Lung Cancer Recognition Using CT-Scan with NCA-XG Boosting & KNN

Likitha Dara- 700743525

Email: lxd35250@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Sabrina Shaik- 700732583

Email: sxs25830@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Ayesha Farhana-700735341

Email: axm53410@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Prudhvi Mahesh Meka- 700738978

Email: pxm89780@ucmo.edu

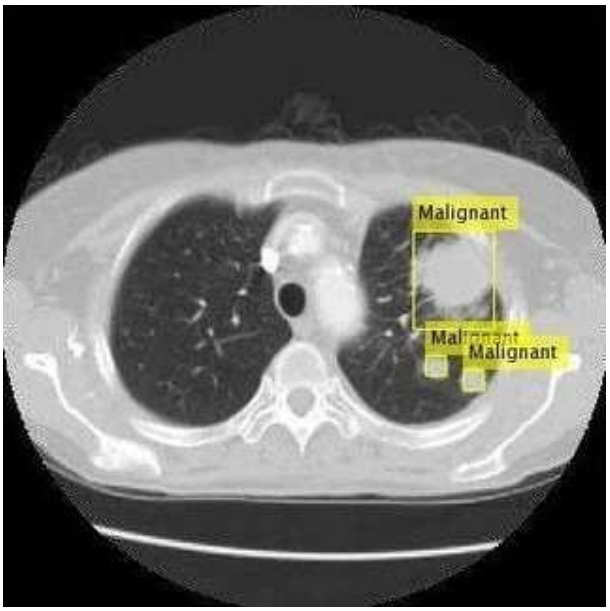
University Of Central Missouri, MO, USA.

Department Of Computer Science.

ABSTRACT

Cancer of the lungs occurs when abnormal cells develop in the lining of the bronchi or alveoli in the lungs. Exposure to carcinogens like those found in tobacco, radiation, and asbestos causes this abnormal growth and multiplication of cells. The primary method for classifying lung cancer is based on the origin of the cancerous cells. The condition can also be diagnosed at the molecular level now, thanks to advances in medical technology. Using this technique, scientists search for abnormalities in the DNA and proteins made by cancer cells. Mutations in EGFR, ALK, KRAS, and ROS1 are only a few of the hundreds of molecular diagnoses available. How rapidly a disease develops and spreads may depend on its molecular subtype. They can also forecast how the disease will react to chemo, targeted therapy, and immunotherapy. Molecular cancer diagnosis allows doctors to create personalized treatment strategies with the best probability of success against each patient's cancer. Lung cancer often produces no noticeable symptoms in its early, treatable stages. However, the disease may harm surrounding tissue, interfering with normal lung function and causing symptoms like haemoptysis (coughing up blood), shortness of breath, or pain as it develops.

Lymphatic metastases are a common route of lung cancer dissemination. Draining from our tissues as a transparent fluid, lymph carries immune cells that aid in the body's defence against illness. It moves through your body via lymphatic vessels. The lymph nodes are connective structures between lymph veins, and they are tiny and bean-shaped. Cancer cells that have spread to the lymphatic system often get caught in them. The circulation is another route by which cancer cells might travel to other organs. Stage IV lung cancer, also called metastatic lung cancer, occurs when the disease spreads to other parts of the body. Lung cancer is still used to describe cancer that has spread to other organs. Treatment for lung cancer differs greatly depending on whether or not the cancer has progressed to the lymph nodes or other organs.



KEYWORDS

Lung-cancer, Computerized Tomography, Machine Learning, Datasets, Algorithm, KNN Classifier, AdaBoost Classifier.

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN

INTRODUCTION

The uncontrolled and abnormal multiplication of cells that originates in one or both lungs and subsequently spreads throughout the rest of the body is the definition of lung cancer. Lung cancer may begin in either lung. Cancer of the lung may develop in either lung. It is possible for cancer to develop in either one of the lungs as a result of smoking.

In healthy tissues, abnormal cells do not proliferate; nevertheless, when these cells are present in sick tissues, they quickly proliferate and grow into tumours. Normal cells do not proliferate. In healthy tissues, abnormal cells do not grow in any significant number. Secondary lung cancer, as opposed to primary lung cancer, which begins in one section of the lungs and does not spread to other areas of the body, begins in another part of the body and travels through the body until it reaches the lungs. Primary lung cancer begins in an area of the lungs and does not spread. The term "primary lung cancer" refers to the kind of lung cancer that affects the most people. Males are diagnosed with primary lung cancer at a rate that is about two to three times greater than the incidence that occurs in females. When present in a patient's body, the early symptoms of lung cancer are frequently symptomatic of the beginning of the illness in the body of the patient. This is because lung cancer tends to develop slowly over time. As the number of people living in today's industrialized cities who are affected with lung ailments continues to rise, there is a rising need for cutting-edge diagnostic processes that are both accurate and prompt. This is a result of the increasing need for advanced medical technology. According to professionals in the medical area, the practice of smoking, which has an effect on the cells of the lungs, is the primary factor in the development of lung cancer. Because cigarette smoke is composed of potentially hazardous substances such as carcinogens, a person who smokes cigarettes will feel the effects of this on the tissues of their lungs almost immediately after beginning to smoke. Cigarette smoking is related with an increased risk of developing lung cancer because of the smoke that is inhaled. Smoking cigarettes is associated with an increased risk of acquiring lung cancer.

