JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org Vol.3, Iss.2 Pages:45-56

LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUE: A CASE STUDY

Vipul Parmar[1], Luvkush[2], Dr. Brajesh Kumar Singh[3]

[1]Master of Technology, [2]Assistant Professor, [3]Professor,
R.B.S. Engineering Technical Campus, Bichpuri, Agra

Article Received: February 2022 Published: August 2022

Abstract

Defining a positive lung cancer screening result and managing lung nodules detected in the scans are two major issues in low-dose computed tomography (CT) screening. We conducted a prospective study using low-dose CT scans as a screening tool to identify factors that may predict whether lung nodules detected during screening will undergo malignancy or be subsequently diagnosed as malignant. In concept image processing, images are captured using cameras/sensors on satellites, spacecraft, and planes, as well as images that are necessary for a wide variety of purposes in everyday life. It has been four to fifty years since various image handling approaches were developed. Automated rockets, space tests, and military surveillance aircraft are used to capture images for the majority of the processes. Image processing frameworks are becoming increasingly popular as powerful computing devices, large memory, design programming, and so on become more accessible. Clinicians must segment and organize clinical pictures when conducting clinical investigations. A CT lung image can be classified as normal or abnormal, depending on the imaging results. A division is then applied to the odd images. A growth segment is then revealed by dividing the odd photographs. Highlights are eliminated from the images as part of the arrangement.

Keywords: Lung disease, Lung Cancer, CT, Image Processing.

INTRODUCTION

Due to the nature of cancer cells, where most of the cells overlap, early diagnosis of lung cancer is a difficult challenge. Digital image analysis is incomplete without classification. It is a computational procedure for grouping photos based on their commonalities. This research uses Histogram Equalization for image pre-processing, feature extraction, and a neural network classifier to determine if a patient's condition is normal or abnormal at an early stage.

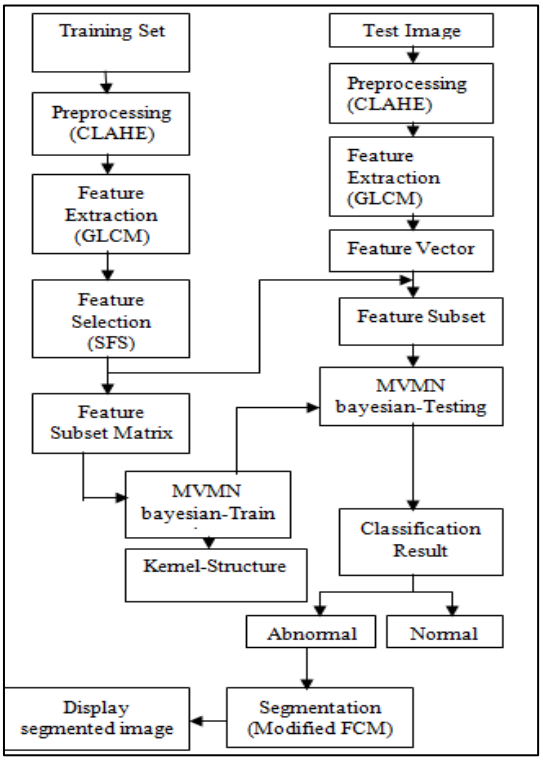

Fig 1 System Architecture

For the most part, we are concentrating on including the extraction stage to improve grouping execution. Surface-based highlights, such as GLCM (Gray Level Co-event Matrix), play an essential role in clinical image analysis. There was a total of 12 distinct measurable highlights extracted. We use sequential forward determination calculation to select the discriminative highlights among them. We lean toward multinomial multivariate Bayesian for the characterization stage after a while. The execution of the classifier will be investigated further. By altering the bunch community and enrollment esteem refreshing basis, the computational pace of the changed weighted FCM calculation is improved.

Manual sputum sample analysis is time-consuming, imprecise, and requires a highly trained individual to avoid diagnostic errors. The segmentation results will be utilized as the foundation for a Computer-Aided Diagnosis (CAD) system for early lung cancer identification, improving the patient's chances of survival. Because most quantitative procedures are based on the nuclear feature, we used a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions because the extreme variation in grey level and relative contrast among the images makes the segmentation results less accurate. The performance of classifiers

is evaluated by an experimental analysis using a dataset. The performance is based on reliable information. The performance is determined by the classifier's correct and wrong classifications. All of the trials were carried out using the WEKA data mining program.

