Research Article

A Novel Bayesian Framework For Multi-State Disease Progression Of Lung Cancer

K. Karthikayani [1], K. Ananthajothi [2*], R. Srividhya Lakshmi [3], Dr.A.Rajalingam [4],

1Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Email : karthikk3@srmist.edu.in.
2(Corresponding author), Department of Computer Science and Engineering, Misrimal Navajee Munoth Jain Engineering College, Chennai, India. Email: kanandjothime@gmail.com
3Department of Computer Science and Engineering, RMK College of Engineering and Technology, Chennai, India.
Email : srividhya.lakshmi89@gmail.com
4Professor, Shinas College of Technology,Al aqur, shinas, Oman.
Email : drrajalingam1234@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

ABSTRACT: CT screening has been commonly used to identify and diagnose lung cancer in its early stages. CT has been shown in clinical studies to reduce lung cancer mortality by 20% as compared to plain chest radiography; however, existing CT screening services face obstacles such as high over diagnosis rates, high costs, and elevated radiation exposure.The study develops computer and deep learning models for predictive lung cancer diagnosis and disease progression prediction in an effort to solve these difficulties. Using a symmetric chain code method and a machine learning system, a novel lung segmentation approach was first developed. The lung nodules connected to the lung wall are included in this process, which minimises over-segmentation error. Finally, to predict the inter disease progression of lung cancer, a Bayesian method was coupled with a prolonged Markov model.The resultant model calculates specific lung cancer state transition data, which can be used to make customised screening recommendations. Extensive trials and results have shown the efficacy of these approaches, paving the way for current CT screening systems to be optimised and improved.

1. INTRODUCTION

Lung cancer is the most common cancer killer in both men and women. Lung cancer has a 5-year survival rate of just 17 percent, but if diagnosed early on, the survival rate jumps to 54 percent. The seminal National Lung Screening Trial (NLST) demonstrated a 20% mortality reduction for people experiencing CT compared to plain chest radiography, making CT the de facto imaging modality for screening and identifying nascent lung cancers[1]. CT produces high-resolution, volumetric datasets that can resolve small and/or low-contrast nodules, as opposed to traditional chest radiography.However, there are many obstacles to accurate detection and successful screening when CT is used in this environment[2].

Established CT screening services face three obstacles, according to clinical studies: high over-diagnosis rates, high costs, and enhanced radiation exposure. The NLST study found that 96.4 percent of all positive screening results were false positives. The costs of an extra health and a performance life year per individual were $52,000 and $81,000, according to the findings in NLST. Furthermore, it was calculated that radiation causes roughly 1-3 lung cancer deaths per 10,000 examined subjects in the trial.The development of computer-aided detection/diagnosis (CAD/CADe) systems and the determination of individualised, optimal screening intervals are seen as critical steps in resolving these issues[3][4].

To make a conscious decision to reliably and effectively identify and diagnose lung cancer, CAD/CADe systems have used a range of machine learning and statistical approaches. For the last two decades, medical image processing has become a hotbed of science. When compared to human reading by thoracic radiologists, CAD/CADe systems have been investigated to assist radiologists in the reading process, potentially enhancing predictive value and reducing the false positive rate in lung cancer screening for small nodules. Screening has also been shown to be more cost-effective by using CAD/CADe systems[5].

With the massive ecosystem of screening data, it is now possible to model the genetic basis of lung cancer progression and, as a result, determine the best screening intervals for each person, making screening programmes more accurate and efficient[6]. As shown in Figure 1, lung cancer usually progresses through three stages: a disease-free state (State 1), a clinical trials state observable through screening but asymptomatic (State 2), and a symptomatic state (State 3). The mean sojourn time (MST) is a measurement of how quickly a disease advances

from preclinical to clinical stages.To approximate MST, a variety of statistical and temporal methods have been developed, including Markov models and discrete equations-based methods. Despite numerous attempts, estimating MST for various subject cohorts using traditional methods remains difficult due to measurement error and data sparsity concerns.

To overcome some of the above issues, this work focuses on lung cancer progression modelling.