The use of deep learning as a possible solution to these problems is a possibility. A model was developed by the researchers at North-western University and the National Institutes of Health using de-identified chest CT screening data from 45,856 participants who participated in the National Lung Screening Trial. After that, the performance of the model was compared to that of six radiologists who all had board certification.

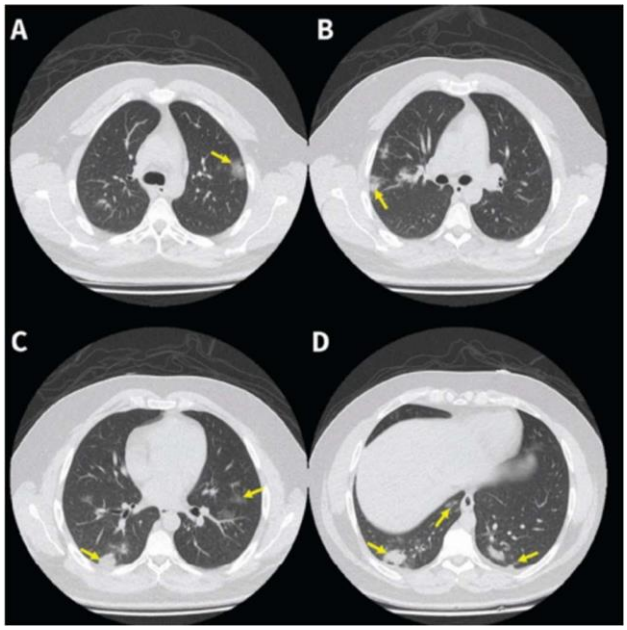
The model performed as good as or better than human radiologists when just a single CT scan was used as a diagnosis tool. The algorithm's performance reached a state-of-the-art level when it reached 94.4 percent AUC. In addition, the model reduced the number of false positives by 11% while simultaneously reducing the number of false negatives by 5%.

The model has the potential to detect moderately malignant tissue in the lungs of a patient in addition to determining the overall level of lung cancer that a patient has. The fact that the deep learning system may be able to include information from earlier scans is a benefit, given that the progression of suspicious tissue may be indicative of the presence of cancer.

The researchers who conducted this study believe their findings show that AI and deep learning have the potential to significantly improve lung cancer screenings.

Even though lung cancer screenings are very important, only around 2% to 4% of individuals who are eligible for them in the United States are actually getting them.

Researchers revealed that machine learning algorithms could detect breast cancer cells that had spread to surrounding lymph nodes. This discovery is an essential component in determining how to treat patients in the most effective manner. The machine learning models outperformed earlier automated techniques and achieved results that were on par with those of human medical professionals.



REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN

MOTIVATION

While lung cancer is the most common form of the disease in men, it is the third most common form of cancer in women. For men, lung cancer is the most prevalent form of the disease. Lung cancer is the most common form of the disease found in males and occurs when abnormal cells in the lungs become cancerous and multiply, causing tumours to form. It is absolutely essential to begin screening for lung cancer at an earlier stage in order to reduce the overall number of deaths that can be attributed to this illness on a global scale. Because the symptoms of lung cancer don't typically present themselves until the disease has already progressed to an advanced stage, it is essential to detect the disease at an early stage by utilizing any and all medical imaging technologies that are currently available. The purpose of this investigation is to develop a classification system for lung cancer that is capable of carrying out diagnostic activities in an automated manner during the earlier stages of the disease. In order to carry out the evaluation, the imaging modalities of the lungs that are provided by computed tomography (CT) are utilized. In order to classify the data that was obtained from this investigation, the NCA-XG Boosting and KNN algorithms were utilized. The VGG19 classifier was applied to the input photographs of the lung after they had been pre-processed, and this was done in order to determine which of the pictures were actually of the lung. This classifier makes use of an adaptive boosting strategy, and the foundation of that strategy is the pretrained approach.

MAIN CONTRIBUTIONS & OBJECTIVES

• Simple to improvise - The software is easy to create and accuracy will rise as we obtain more picture samples.
• Database administration - Convenient data management techniques in a single library format.
• Application – Computerized Tomography, or CT scan, is the most effective approach for detecting illness at the tissue level, and with the assistance of modern machine learning algorithms, it is also the simplest and most time-efficient.
• Availability - CT scans are now accessible in all primary and secondary health care facilities.
• Adaptability - It is not essential to master machine learning methods in order to use CT scans, since the technology is currently in use and has a flexible user interface that requires some training.