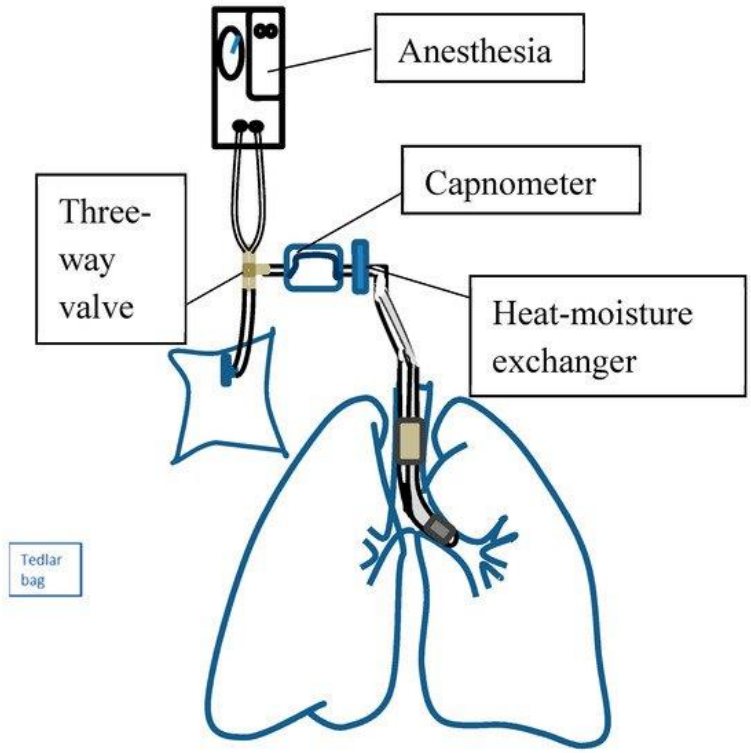

Fig 2 Schematic of the system architecture and sample library

A variety of factors cause cancer. If lung imperfections are discovered early on, there is a chance to improve the diagnosis and save lives. Blemishes - denseness ranging from 2.9 to 30mm, with additional variations in position, setting, non-solid, strong and sub-strong. CT may reveal many of these. CT is turned off during the initial investigation of the flaws. According to LCST (LUNG CANCER CELLS TESTING ROUTES), lung cancer cells are lowered by 20%, and lung cancer cell death is reduced by roughly five years in CT compared to higher.

This short essay aims to examine current knowledge of lung blemish location by CT scans as we go into the era of traditional CT-based lung cancer cell monitoring. After performing a CT scan on a patient, radiologists must examine the information in the form of photos based on nodule morphology and procedure. This should be done by clinical methods and should not include factors such as exhaustion or misinterpretation of information. The radiologists must evaluate the information diagnosed to improve the information in the form of photos or even progress picture analysis. The radiologists must evaluate the information diagnosed to improve the information in the form of photos or even progress picture analysis. According to studies, CT has a detection rate for lung cancer that is 2.6 times faster than analog radiography.

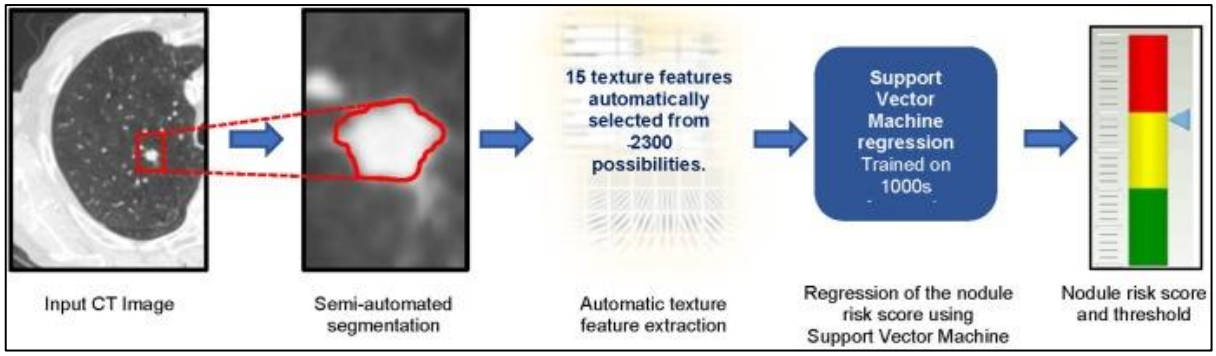

Fig3 A block diagram of the LungX winning system

A. Medical Image Classification

Classification Process

1) Training/Clustering Stage: the process of defining criteria by which patterns are recognized, and developing a numerical description for each class.