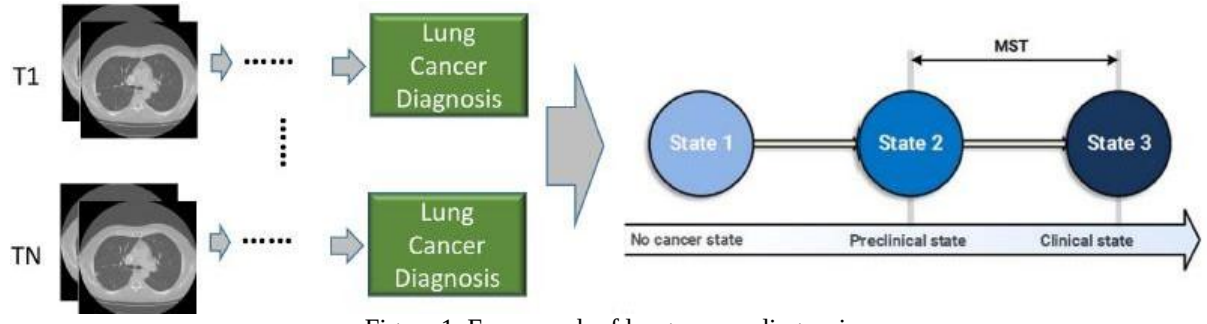

Figure 1. Framework of lung cancer diagnosis

The mean sojourn period (MST), which measures how quickly lung cancer develops from a preclinical to detectable clinical condition, determines how quickly a lung cancer can be identified using imaging. As a result, MST is commonly used in population screening for determining the optimum interval between screens and evaluating the degree of overdiagnosis. MST varies depending on imaging techniques and patient cohorts, with higher MST patients[7]. MSTs should have longer screening times so they are at a reduced risk of cancer.Personalized temporal models that estimate MST can thus help shift screening guidelines away from a one-size-fits-all approach and toward more tailored policies. However, using retrospective screening data to estimate MST has a number of drawbacks[8].

For starters, clinical findings for disease states are often subject to interpretation error, such as when physicians miss a cancerous nodule. Any MST prediction will be skewed if such observation error is not modelled. Second, in clinical practise, missing or incomplete findings are normal. Some patients, for example, can skip a scheduled screening exam or receive treatment at a facility where data is not shared. Third, the time between screening tests is often inconsistent[9].As a consequence, the estimation of continuous time data results in the loss of important data. Fourth, some detected disease states may have a very limited sample size (i.e., sparse), making evaluation difficult. When a patient has an early stage cancer, for instance, they will normally have an injection to remove them from further examination.As a result, there are fewer individuals for which uncertainties can be measured for transitions to later states. To overcome such issues, disease progression variables for periodic screening data are needed. The most significant contribution is the development of a new mathematical multi-state disease progression model structure[10]. A continuous time markov model is used to model disease progression and observation error in this model.To deal with issues like incomplete observations and data sparsity, a Bayesian approach is used. This estimation model allows for more precise determination of appropriate screening times. It also acts as a basis for moving toward tailored screening frameworks. Section II describes the related study, Section III reviews the system model , Section IV depicts the results followed by conclusion in Section V.

2. RELATED STUDY

To diagnose early stage lung cancer, computed tomography (CT) has become the most commonly used screening method. The seminal NLST study published in 2011 found that CT screening reduced mortality by 20% compared to simple chest radiography in people with lung cancer. Following this evidence, the United States Preventive Services Task Force issued a Grade B guideline that annual lung cancer screening with CT be undertaken in adults aged 55 to 80 who have a 30 pack-year smoking habits (number of packs of cigarettes smoked per day based on the number of years an individual has smoked) smoking history and either still smoke or have quit within the previous 12 months.This policy has prompted the development and introduction of new CT-based lung cancer screening programmes. CT scan perception is both time-consuming and potentially difficult[11].

Lung cancer in its early stages presents as pulmonary nodules, which appear on CT scans as small circular or oval-shaped opacities with diameters less than 30mm. With a growing percentage of CT scans to read and a higher resolution, reading such large sets of data can cause visual fatigue and/or pressure, which can lead to a reduction in diagnostic accuracy. Furthermore, since interpretation is highly reliant on prior experience, less trained radiologists have significant variability in detecting subtle lung cancers. For the detection of lung nodules, there has been significant variation in performance amongst radiologists.CT scan analysis is further complicated by the complex airway and vessel layout[12].

