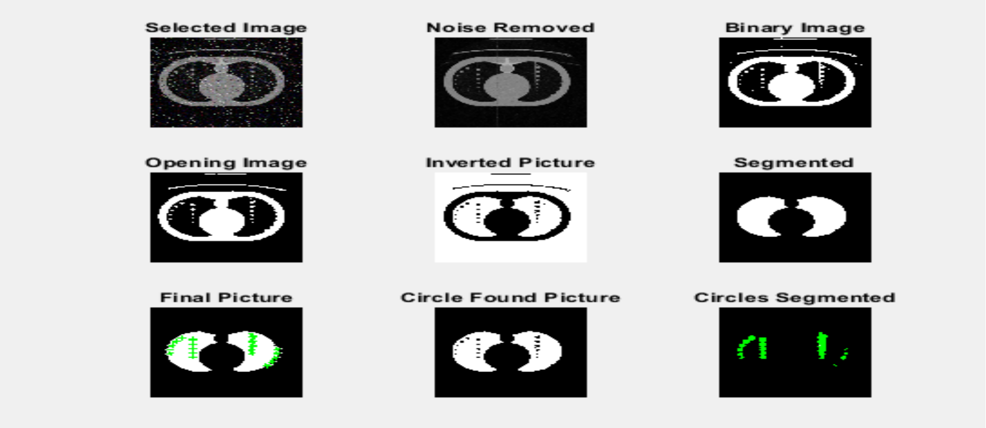


Figure 3.3: Results of Lung figure 2

Numbers of Circles = 21

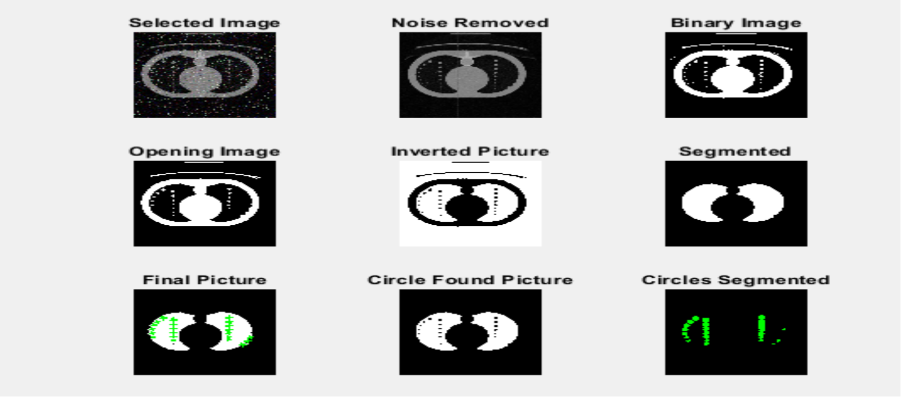


Figure 3.4: Results of Lung figure 3

Numbers of Circles = 22

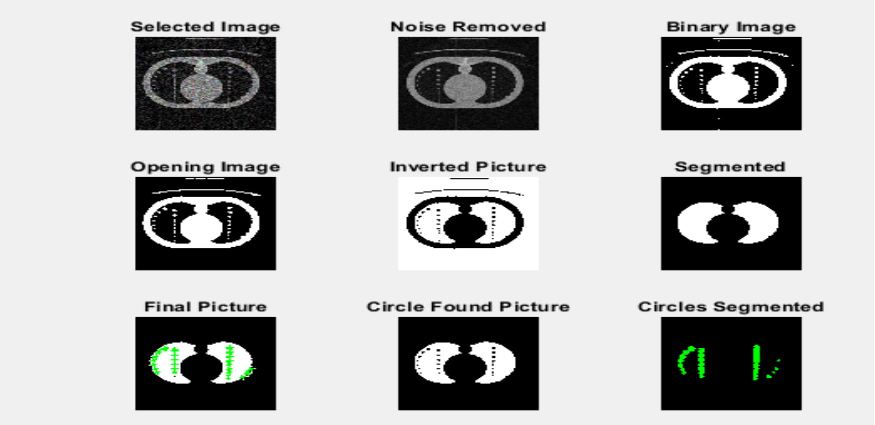


Figure 3.5: Results of Lung figure 4

Numbers of Circles = 26

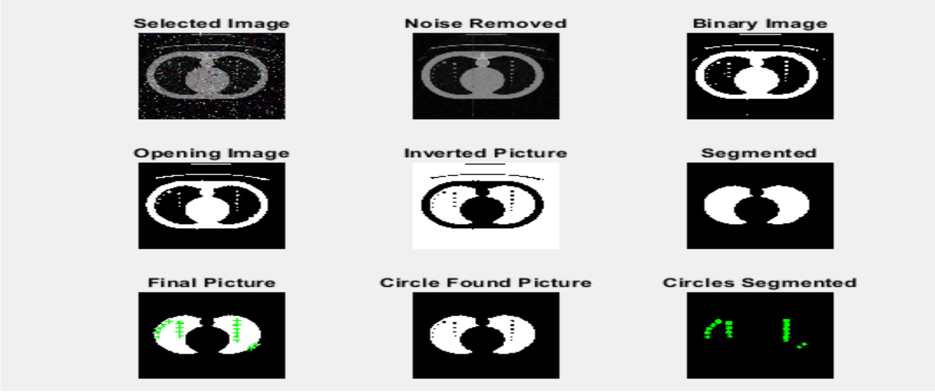


Figure 3.6: Results of Lung figure 5

Numbers of Circles = 17

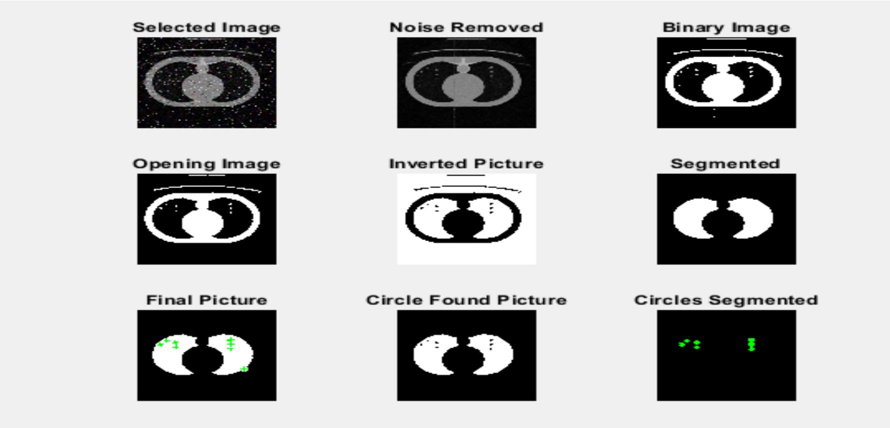


Figure 3.7: Results of Lung figure 6

Numbers of Circles = 8

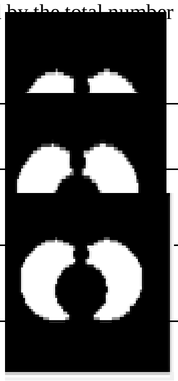
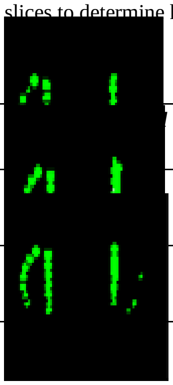
Figure 3.2 to Figure 3.7 show the results of pre-processing (noise removal), segmentation and circles segmented.

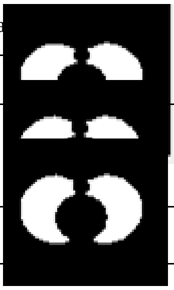
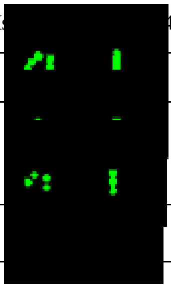
The segmented picture and the quantity of circles found in each image are displayed in table 3.1.

For the purpose of automatically learning feature representation for retrieval, we created a Feed Forward Neural Network model. For every CT scan, the model was trained to identify a circle.

Every lung nodule consists of many slices, each of which has a feature vector connected to it. The distances between every slice of nodule A and every slice of nodule B are totaled together and divided by the total number of slices to determine how similar the two nodules are.