2) Classification Stage: each pixel in the image data set is categorized into the class it most closely resembles based on a mathematical decision rule.

3) Output Stage: results are presented in a variety of forms (tables, graphics, etc.)

• Supervised-image analyst "supervises" the selection image regions that present patterns/features that the analyst can recognize: Prior Decision.

• Unsupervised - statistical "clustering" algorithms used to cluster the pixels, more

• Computer-automated: Posterior Decision.

B. Machine Learning and Classification

ML comes under the umbrella of computer science that enables computers to learn without being explicitly programmed. We feed a training set into a machine learning system (like SVM, Artificial Neural Network, Logistic Regression, Linear Regression etc). The learning algorithm then generates a function referred to as the hypothesis for historical reasons. The hypothesis' role is to take a fresh input and produce a predicted output or class. The training is used to learn the parameters that define the hypothesis. A classifier will utilize the feature set to determine whether each pixel is a tumor pixel or a regular pixel after it has been computed for each pixel. The classification stage comprises two parts: training and classification. Training and testing phase are both included in the categorization step. Pixel features and their related manual labels are the input in the training phase, and the output is a model that predicts the appropriate label using the features.

Although random sub-sampling can be employed, spatial information can be used to construct a more strategic sub-sampling that does not impair the learned model as much as random sub-sampling. Because few pixels outside the brain will be needed in training, a non-random sub-

sampling technique that employs spatial information is to sub-sample proportionally to the pixel's prior odds of being part of the brain mask (assuming that a brain mask prior probability is used).

Non-random sub-sampling with spatial information could also be utilized to sub-sample common areas with vast distances from tumor pixels. These should exhibit fairly typical behavior and will not aid much in creating a model that correctly identifies confusing occurrences. "The support vector machine (SVM) is a universal constructive learning technique based on Vapnik's 1995 statistical learning theory." Mulier and Cherkassky (1998).

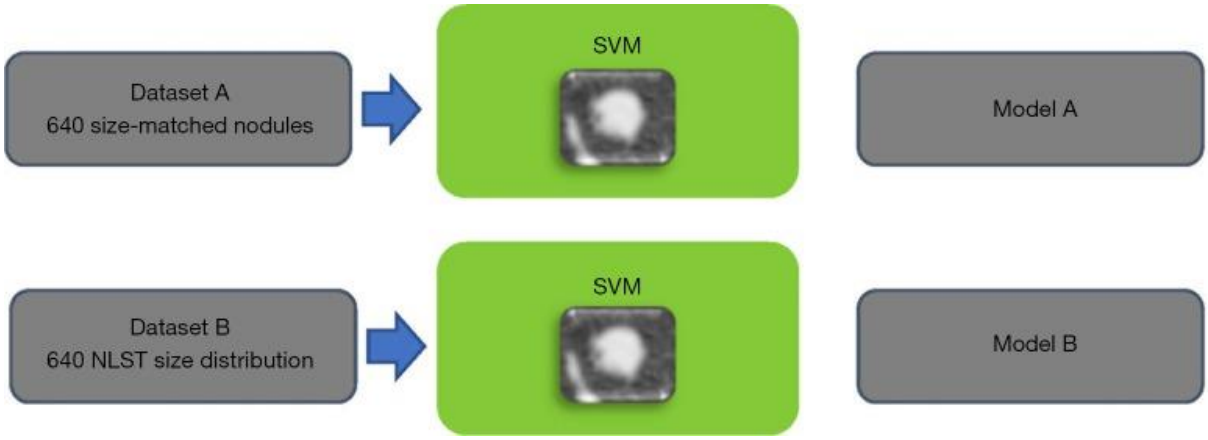

Fig 4 Nodule size within a machine learning model

Two groups have recently investigated Support Vector Machines (SVM) classification for the task of brain tumor segmentation. SVM models are more appealing than ANN models for binary classification because they have more robust (theoretical and empirical) generalization properties, achieve a globally optimal solution, and allow modeling nonlinear dependencies in the features. Standard pattern recognition methods are used to classify MR brain images, comprising procedures such as feature extraction, feature dimensionality reduction, and feature-based classification. Feature extraction is the most important of these three steps.