CT screening produces a high number of false positives, even among seasoned radiologists, resulting in a serious over-diagnosis. The high false positive rates due to benign nodules are essential to minimize the benefits of CT screening for early lung cancer detection; minimizing false positives and identifying patients who need medication could reduce costs and morbidity correlated with alone and intervention. As a result, distinguishing benign from malignant nodules is becoming increasingly important.Distinguishing benign from malignant nodules, as well as indolent vs. aggressive cancers, poses significant challenges. As a result, computer-aided diagnosis (CAD) systems have been extensively researched to help physicians solve this issue[13].

CAD is used in all parts of the body, such as the head, abdomen, heart, breast, kidney, spine, and joints, and has been established in screening techniques such as magnetic resonance imaging (MRI), CT, projectional radiography, nuclear medicine, ultrasound, and digital pathology imaging. To make the final decision, these CADs typically use a traditional machine learning scheme, which includes segmentation of the region of interest, extraction of features, and labeling.Conventional machine learning approaches are unable to process natural data in its raw form (raw image pixels), necessitating the use of feature design and technical requirements to define and derive significant value from the raw data into learnable depictions. Representation learning, on the other hand, is a class of methods that can automatically identify the best representation of raw data and derived features to help with classification, prediction, and detection tasks[14].

Deep learning methods have recently been applied to a number of medical image processing tasks after achieving considerable success in object detection tasks in image features. However, in the clinical domain, a major obstacle is the lack of labelled datasets. Most of the success of normal image recognition tasks is due to the availability of large numbers of labelled training data (e.g., millions of tagged images), which allows the training of complex and deep neural networks with thousands of parameters.The volume of branded training knowledge in the healthcare domain is much lower. The 3D structure of most medical images is another concern in clinical imaging techniques. A standard magnetic resonance imaging (MRI) or computed tomography (CT) scan has several slices, with disease regions inside the scan being relatively small in comparison to the entire picture stack (and of varying 3D size)[15].

Since the cross-sectional scanning slice thickness is often much larger than the in plane pixel dimension, collecting 3D data quickly in a deep neural network architecture is difficult. Designing deep-learning-based systems for 3D medical images has become more difficult. A better understanding of lung cancer's growth and complexities, such as the time it takes to reach a particular stage of the disease, can contribute to more effective prevention, treatment, and treatment, as well as early diagnosis.One of the most effective ways to diagnose early stage lung cancer is by routine imaging screening. Longitudinal data gathered as a result of screening can be used to develop improved methods for describing disease progression and generating forecasts for individualised diagnosis and monitoring policies. An increasing number of people who are thought to be at high risk of cancer are now regularly screened at the population level.However, risks such as radioactive contamination, overtreatment, and defensive medicine highlight the need for more accurate temporal models that predict who should be monitored and how often[16].

The mean sojourn time (MST), a period of time during which a tumour can be identified by imaging but no clinical symptoms are present, is an important factor to consider when developing screening policies. Continuous Markov models (CMM) with Maximum likelihood estimation (MLE) have been used to estimate MST for a long time. However, risks like radiation exposure, overdiagnosis, and overtreatment highlight the need for more accurate temporal models that can forecast outcomes.Traditional methods, on the other hand, presume no observation error when interpreting imaging data, which is impossible and can skew MST estimates[17]. Furthermore, when data is sparse, the MLE can not be stably calculated. This thesis proposes a probabilistic modelling method for periodic cancer screening data to address these flaws. [18] [20] an iterative measure of impact measure that estimates target class influence based on multiple levels and probability of information presented by the author. [19] To improve overall accuracy, assign positive reductions and experiment with different patch sizes.

3. SYSTEM MODEL

A disease-free state, a clinical trials state observable through screening but show no symptoms, and a diagnosable state are the three stages of lung cancer progression. The model assumes that in order to reach State 3, a patient in State 1 must first pass through State 2. When a patient is screened, one of two conditions can be noted: the patient is in the clinical trials state if the screening test is positive and confirmed by a diagnostic evaluation; alternatively, the individual is in the illness free state.