Table 3.1: Discovery of circles

Figure No.	***Original image***			***No. of Circles***
	Lung1			**14**
	Lung2			**21**
	Lung3			**22**

	Lung4			**26**
	Lung5			**17**
	Lung6			**8**

3.2 Classification Results

Table 3.2: Classification Results

Output	**Details**
Number	**Lung Disease Diagnosis: 0** **= No.** **1= low probability.** **2 ≥ 1** **3 ≥ 2** **4 = high probability.**

The results obtained at the end of the cross validation procedure are the following:

Table 3.3: Outcome of Various Algorithms

Algorithms	**Value**
Majority	**0.545**
Decision tree	**0.557**
Random forest	**0.587**
Naive Bayes	**0.534**
CN2	**0.522**

It is apparent in figure 3.8 how low the acquired accuracy numbers are and how poorly the objective (accuracy 85%) is met. The class's characteristics, however, point to a considerable dataset adjustment that could have a significant positive impact on classification accuracy. In fact, the class qualitatively shows both the proportion of stenosis in the major cardiac vessels as well as the qualitative chance of having the condition. A percentage of stenosis less than 50% exists when the class is 0, a percentage of stenosis higher than 50% exists when the class is greater than or equal to 1, and the percentage of stenosis increases as the integer value of the class increases. Furthermore, it is clear from the literature that an occlusion greater than 50% poses major health hazards to the patient, who is diagnosed with lung disease and requires surgery to reopen the blocked conduit.

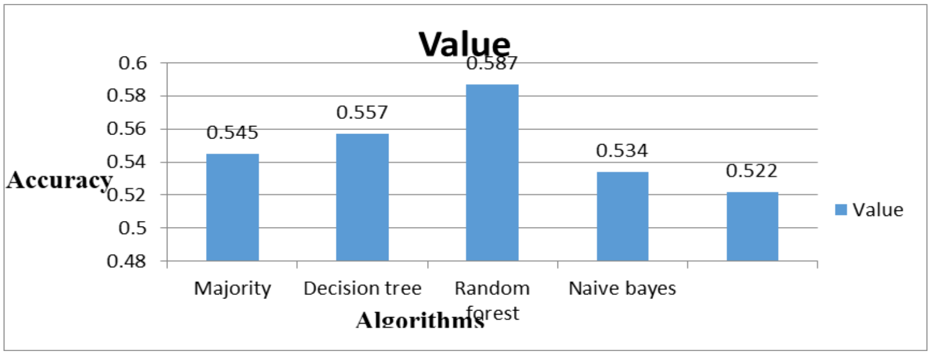


Figure 3.8: Outcome of Various Algorithms

Table 3.4: Classification Results

Output	**Details**
Number	Diagnosis: 0: stenosis of major lung nodule < 50%, 1: stenosis of major lung nodule >50%.

An evaluation of the frequencies of the class has been carried out, which could indicate which quality indices to calculate in the modeling and evaluation phase.

Table 3.5: Classification Results

Number	**Frequency Relative**
0	**0.54**
1	**0.46**

These frequencies suggest that the choice of patients within the dataset was not carried out randomly on a sample of subjects belonging to the population. It would be unthinkable, in fact, that 46% of the population is affected by lung disease as shown in figure 3.9.

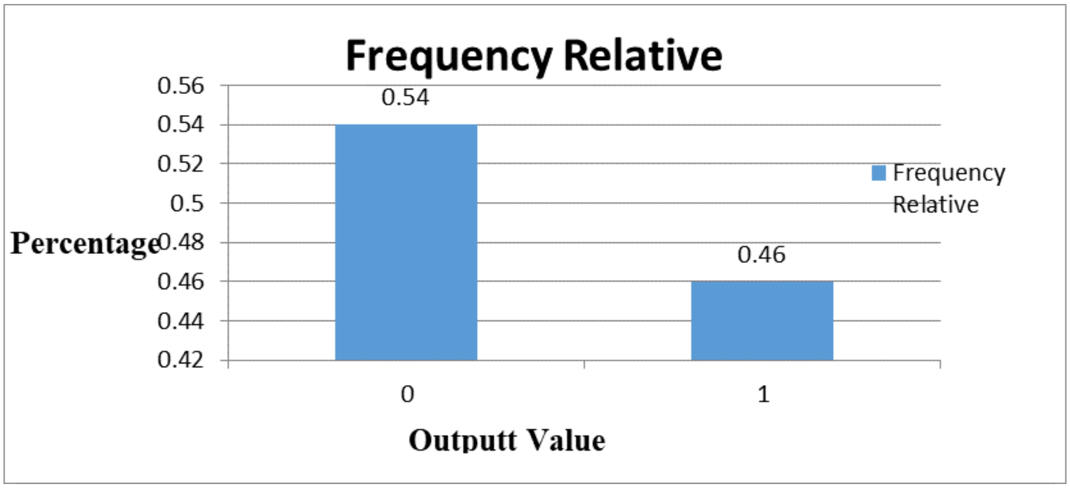


Figure 3.9: Outcome of Frequency Relative

The testing procedure is performed on the 75% of data remaining from the feature selection. Quality indices of the classifiers obtained after a 10-fold cross validation procedure were obtained. In particular, indices such as sensitivity, specificity, accuracy and AUC were used.