After extracting practical features, feature dimensionality reduction and feature-based classification can be accomplished quickly and efficiently utilizing machine learning algorithms. For example, Principal Component Analysis (PCA) and feature selection techniques can be used for feature dimensionality reduction. In contrast, the support vector machine (SVM) based on Lung Cancer Detection and Classification using Machine Learning & Multinomial Bayesian classifier or the linear discrimination analysis method can be used for classification. The goal of classification is to sort items into groups based on corresponding feature values. This is accomplished by the classifier deciding on a classification based on the value of the linear combination of characteristics.

REVIEW OF LITERATURE

They examined potential influences on sensors and assessed the dependability of the sensors. The conductive polymer sensor, a chemiresistive sensor type sensitive to temperature,

humidity, and baseline drift [1], was utilized in this investigation. In this investigation, we used the E-nose in the same space, which was kept at a constant 20 to 25 °C temperature and between 50 and 65 % environmental humidity.

They adopted a heat-moisture exchanger to the breath sampling device. In our pilot study, we measured the humidity of the sampled air before and after passage through the heat-moisture exchanger with a humidity meter (Rotronic HygroPlam, Bassersdorf, Switzerland, Supplementary. The mean relative humidity (R.H.) was 22.3% at 24 °C. To prevent a sensor drift influence, we visually examined all 32 raw sensor responses of the 10 measurements [2].

The ICC values of the 32 sensors were all greater than 0.99, and the coefficients of variation (CVs) were within 0.08–0.32%, indicating excellent reliability [3].

All procedures in this study were standardized to prevent any factors that could influence the VOC concentration. We collected alveolar air from the endotracheal tube to prevent contamination from the dead space in the respiratory or digestive tracts. All subjects were required to refrain from eating or smoking for 12 h before sampling. We used a fixed flow rate to obtain a steady concentration of VOCs to prevent the influence of the flow rate. To prevent the influence of contaminated sampling bags, we followed a standardized cleaning protocol according to recommendations from the European Respiratory Society [4].

Each bag was flushed with nitrogen five times and then heated to 45 °C for approximately 12 h; all the procedures were repeated overnight, which has been shown to provide good recoveries [5]. Because the breath air samples were collected from anesthetized subjects, anesthesia drugs might have influenced them. However, all subjects were administered the same anesthetic air (1–2% sevoflurane), and the anesthetic dose was adjusted for each subject's body weight (2 mg/kg for propofol, 2 μg/kg for fentanyl). We conservatively think the use of anesthetics might not confound our results.

Some studies support that volatile metabolites are generated from tumor tissue [6]. In contrast, others suggest that VOCs are released from the systemic circulation and released to the alveolar air by the gas exchange at the blood–gas interface in the lungs. [7]

To determine the origin of volatile metabolites, we separately analyzed VOCs from healthy and diseased lungs in the same patient among 24 subjects. After excluding two subjects with benign tumors, three subjects with metastatic lung cancer, and two subjects in which both lungs were affected, 18 subjects were used for the analysis [8,9].

However, the VOCs from healthy and diseased lungs were not discriminated well (Supplementary Material). These results are consistent with those of Capuano et al., indicating that volatile biomarkers might be produced not only by tumor tissue but also by an epi-phenomenon that accompanies lung cancer development, probably due to the chronic load and burden of VOCs in overall lung tissue [10].

Though GC-MS can be used to precisely identify the chemical components necessary to discover pathophysiological mechanisms, the E-nose can advantageously be used in clinical applications because it is simple to use, provides real-time analysis, and is hand-held in size.

Moreover, the methods for analyzing full-scan GC-MS signals have not been well established, and exhaled breath VOC libraries are currently being built but are not yet complete [11]. these methods without standards often identify VOCs that are not replicable in different studies [12]. this study are thus needed to explore actual biomarkers in the breath of lung cancer patients.

RESULTS

They examined potential influences on sensors and assessed the dependability of the sensors. The conductive polymer sensor, a chemiresistive sensor type sensitive to temperature, humidity, and baseline drift [1], was utilized in this investigation. In this investigation, we used the E-nose in the same space, which was kept at a constant 20 to 25 °C temperature and between 50 and 65 % environmental humidity.