RELATED WORK

Machine learning improves AI by enabling in-component learning from experience or extrapolation of data. The software runs complicated decision-making methods as it grows and learns from prior acts. Below is a summary of the published studies that investigate the use of machine learning techniques in the detection of lung cancer.

- Examines the prediction of post-operative life expectancy in lung cancer patients using predictive data mining methods to evaluate Decision Tree, Naive Bayes, and Artificial neural network algorithms. A stratified 10-fold cross-validation comparison study was performed on the aforementioned methods, and the accuracy of each classifier was determined.

- This paper compares classification algorithms for the identification of brain tumours. Using volumetric and location information, the total accuracy rate was computed based on 2 classification classes, including logistic regression and quadratic discriminant, and 3 classification classes, including Linear SVM, Coarse Gaussian SVM, Cosine KNN, and Complex and median tree.

- In this article, distinct findings are generated for each classifier on the collected lung cancer dataset. The KNN, SVM, NN, and Logistic Regression classifiers were implemented, and the appropriate accuracy rates were obtained. Support Vector Machine offers the greatest degree of precision, at 99.3%. The application of the suggested technique to medical datasets assisted physicians in making more accurate decisions.

- Several segmentation techniques, including Naive Bayes and Hidden Markov Model, were addressed. It is explained in detail how and why different segmentation algorithms are used in the identification of lung cancer.

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN

- Instructions on how to create a basic flowchart for an algorithm that may identify brain tumours were provided. Classification strategies for two distinct varieties of data mining approaches were discussed.

1. Statistical methods: Naive Bayes and the Support Vector Machine

2. Decision tree and neural network data compression techniques

3. Many data sets were the subject of discussion.

We have the BRATS Dataset, the OASIS Dataset, and the NBTR Dataset.

In other case, the authors conducted an experiment to examine the influence of referral course and side effects on delays in a quick outpatient diagnostic programme for patients with suspected lung cancer, as well as whether delays were connected with disease stage and prognosis. There has been a comprehensive investigation of the features of tumours, their structure, and the many postponements that have happened. For this investigation, a total of 565 patient restoration schematics were gathered. 51% of the participants had lung growths, whereas 8.5% had a variety of injuries, and 19.6% of the participants had non-life-threatening radiological abnormalities. In the instance of haemoptysis, first-line wait times were much shorter than in other situations. During the rulemaking procedure, a RODP was created to aid the analytical process. According to estimates, the great majority of patient postponements are due to delays in the first and second treatment lines.

They examined many methods for evaluating lung growth. Among these were the use of artificial neural networks, image processing, linear dependency analysis (LDA), and self-organizing maps (SOM). In conclusion, it is suggested that support vector machines be employed as a tool for characterisation. When using machine learning, support vector machines may be used to examine data and identify patterns. At the outset of their investigation, [10] created a method to identify lung development. Data pre-processing is performed in this way to begin the image enhancement procedure. When the datasets are prepared for testing under information mining and neural systems, which are both essential for differentiating amongst rehabilitative treatments, they may be tested at this stage. Researchers were able to achieve the required outcome by using back-propagation neural networks to categorize information images as malignant or harmless (BPNN). When making a diagnosis, medical personnel choose which stage of cancer will be most beneficial to them.

This study used network-based biomarker discovery and quality set improvement methodologies to identify and validate traits associated with lung cancer development and associated pathways. In addition to the characteristics anticipated by past research in these areas, they discovered that the data showed a vast array of novel and unexpected characteristics associated with hypothesized physiological capacity in smoking. developed a network-based technique for dealing with observable confirmation of smoking, classifying between the qualities associated with lung tumour survival and those associated with non-smoking groups, and identifying all the qualities associated with lung tumour survival and non-smoking groups. It has been shown that a six-quality smoking score may predict the risk of lung enlargement and the probability of survival. If this quality mark is used, smokers may be able to see and recognize lung expansion.

To explore lung expansion, they used information mining and streamlining techniques to obtain insights from a huge number of datasets. It may be used to identify and exploit patterns of malignancy in datasets. These patterns, which are identified in databases, may then be utilized to forecast the fate of a disease based on the precise therapy cases kept in databases. The authors showed the identification of neural system enlargement using computed tomography images and a previously described computer-aided diagnostic (CAD) order approach. To recreate the lung, CT scan highlights were stitched together and then reconstructed. The mean, standard deviation, skewness, and kurtosis, as well as the fifth and sixth central moments, were used to identify whether or not the data included malignant cells. To enhance grouping, forward- and backward-feeding neural networks are utilized to organize objects.