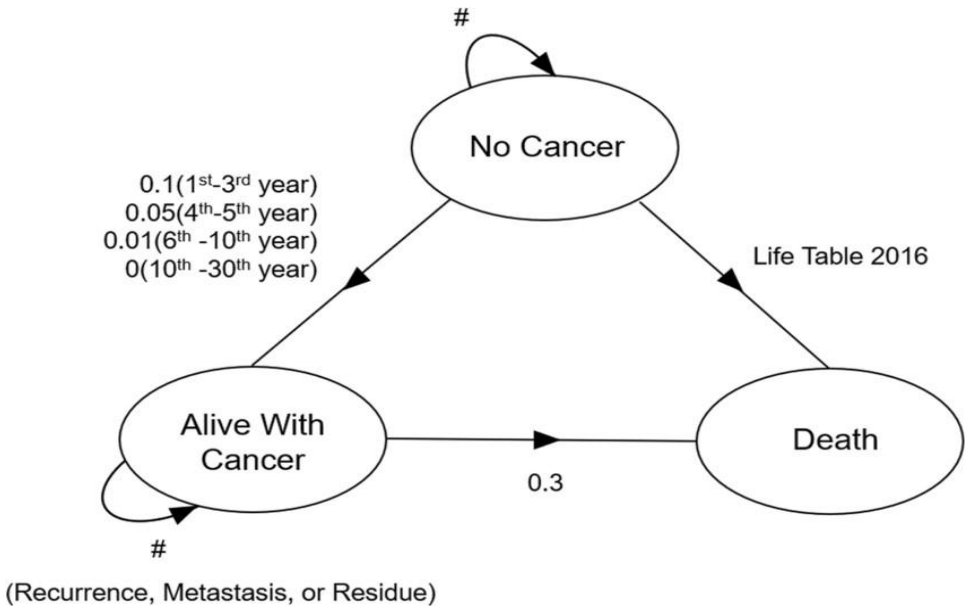

Figure 2. Proposed Bayesian model transition diagram

As a result, the second state (preclinical) can be distinguished by two factors: a positive screening test and a verified positive pathology evaluation. Patients in the preclinical stage, on the other hand, include both false-negatives and true-negatives, both of which will advance to the clinical stage. The individual is in the clinical condition when cancer is first diagnosed through evolving lung cancer symptoms (rather than by screening). Patients who do not advance to the clinical state and are not considered to be preclinical during screening will replicate the procedure in subsequent rounds in inter screening facilities.Interval cases are those that are diagnosed with lung cancer symptoms before undergoing another round of screening.

Since the direct change from the disease-free to the preclinical state is clinically unobservable, MST is hard to ascertain. Patients will be treated after being detected in a clinical trials state (a positive cancer screening), avoiding the normal transition from preclinical to clinical. If no control group is open, interval cases become the only source of knowledge for estimating MST. The assessment of MST is influenced by detection sensitivity, just as the discovery of interval cancers is affected by false-negative screening results; a skewed estimate of sensitivity will impact the measurement of MST.Sensitivity refers to the chance of identifying clinical trials of cancer by screening.

Table 1. Participants breakdown

	Participants	Screening-Detected Cases	Negative Screening Cases	Interval Cancers
First screening	26	14	24	9
Second screening	23	7	21	6
Third Screening	19	4	16	20

The model was created for three types of screenings, including interval cancer and post-screening cancer. The time ranges between the first and second screenings, as well as the second and third screenings, are respectively t_{12} and t_{23}. Assume that all participants' t_{12} and t_{23} are the same.Only elevated lung cancer subjects with a minimum of 20 pack-years of smoking cigarette history were included in the National Lung Screening Trial dataset (NLST) used in this study. Covariates are classified from ethnicity, smoking history, and health history, including age, gender, family background of cancer, waist circumference, disease history, cancer history, current or former smoker, number of packs of cigarettes per year, and smoke years, to further split the dataset into sub-cohorts and explore cancer progression variations.Each covariate's proportions in the no cancer, non-symptomatic cancer, symptomatic cancer, and post-screening cancer groups are plotted and compared to classify the significant covariates in this high-risk population for further inequity.

4. EVALUATION AND RESULTS

Two measures are used to test the proposed model and equate it to other methods. To begin, Pearson's chi-square is used to assess the proposed model's suitability and validate analytical results by determining if there are substantial discrepancies between observed and expected counts. A p-value greater than 0:05 indicates that there is no substantial difference, suggesting a good t and reliable parameter estimation.The Bayesian method then uses posterior predictive p-values (PPPV) to search for inconsistencies between model predictions and observed counts. A PPPV with a p-value greater than 0 implies a strong t. The proposed model suits the data better than the model without measurement error by using Pearson's chi-square and PPPV.