Random Forest

Having 9, p is equal to 3 because P is set to be equal to the square root of the number of characteristics. Additionally, if the number of examples is 8 or less, stop the growth of the trees by selecting the value that produces the best AUC.

Table 3.6: AUC Estimation

Size Limit	**AUC**
<=6	**≤0.889**
7	**0.891**
8	**0.911**
9	**0.891**
10	**0.901**
>=11	**0.897**

CN2

The optimum rule length is ten. It would seem that the results in terms of AUC are unaffected by the size of the beam. Consequently, it was decided to set this dimension to 5, which is its default value. Entropy was also selected as the evaluation metric due to the existence of continuous data. In reality, even while a Laplace estimate is preferred; using one would necessitate preventive discretization, whereas for the entropy, discretization is done to maximize it. Finally, it is found that by selecting 0.17 as the minimum coverage while analyzing the rules' minimal coverage that the AUC is improved.

Table 3.7: Evaluating Minimum Coverage of Rules

Cover	**Approx.**
0.12	**0.673**
0.17	**0.697**
0.23	**0.677**
> 0.23	**≤0.655**

The following average outcomes were achieved following the cross validation process:

Table 3.8: AUC Evaluation of Various Algorithms

Algorithms	**Sensitivity**	**Specificity**	**AUC**
Naïve Bayes	**0.785**	**0.863**	**0.899**
Decision Tree	**0.731**	**0.817**	**0.873**
CN2 Rule Inducer	**0.736**	**0.832**	**0.841**
SVM	**0.764**	**0.821**	**0.778**

The evaluation parameter is thought of as the AUC. We begin by selecting them based on the AUC value in order to determine which classifier (or classifiers) should be regarded as the best for such a situation. The algorithms with maximum accuracy are chosen in accordance with the basic goal, so Nave Bayes as shown in figure 3.10.

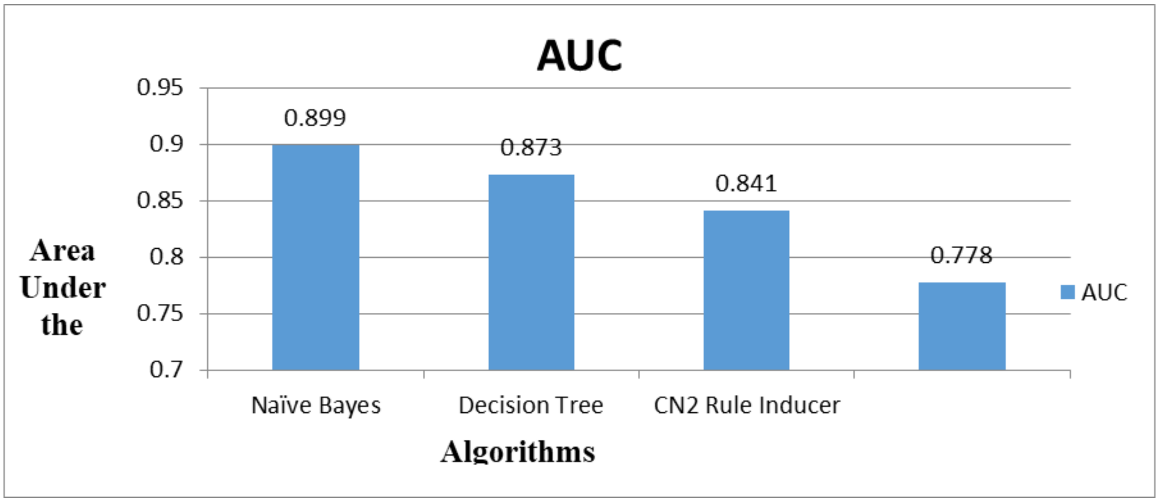


Figure 3.10: AUC Evaluation of Various Algorithms

CONCLUSION

For volumetric CT data sets, image processing and imaging techniques may improve the radiologist's capacity to identify small lung knots. For instance, it has been proposed that the reconstruction of CT scans with brief interscan intervals and the more meticulous description of nodules using photographs than film-based examination technique might expedite the identification of inconsequential nodules. Nodule discovery is a method for working with images. The objective is to identify the precise position and form of the illogical formations in the lungs called as nodules. A "nodule" is a small, radioactively more concentrated lung wound or worm-shaped damage that is attached to the lung border, the pleura, than the lung parenchyma.