There is no doubt that the authors have been working for quite some time on the use of different artificial intelligence algorithms for illness detection and medicine provision. Using an artificial neural network (ANN), breast cancer data may be analysed. Similar to ANNs, multilayer feedforward neural networks may be used to identify the onset of lung cancer utilizing data from microarrays and the UCI machine learning library. The preparation of the system employs the back-propagation rule. Using cross approval, it is possible to compare datasets with variable amounts of hidden layers and hubs linking to the same dataset. If an event from the UCI dataset (breast cancer) happens, it is anticipated that the different mixes of masked layers and linked hubs would increase the accuracy of this framework. As the number of hubs and hidden layers in the NCBI dataset continues to grow, the accuracy of the analysis improves. Using a comparable brain system, it is possible to forecast the course of a patient's condition. This is possible via the use of an automated decision system.

PROPOSED FRAMEWORK

- Detailed design of features

The vast majority of machine learning engineers spend a significant amount of time pre-processing or cleaning the data before beginning the process of building a model from the ground up. A few examples of data pre-processing procedures include the identification and treatment of outliers, the handling of missing values, and the elimination of undesired or noisy data.

Pictures that have been reduced to the most fundamental level of abstraction are what are referred to as "image pre-processing," which is the same thing as "image processing." Entropy is a measure of information, and this method does not add to the amount of picture information that is stored in the image; rather, it decreases the amount of image information that is stored in the image. The objective of pre-processing is to improve the quality of the image data by removing undesired distortions and enhancing certain visual properties that are necessary for the subsequent processing and analysis of the image. Pre-processing is done before the actual processing and analysis of the image. The actions that make up pre-processing may be broken down into the two categories that are detailed down below. The pre-processing steps may be broken down into two different categories:

1. Image Filtering and Segmentation.

2. Fourier Transform and image restoration.

Pre-processing is necessary for input CT scans in order to cut down on the amount of noise that is already there and to prepare the pictures to be used in following processes such as image segmentation. As a direct consequence of this, the input pictures will exhibit less distortion, and the relevant characteristics of the inputs will be emphasized. There are four different kinds of cancer nodules that are taken into account in the input database. These include the well-circumscribed type, the juxta-pleural type, the vascularized type, and the pleural-tail type. Primary and secondary phase cancer nodules are also taken into account. The CT picture of the lung harbouring cancer that was utilized as input may be seen in Figure 2 above.

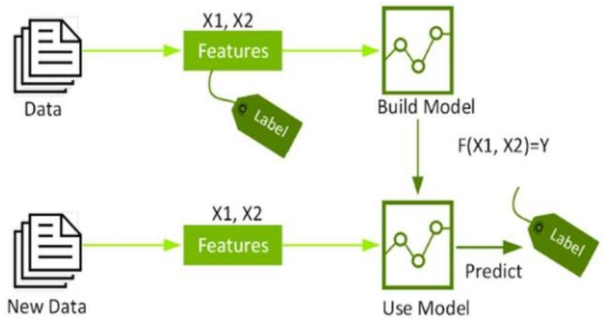
NCA-XG Boosting:

The decision-tree-based ensemble Machine Learning approach known as XG Boost makes use of a gradient-boosting framework to increase both the accuracy and the speed of its predictions. XG Boost was developed by Microsoft Research and was given its current moniker in honour of the founder of the corporation. When it comes to solving prediction issues that include unstructured data, artificial neural networks often outperform all other algorithms and frameworks (images, text, etc.). Decision tree-based algorithms are widely acknowledged as the most effective method for the management of structured or tabular data sets that range in size from relatively small to relatively medium.

Two common types of ensemble tree algorithms are XG Boost and Gradient Boosting Machines (GBMs). Both of these ensemble tree approaches improve the performance of weak learners (in general, CARTs) by using the gradient descent architecture.

Since it has helped people and teams win almost every structured data competition on Kaggle, people and teams have developed a unique love for the tool known as XG Boost. Participants in these competitions are asked to submit data, after which statisticians and data miners compete to see who can construct the most reliable models for analysing and interpreting the data. Python was first used in the development of XB Boost, and then R took over. Because of the overwhelming demand for its services, XG Boost has begun providing package implementations for a variety of languages, including Java, Scala, Julia, and Perl, amongst others. The popularity of XB Boost has grown among the Kaggle community as a direct result of the new implementations that have been made possible as a result of these.

XB Boost has been integrated with a variety of other tools and packages, including scikit-learn for Python and caret for R users. Due to its integration, distributed processing frameworks such as Apache Spark and Task may also be utilized with XB Boost. This year, InfoWorld presented XB Boost with its Technology of the Year award, which it earned with flying colours.