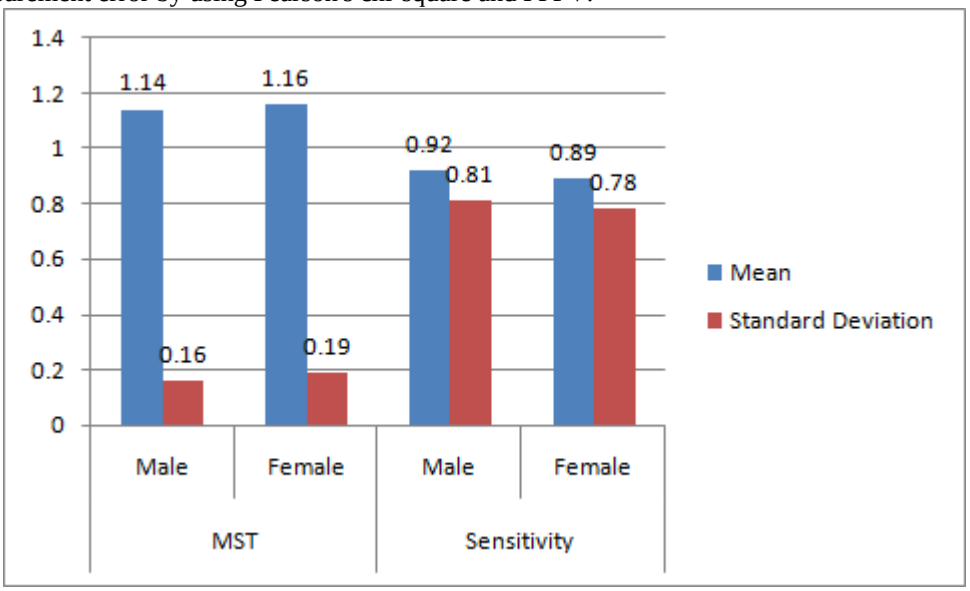

Figure 3. Summaries of the posterior for the two gender groups

The model provides realistic and reliable results when applied to lung cancer screening data from the NLST dataset. For modelling the transitions of discrete health states, the Bayesian model is a natural option. Through integrating longitudinal patient data and variable monitoring intervals, it can model transitions over time. Two relevant unknown parameters in the model are sensitivity and MST. These two criteria, however, are intertwined and difficult to separate. MST can only be measured without a control group based on the prevalence of interval cancer cases. More false-negative cases, on the other hand, lead to more interval cases, due to shorter MST.MST and sensitivity should therefore be modelled together, and projections should be sensitive to minor variations in interval cancer counts. Owing to the fact that both are not estimated together, this approach is often prone to error.

5. CONCLUSION

A predictive bayesian disease assessment model is proposed that uses a novel Bayesian method to achieve more accurate and stable disease progression estimation when accounting for observation error. The model lays the groundwork for making individualised screening recommendations for a group of people based on their covariates. This study shows how to measure MST more reliably for diverse cohorts with sparse observations when accounting for observation error.The emphasis of future research will be on developing an individualised bayesian system that models each patient's data separately. Both average and individualised patient migration times were examined in the reduced model with sensitivity of 1, and the estimates for MST were very similar.

REFERENCES

1. Adi, Kusworo & Widodo, Catur & Widodo, Aris & Gernowo, Rahmat & Pamungkas, Adi & Syifa, Rizky. (2017). Naïve Bayes Algorithm for Lung Cancer Diagnosis Using Image Processing Techniques. Advanced Science Letters. 23. 2296-2298. 10.1166/asl.2017.8654.
2. Armero, Carmen & Cabras, Stefano & Castellanos, Maria & Perra, S & Quirós, Alicia & Oruezabal Moreno, Mauro & Sánchez-Rubio, J. (2012). Bayesian analysis of a disability model for lung cancer survival. Statistical methods in medical research. 25. 10.1177/0962280212452803.
3. Bharath, AC & Kumar, Dhananjay. (2014). An improved Bayesian Network Model Based Image Segmentation in detection of lung cancer. 2014 International Conference on Recent Trends in Information Technology, ICRTIT 2014. 1-7. 10.1109/ICRTIT.2014.6996143.