They adopted a heat-moisture exchanger to the breath sampling device. In our pilot study, we measured the humidity of the sampled air before and after passage through the heat-moisture exchanger with a humidity meter (Rotronic HygroPlam, Bassersdorf, Switzerland, Supplementary. The mean relative humidity (R.H.) was 22.3% at 24 °C. To prevent a sensor drift influence, we visually examined all 32 raw sensor responses of the 10 measurements [2].

The ICC values of the 32 sensors were all greater than 0.99, and the coefficients of variation (CVs) were within 0.08–0.32%, indicating excellent reliability [3].

All procedures in this study were standardized to prevent any factors that could influence the VOC concentration. We collected alveolar air from the endotracheal tube to prevent contamination from the dead space in the respiratory or digestive tracts. All subjects were required to refrain from eating or smoking for 12 h before sampling. We used a fixed flow rate to obtain a steady concentration of VOCs to prevent the influence of the flow rate. To prevent the influence of contaminated sampling bags, we followed a standardized cleaning protocol according to recommendations from the European Respiratory Society [4].

Each bag was flushed with nitrogen five times and then heated to 45 °C for approximately 12 h; all the procedures were repeated overnight, which has been shown to provide good recoveries [5]. Because the breath air samples were collected from anesthetized subjects, anesthesia drugs might have influenced them. However, all subjects were administered the same anesthetic air (1–2% sevoflurane), and the anesthetic dose was adjusted for each subject's body weight (2 mg/kg for propofol, 2 μg/kg for fentanyl). We conservatively think the use of anesthetics might not confound our results.

Some studies support that volatile metabolites are generated from tumor tissue [6]. In contrast, others suggest that VOCs are released from the systemic circulation and released to the alveolar air by the gas exchange at the blood–gas interface in the lungs. [7]

To determine the origin of volatile metabolites, we separately analyzed VOCs from healthy and diseased lungs in the same patient among 24 subjects. After excluding two subjects with benign tumors, three subjects with metastatic lung cancer, and two subjects in which both lungs were affected, 18 subjects were used for the analysis [8,9].

However, the VOCs from healthy and diseased lungs were not discriminated well (Supplementary Material). These results are consistent with those of Capuano et al., indicating that volatile biomarkers might be produced not only by tumor tissue but also by an epi-phenomenon that accompanies lung cancer development, probably due to the chronic load and burden of VOCs in overall lung tissue [10].

Though GC-MS can be used to precisely identify the chemical components necessary to discover pathophysiological mechanisms, the E-nose can advantageously be used in clinical applications because it is simple to use, provides real-time analysis, and is hand-held in size. Moreover, the methods for analyzing full-scan GC-MS signals have not been well established, and exhaled breath VOC libraries are currently being built but are not yet complete [11]. these methods without standards often identify VOCs that are not replicable in different studies [12]. this study are thus needed to explore actual biomarkers in the breath of lung cancer patients.