An algorithm is used to find patterns in a labelled data set for model training, and then that model is used to a new dataset to make label predictions. While machine learning algorithms may be constructed to handle raw data, the initial stage is the feature selection phase. When working with massive volumes of high-dimensional data, the Neighbourhood Component Analysis (NCA) technique is an effective tool for choosing important feature points. The closest neighbour feature weighting technique is the foundation of this algorithm, and it will be discussed in further depth below. The NCA technique has been shown to be successful on several microarray datasets for malignancies such as colon cancer, brain tumour, leukaemia, lung cancer, and prostate cancer when used as a feature selection tool. An efficient classifier is essential for automated cancer subtype categorization.

As a machine learning classifier, we apply the ensemble-based XB Boosting algorithm. XB Boost is a quick and efficient implementation of performance-optimized gradient-boosted decision trees. Implementation of the method was intended to be as efficient as feasible in terms of processing time and memory use. One of the design objectives was to maximize the utilization of available resources for training the model.

KNN CLASSIFIER

Because it is sometimes helpful to include more than one neighbour, this method is also referred to as k-Closest Neighbour (k-NN) Classification, whereby k closest neighbours are utilized to decide the class. Due to the \fact that training examples are needed at runtime, i.e., they must be in memory \sat runtime, it is also known as Memory-Based Classification. Since induction is delayed until execution time, this approach is classified as Lazy Learning.

Classification is also known as Example-Dependent Classification or Case-Reliant \classification as it is based on training instances. Following this evaluation of distance, the k nearest neighbours is selected. The k nearest neighbours may then be used in a variety of ways to determine q's category. Assigning the majority class to the nearest neighbours of the query is the quickest and easiest approach.

When selecting the query's class, it is often smart to assign greater \weight to the query's closest neighbours. A fairly general approach is to use distance-weighted voting, in which neighbours vote on the class of the query case with votes weighted by the inverse of their distance from the query.

ADABOOST CLASSIFIER

AdaBoost (Adaptive Boosting) is a popular boosting technique that combines many weak classifiers to build a single robust classifier. Yoav Freund and Robert Schapiro are the authors of

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN

AdaBoost. By collecting several weak classifiers and learning from their misclassified objects, we may create a robust model. Classifiers include Decision Trees, Logistic Regression, and other techniques. What do weak classifiers entail? A weak classifier performs better than random guessing but is incapable of classifying things. A bad classifier may predict that those beyond the age of 40 cannot run a marathon while those under the age of 40 can. Even with an accuracy of above 60%, you would misclassify several data items!

AdaBoost is capable of learning from any categorization and recommending a more precise model. For this reason, it is called the "best out-of-the-box classifier." Consider AdaBoost and Decision Stumps. Decision Stumps are "young" trees in the Random Forest. One leaf has two nodes. AdaBoost instead employs stumps. Stump-based decisions are poor. A mature tree forecasts the target value by merging all variable options. A stump may only use one variable for decision-making. To appreciate the inner workings of the AdaBoost algorithm, we will consider a number of characteristics to determine whether a person is "fit" (healthy).

DATA DESCRIPTION

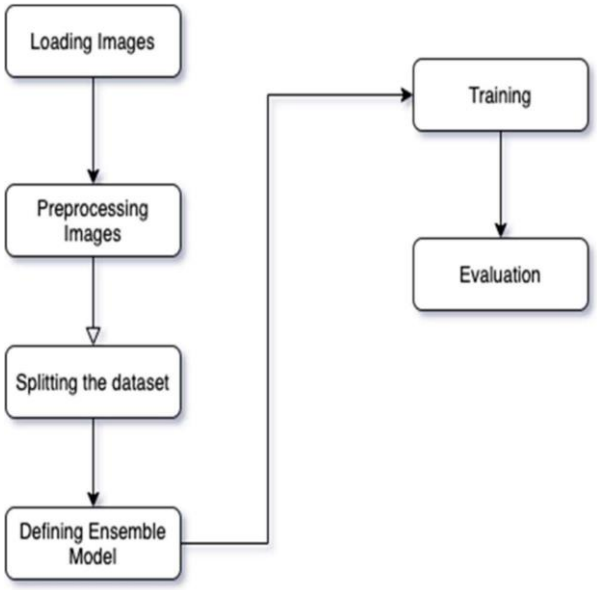
The pre-processing or cleaning of data is an important part of a Machine Learning Engineer's job, and the vast majority of Machine Learning Engineers put in a lot of work before building a model from scratch. Some examples of data pre-processing techniques are finding and dealing with outliers, dealing with missing values, and getting rid of unwanted or noisy data.