With notable controllability, the local histogram mapping job may be chosen using the CLAHE (Contrast Limited Adaptive Histogram Equalization) approach. Using this method, the image is divided into pertinent portions, which are subsequently equalized on the histogram. It modifies the brightness standards of the image while maintaining a nonlinear technique to optimize the contrast for every pixel in the image. More specifically, this thesis developed a system and associated software tool that can use digital chest imaging as input and use its features to automatically identify the lungs and the likely existence of tumor lesions.

The obtained accuracy, which reflects the experimental results, ranges from 85 to 95%. The identification of tumors in slices that do not include the lungs has led to false positives; in the future, it might be possible to automatically eliminate a specific percentage of images to increase accuracy and lower the risk of false positives in areas that are categorically not the lungs.

If the chosen learning algorithm were provided in a clear and useable manner, physicians might find it to be of tremendous assistance. One such implementation is developing a straightforward graphical user interface that can be used on desktop or mobile devices and allows users to enter values for the nine variables under consideration using drop-down menus. The system provides the diagnosis and relative probability after the entry is complete so the doctor can compare them. Once a diagnosis has been requested, the program automatically sends the values entered to regional, national, and international databases so that the algorithm can be periodically updated with more data, resulting in optimal performance. Furthermore, the availability of datasets pertaining to various geographic regions makes statistics particularly efficient due to their simplicity in integrating with the most common information systems.

The acquired accuracy in the experimental findings ranges from 85 to 95%. In certain instances, certain slices devoid of lungs have been mistakenly identified as having a tumor (false positives); going forward, it will be possible to automatically eliminate a portion of the images to improve accuracy and lessen the likelihood of creating false positives in regions that are unquestionably not the lungs.

REFERENCES

1- Daniel Perez and Yuzhong Shen, "Deep Learning for Pulmonary Nodule CT Image Retrieval - An Online Assistance System for Novice Radiologists", 2017 IEEE International Conference on Data Mining Workshops

2- Xue Chen Li , Linlin Shen and Suhuai Luo, "A Solitary Feature-Based Lung Nodule Detection Approach for Chest X-Ray Radiographs", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 22, NO. 2, MARCH 2018

3- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I.: Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13:8{17, 2015.

4- Beachey, W.: Respiratory care anatomy and physiology: foundations for clinical practice. Elsevier Health Sciences, 2018.

5- Broaddus, V. C., Mason, R. C., Ernst, J. D., King, T. E., Lazarus, S. C., Murray, J. F., Nadel, J. A., Slutsky, A., and Gotway, M.: Murray & Nadel's Textbook of Respiratory Medicine E-Book. Elsevier Health Sciences, 2015.

6- Hall, J. E.: Guyton and Hall textbook of medical physiology e-Book. Elsevier Health Sciences, 2015.

7- Corrin, B. and Nicholson, A. G.: Pathology of the Lungs E-Book: Expert Consult: Online and Print. Elsevier Health Sciences, 2011.

8- Cancer de_nition. https://www.cancer.gov/publications/dictionaries/ cancer-terms/def/cancer. Accessed: 2018-03-14.

9- Kerr, J. F., Wyllie, A. H., and Currie, A. R.: Apoptosis: a basic biological phenomenon with wideranging implications in tissue kinetics. British journal of cancer, 26(4):239, 1972.

10- Prakash Ramani, Nitesh Pradhan, “Classification Algorithms to Predict Lung Diseases—A Survey “, Computer Vision and Machine Intelligence in Medical Image Analysis, pp 65-71, 2020

11- Sarangam Kodati, R. Vivekanandam G. Ravi, “Comparative Analysis of Clustering Algorithms with Lung Disease Datasets Using Data Mining Weka Tool” , Soft Computing and Signal Processing, pp 111-117, 2019

12- T.Nagamani, S.Logeswari, B.Gomathy, “Lung Disease Prediction using Data Mining with Mapreduce Algorithm “, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-3, 2019

13- Akitoshi Shimazaki, Daiju Ueda, Antoine Choppin, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method", Scientific Reports, 2022, 12:727

14- Imran Nazir, Ihsan ul Haq, "Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion", Hindawi Journal of Sensors Volume 2023, Article ID 6683438, 19 pages https://doi.org/10.1155/2023/6683438