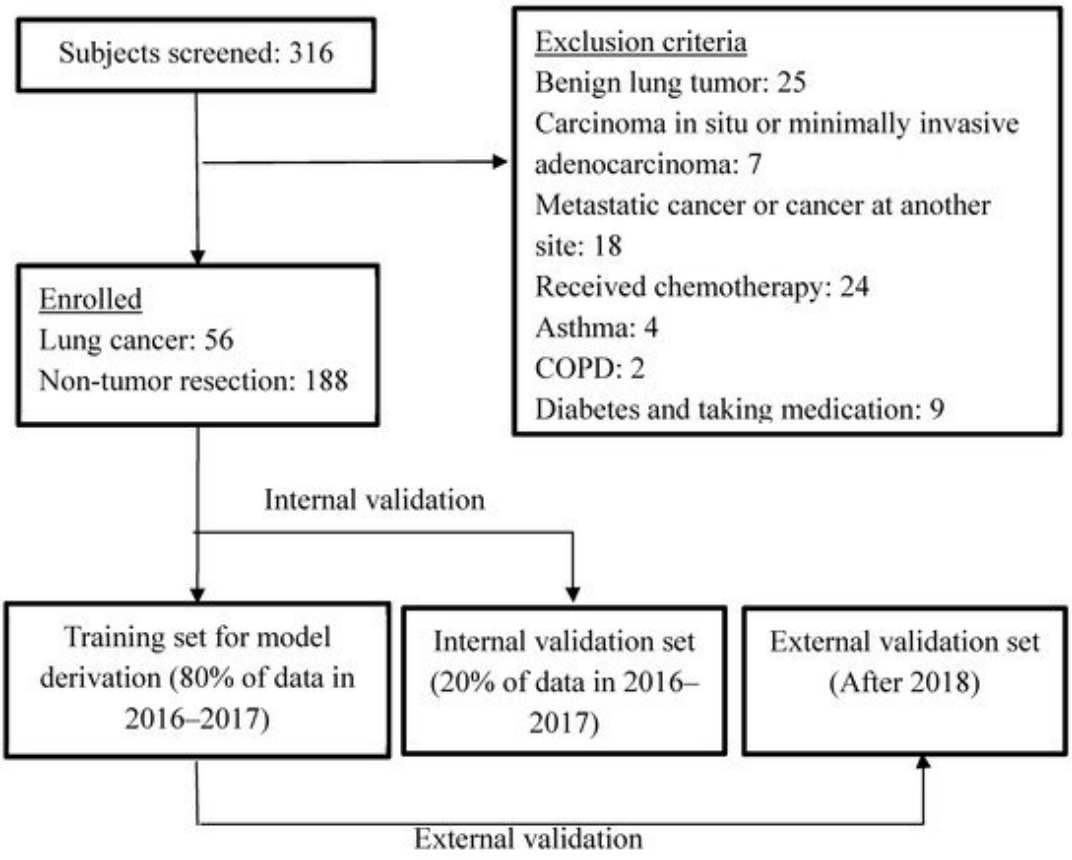

Fig 5 Flow diagram of the inclusion and exclusion of the study

Employed an independent external validation set and conducted repeated double cross-validation. The repeated double cross-validation used two nested loops. The inner loop used the study subjects enrolled between 2016 and 2017 as a calibration set for model selection and parameter optimization and was divided into a training set (80%) and an internal validation set (20%). The outer loop used the prediction model established from the calibration set to externally validate the study subjects enrolled in 2018.

LDA and SVM are used to estimate the receiver operating characteristic curves for lung malignancies in the internal and external validation sets. Both linear and non-linear techniques exhibit great accuracy according to the internal validation. The external validation shows a small decline in accuracy.

DISCUSSION

This study demonstrated that the E-nose might accurately identify early-stage lung cancer using a susceptible (at the ppb level) chemical sensor and a sophisticated data analysis approach.

We thoroughly examined the breath test's reliability. Lung cancer is frequently found using conductive polymers, quartz crystal microbalances, and metal oxide sensors. We chose a conductive polymer sensor for this investigation. This sensor is appropriate for detecting VOCs connected to lipid peroxidation, such as ethanol or isopropanol [12]. We discovered that the study's design and the choice of controls might have more of an impact on the detection accuracy than the types of sensors (Supplementary Material). The diagnostic performance of the test was inferior when the controls had additional comorbidities compared to results obtained when a healthy population was employed as the control.