Image pre-processing, which is the same as image processing, is the term for images at the most basic level of abstraction. According to entropy as a measure of information, this process doesn't add more information to the image; instead, it takes away information. The goal of pre-processing is to improve the quality of the image data by getting rid of unwanted distortions and improving some visual qualities that are important for processing and analysing the image after it has been taken. Procedures for pre-processing can be put into two groups, which are listed below. There are two different kinds of pre-processing steps:

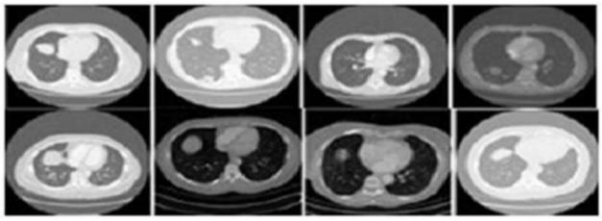
1. Filtering and dividing up images

2. The Fourier transform and restoring an image

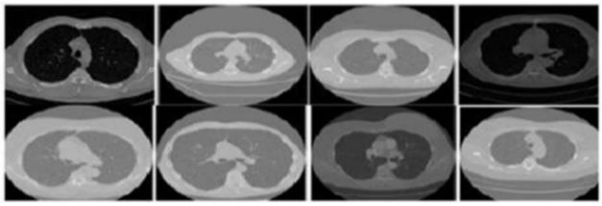
CT images must be pre-processed to get rid of noise and make them ready for later steps like image segmentation. Because of this, input images will be less distorted, and the right parts of inputs will be brought out more. MATLAB is used to prepare CT images before they are used. In the input database, the study looks at both primary and secondary phase cancer nodules. There are four different types of nodules to look at: well-circumscribed, juxta-pleural, vascularized, and pleural-tail. Figure 2 shows the CT image of the lung with cancer that was used to train the model.



The Collected datasets can be obtained from the following link: https://drive.google.com/file/d/1AZ2UQJToA3J8k8h0XJWD12- v2w8p9H2Q/view?usp=share_link

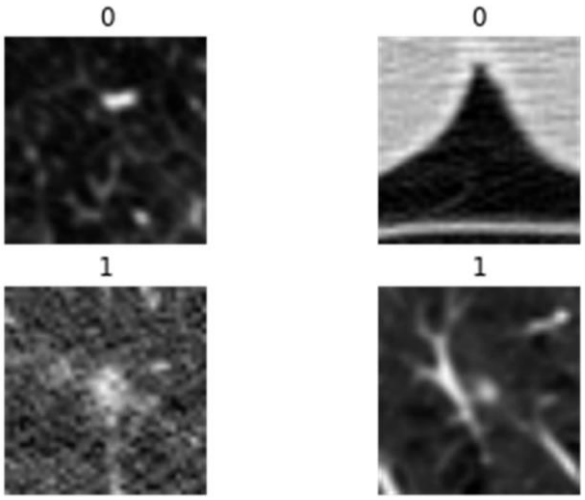


(a)

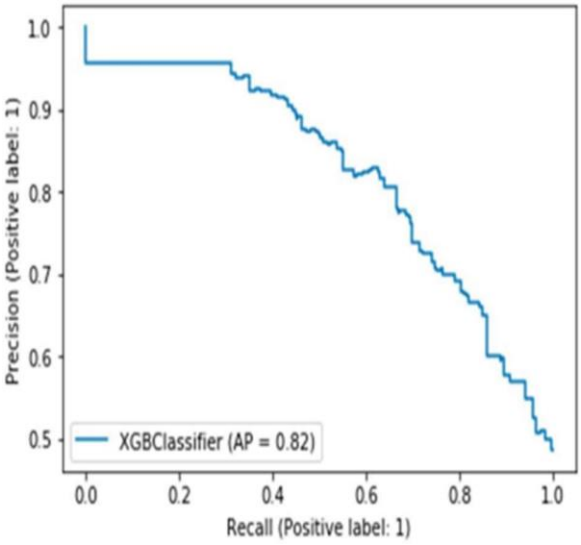


(b)

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN



These are the Sample Images of a) Malignant and b) Benign from the LIDC dataset.

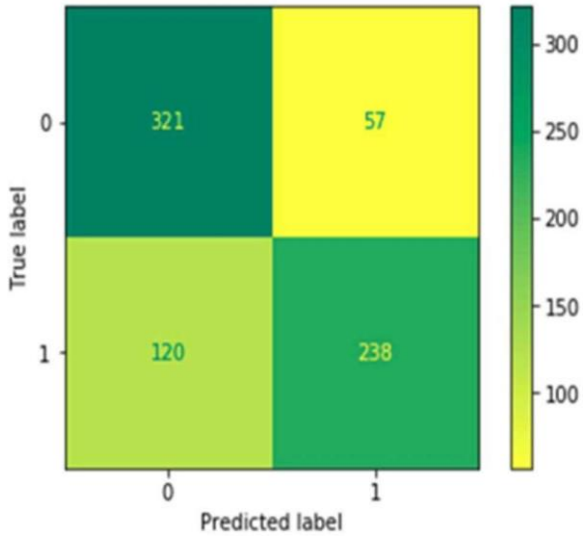


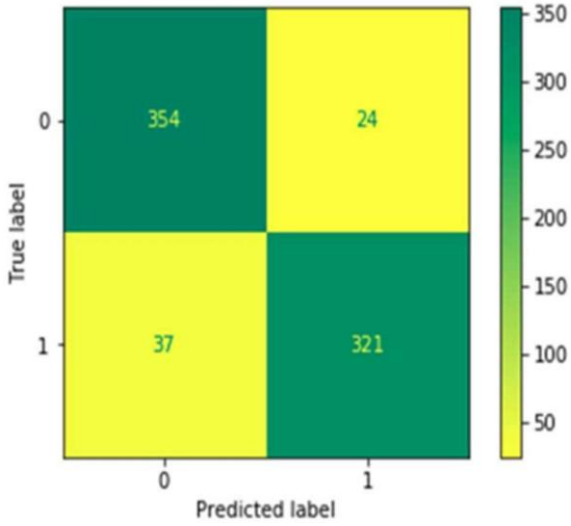
RESULTS

The performance measurements are utilized to establish the most efficient way of instruction (accuracy, sensitivity, and specificity). This system attained the greatest levels of performance, with average rates of 92 percent for accuracy, 92 percent for sensitivity, and 92 percent for specificity across all three performance criteria. By the KNN model, 76 percent, whereas NCA-XB Boosting is the lowest of the three at just 76 percent.

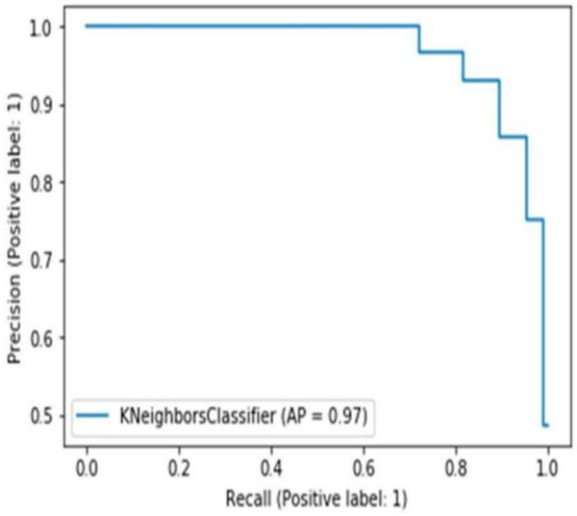
	precision	recall	f1-score	support
0	0.73	0.85	0.78	378
1	0.81	0.66	0.73	358
accuracy			0.76	736
macro avg	0.77	0.76	0.76	736
weighted avg	0.77	0.76	0.76	736

Results of NCA-XG Boosting



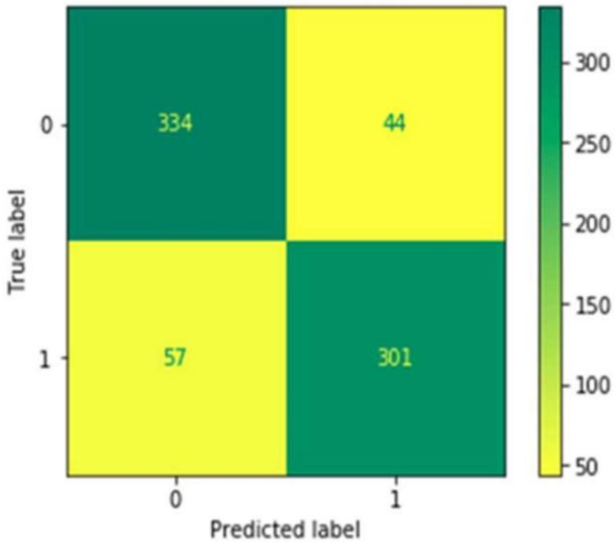


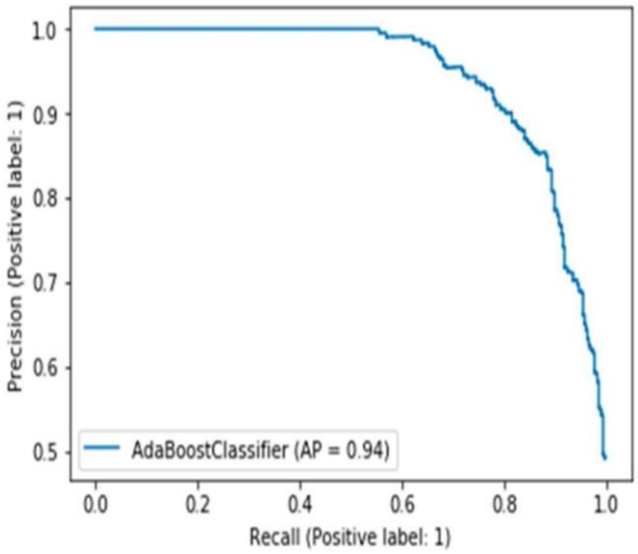
REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN



	precision	recall	f1-score	support
0	0.91	0.94	0.92	378
1	0.93	0.90	0.91	358
accuracy			0.92	736
macro avg	0.92	0.92	0.92	736
weighted avg	0.92	0.92	0.92	736