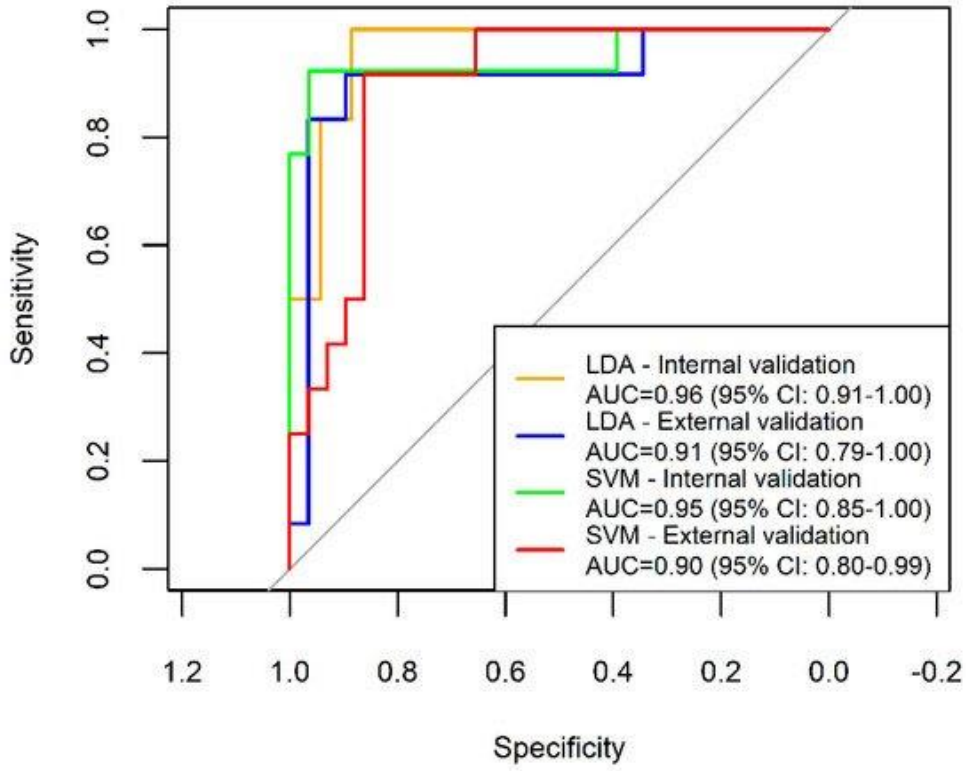

Fig 6 The LDA and SVM-derived internal and external validation sets' receiver operating characteristic curves for lung tumors

Both linear and nonlinear techniques exhibit great accuracy according to internal validation. The external validation shows a slight decline in accuracy. Case-control study diagnostic tests frequently exhibit control heterogeneity, sometimes known as "spectrum bias," which can result in overly optimistic accuracy estimates. Studies in the E-nose development stage must generally use a trustworthy test as a gold standard to evaluate the performance of a novel test,

such as LDCT or pathology reports. The accuracy attained may not be directly transferable to places where illness incidence varies because most investigations are done using a hospital-based case-control study methodology. A crucial phase in the creation of the prediction model is validation.

We used two nested loops recommended by Marco and strictly performed double cross-validation to evaluate the prediction accuracy. The calibration set was divided into a training set and an internal validation set by the inner loop, which employed internal validation processes for model selection and parameter optimization. The calibration and external validation sets were divided into outer loops to calculate the prediction performance. To confirm the repeatability of the test, external validation in a separate dataset from the intended target population of the screening technique is therefore required. As a result of our external validation of the breathalyzer in this study's target population, the test is appropriate for participants from the hospitals.

A good collaberation between sensor engineers, medical doctors, and statisticians is essential to accelerate the development of sensor technology in clinical use.

CONCLUSION

Using CT imaging data, we have covered the most common methods for nodule categorization and lung cancer prediction. In our experience, the present state-of-the-art is accomplished utilizing CNNs trained with Deep Learning, with classification performance in the low 90s AUC points, given sufficient training data. When analyzing system performance, it is critical to understand the limits of the training and validation data sets used, such as whether the patients were smokers or non-smokers and whether patients with a current or prior history of malignancy were included. After determining an acceptable level of performance, the next step is to evaluate such CADx systems in a clinical situation. However, before doing so, we must first describe how the CADx output should be used in clinical decision-making. Who should use a system like this, and how should it be integrated into their decisions? Should the algorithm create an absolute risk of malignancy, and if so, how should this be expressed? Should it be included in clinical opinion, and if so, how much weight should doctors or patients give it? Should the algorithms be integrated into or tailored to current recommendations like Lung-RADS or BTS? Should the program take into account variations in nodule size as time passes? Should the method account for changes in nodule volume if nodules are tracked over time, or should this be examined separately? Is success measured by a decrease in the number of false positive scans that require further investigation or intervention or by finding all lung tumors earlier than current recommendations allow? When determining the algorithm's worth, who should be compared to it? Should specialists or general radiologists be used as a comparison, as it may be tough to outperform an expert but beneficial to a generalist, as professionals do not read most scans? Such questions have received relatively little attention.

REFERENCES

1. Weir, J.P. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. 2005, 19, 231–240.
2. Bruton, A.; Conway, J.H.; Holgate, S.T. Reliability: What is it, and how is it measured? Physiotherapy 2000, 86, 94–99.
3. Lourenco, C.; Turner, C. Breath analysis in disease diagnosis: Methodological considerations and applications. Metabolites 2014, 4, 465–498.
4. Horvath, I.; Barnes, P.J.; Loukides, S.; Sterk, P.J.; Hogman, M.; Olin, A.C.; Amann, A.; Antus, B.; Baraldi, E.; Bikov, A.; et al. A European Respiratory Society technical standard: Exhaled biomarkers in lung disease. Eur. Respir. J. 2017, 49, 1600965.
5. Mochalski, P.; Wzorek, B.; Sliwka, I.; Amann, A. Suitability of different polymer bags for storage of volatile sulphur compounds relevant to breath analysis. J. Chromatogr. B 2009, 877, 189–196.
6. Kischkel, S.; Miekisch, W.; Fuchs, P.; Schubert, J.K. Breath analysis during one-lung ventilation in cancer patients. Eur. Respir. J. 2012, 40, 706–713
7. Wang, C.; Dong, R.; Wang, X.; Lian, A.; Chi, C.; Ke, C.; Guo, L.; Liu, S.; Zhao, W.; Xu, G.; et al. Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation. Sci. Rep. 2014, 4, 7312.
8. Amorim, L.C.A.; Cardeal, Z.D.L. Breath air analysis and its use as a biomarker in biological monitoring of occupational and environmental exposure to chemical agents. J. Chromatogr. B 2007, 853, 1–9.
9. Capuano, R.; Santonico, M.; Pennazza, G.; Ghezzi, S.; Martinelli, E.; Roscioni, C.; Lucantoni, G.; Galluccio, G.; Paolesse, R.; Di Natale, C.; et al. The lung cancer breath signature: A comparative analysis of exhaled breath and air sampled from inside the lungs. Sci. Rep. 2015, 5, 16491.
10. Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, " Artificial Neural Network based Cancer Cell Classification (ANN – C3)", Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.
11. Almas Pathan, Bairu.K.saptalkar, "Detection and Classification of Lung Cancer Using Artificial Neural Network", International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1.
12. Dr. S.A.PATIL, M. B. Kuchanur, " Lung Cancer Classification Using Image Processing," International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
13. Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques," Leonardo Electronic Journal of Practices and Technologies Issue 20, January-June 2012, p. 147-158.
14. Fritz Albregtsen, "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," International Journal of Computer Applications, November 5, 2008.
15. Taranpreet Singh Ruprah, "Face Recognition Based on PCA Algorithm," Special Issue of International Journal of Computer Science & Informatics (IJCSI), 2231–5292, Vol.- II, Issue-1, 2.
16. ZAKARIA SULIMAN ZUBI1, REMA ASHEIBANI SAAD, "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer", Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, LIBYA, 2011.
17. Balaji Ganeshan, Sandra Abaleke, Rupert C.D. Young, Christopher R. Chatwin, Kenneth A. Miles, "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage," Cancer Imaging , v.10(1): 137–143, 2010 July 6.
18. Lynne Eldridge MD. (2013, March 22). Lung Cancer Survival Rates by Type and Stage [Online]. Available:http://lungcancer.about.com/od/whatislungcancer/a/lungcancersurvivalrates.htm.
19. Morphological Operators, CS/BIOEN 4640: Image Processing Basics, February 23, 2012.
20. Image Processing – Laboratory 7: Morphological operations on binary images, Technical University of Cluj-Napoca, Computer.
21. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. N Engl J Med 2013;369:910-9. 10.1056/NEJMoa1214726 [PMC free article]
22. Gould MK, Ananth L, Barnett PG, et al. A Clinical Model To Estimate the Pretest Probability of Lung Cancer in Patients With Solitary Pulmonary Nodules. Chest 2007;131:383-8. 10.1378/chest.06-1261 [PMC free article]
23. Swensen SJ, Silverstein MD, Ilstrup DM, et al. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med 1997;157:849-55. 10.1001/archinte.1997.00440290031002
24. Deppen SA, Blume JD, Aldrich MC, et al. Predicting lung cancer prior to surgical resection in patients with lung nodules. J Thorac Oncol 2014;9:1477-84. 10.1097/JTO.0000000000000287
25. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax 2015;70 Suppl 2:ii1-54. 10.1136/thoraxjnl-2015-207168
26. Revel MP, Bissery A, Bienvenu M, et al. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 2004;231:453-8. 10.1148/radiol.2312030167

Cite this article:

Vipul Parmar, Luvkush, Dr. Brajesh Kumar Singh, "Lung Cancer Detection Using Machine Learning Technique: A Case Study", Journal of Multidimensional Research and Review (JMRR), Vol.3, Iss.2, pp.45-56, 2022