Results of KNN Boosting





	precision	recall	f1-score	support
0	0.85	0.88	0.87	378
1	0.87	0.84	0.86	358
accuracy			0.86	736
macro avg	0.86	0.86	0.86	736
weighted avg	0.86	0.86	0.86	736

Results of AdaBoost

ANALYSIS & CONCLUSION

As a consequence of deteriorating living conditions, such as sedentary lifestyles, bad diets, and increasing smoking, cancer rates have grown considerably over the last century. Therefore, researchers and experts have adopted measures to tackle this fatal illness. The findings of scientific study indicate that early discovery of this condition makes it simpler to cure and reduces the mortality risk associated with it. This research proposes using an autonomous method based on probabilistic neural networks to diagnose CT-based lung pictures in medical imaging as accurately as feasible. The classification and diagnosis accuracy of the suggested technique was good due to the extraction of high-level characteristics using deep neural networks. In terms of precision and accuracy, the KNN model surpasses its rivals. The model may be adjusted using feature selection and a stacking-based strategy, which can be coupled, to increase its accuracy. In order to enhance the performance of the model, we may, if required, increase the number of photos in the dataset.

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN

REFERENCES

[1] R. Navid, A. Mohsen, K. Maryam et al., "Computer-aided diagnosis of skin cancer: a review," Current Medical Imaging, vol. 16, no. 7, pp. 781–793, 2020.

[2] L. Hussain, W. Aziz, A. A. Alshdadi, M. S. Ahmed Nadeem, I. R. Khan, and Q.-U.-A. Chaudhry, "Analyzing the dynamics of lung cancer imaging data using refined fuzzy entropy methods by extracting different features," IEEE Access, vol. 7, pp. 64704–64721, 2019.

[3] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez, "Optimal deep learning model for classification of lung cancer on CT images," Future Generation Computer Systems, vol. 92, pp. 374–382, 2019.

[4] Armato SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38:915–31.

[5] Askarzadeh A. A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Comput Struct. 2016;169:1–12.

[6] Cascio D, Magro R, Fauci F, Iacomi M, Raso G. Automatic detection of lung nodules in CT datasets based on stable 3D mass–spring models. Comput Biol Med. 2012;42:1098–109. [7] Chen H, Zhang J, Xu Y, Chen B, Zhang K. Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans. Exp Sys Appl. 2012;39:11503–9.

[8] Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553:446.

[9] Cancer. Accessed: Apr. 30, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Cancer

[10] P. Chaudhari, H. Agarwal, and V. Bhateja, "Data augmentation for cancer classification in oncogenomics: an improved KNN based approach," Evol. Intell., pp. 1–10, 2019.

[11] S. F. Khorshid and A. M. Abdulazeez, "BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 1927–1951, 2021.

[12] F. Q. Kareem and A. M. Abdulazeez, "Ultrasound Medical Images Classification Based on Deep Learning Algorithms: A Review."

[13] D. Q. Zeebaree, A. M. Abdulazeez, D. A. Zebari, H. Haron, and H. N. A. Hamed, "Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features." [14] J. R. F. Junior, M. Koenigkam-Santos, F. E. G. Cipriano, A. T. Fabro, and P. M. de Azevedo-Marques, "Radiomics-based features for pattern recognition of lung cancer histopathology and metastases," Comput. Methods Programs Biomed., vol. 159, pp. 23–30, 2018.

[15] I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, 2021.

[16] P. Das, B. Das, and H. S. Dutta, "Prediction of Lungs Cancer Using Machine Learning," EasyChair, 2020.

[17] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," Neural Comput. Appl., vol. 31, no. 10, pp. 6863–6877, 2019.

[18] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021.

[19] H. A. Hussein and A. M. Abdulazeez, "COVID-19 PANDEMIC DATASETS BASED ON MACHINE LEARNING CLUSTERING ALGORITHMS: A REVIEW," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 2672–2700, 2021.

[20] D. M. Abdullah and N. S. Ahmed, "A Review of most Recent Lung Cancer Detection Techniques using Machine Learning," Int. J. Sci. Bus., vol. 5, no. 3, pp. 159–173, 2021.

[21] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer," in 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1–4.

[22] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 145–150.

[23] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Trainable model based on new uniform LBP feature to identify the risk of the breast cancer," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 106–111.

[24] H. Tang, J. Zhao, and X. Yang, "Explore machine learning for analysis and prediction of lung cancer related risk factors," in Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, 2018, pp. 41–45.

REPOSITORY LINK: https://github.com/AishaFar/Lung-Cancer-Recognition-Using-CT-Scan-with-NCA-XG-Boosting-KNN