4. Dimitoglou, George & Adams, James & Jim, Carol. (2012). Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability. 4.
5. Dwivedi, Mr & Borse, Rushikesh & Yametkar, Mr. (2014). Lung Cancer detection and Classification by using Machine Learning & Multinomial Bayesian. IOSR Journal of Electronics and Communication Engineering. 9. 69-75. 10.9790/2834-09136975.
6. Hui, Bei & Zhou, HongTing & Jiang, YuNan & Ji, Lin & Chen, Jia. (2017). The Research of Postoperative Life Expectancy of Lung Cancer Based on Semi-Naive Bayesian. 17-19. 10.26480/iscsai.01.2017.17.19.
7. Jabbari, Fattaneh & Villaruz, Liza & Davis, Mike & Cooper, Gregory. (2020). Lung Cancer Survival Prediction Using Instance-Specific Bayesian Networks. 10.1007/978-3-030-59137-3_14.
8. Jiang, Weiqin & Shen, Yifei & Ding, Yongfeng & Ye, Chu-Yu & Zheng, Yi & Zhao, Peng & Liu, Lulu & Tong, Zhou & Zhou, Linfu & Sun, Shuo & Zhang, Xingchen & Teng, Lisong & Timko, Michael & Fan, Longjiang & Fang, Weijia. (2017). A naive Bayes algorithm for tissue origin diagnosis (TOD- Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. International Journal of Cancer. 142. 10.1002/ijc.31054.
9. Liu, Yan & CHANG, Wen-jun & Cao, Guangwen. (2014). Application of Bayesian classifier in diagnosis of lung cancer by multiple autoantibody biomarkers. Academic Journal of Second Military Medical University. 33. 1358-1364. 10.3724/SP.J.1008.2013.01358.
10. Luo, Yi & Mcshan, Daniel & Ray, Dipankar & Matuszak, Martha & Jolly, Shruti & Lawrence, Theodore & Kong, feng-ming & Ten Haken, Randall & El Naqa, Issam. (2018). Development of a Fully Cross-Validated Bayesian Network Approach for Local Control Prediction in Lung Cancer. IEEE Transactions on Radiation and Plasma Medical Sciences. PP. 1-1. 10.1109/TRPMS.2018.2832609.
11. Oh, Jung Hun & Craft, Jeffrey & Lozi, Rawan & Vaidya, Manushka & Meng, Yifan & Deasy, Joseph & Bradley, Jeffrey & El Naqa, Issam. (2011). A Bayesian network approach for modeling local failure in lung cancer. Physics in medicine and biology. 56. 1635-51. 10.1088/0031-9155/56/6/008.
12. Patra, Radhanath. (2020). Prediction of Lung Cancer Using Machine Learning Classifier. 10.1007/978-981-15-6648-6_11.
13. Priya, T. & Meyyappan, T.. (2021). Disease Prediction by Machine Learning Over Big Data Lung Cancer. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 16-24. 10.32628/CSEIT206669.
14. Sangeetha, K. Ananthajothi.(2020).Machine Learning Tools for Digital Pathology - The Next Big Wave in MEDICAL SCIENCE.Solid State Technology.VOl.63.No 4.
15. Sesen, Mustafa & Kadir, Timor & Alcantara, Rene-Banares & Fox, John & Brady, Michael. (2012). Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer. AMIA .Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2012. 838-847.
16. Sesen, Mustafa & Nicholson, Ann & Banares-Alcantara, Rene & Kadir, Timor & Brady, Michael. (2013). Bayesian Networks for Clinical Decision Support in Lung Cancer Care. PloS one. 8. e82349. 10.1371/journal.pone.0082349.
17. Taher, Dr. Fatma & Werghi, Naoufel & Al-Ahmad, Hussain. (2012). Bayesian classification and artificial neural network methods for lung cancer early diagnosis. 773-776. 10.1109/ICECS.2012.6463545.
18. Ananthajothi, K., Subramaniam, M. Multi level incremental influence measure based classification of medical data for improved classification. Cluster Comput 22, 15073–15080 (2019). https://doi.org/10.1007/s10586-018-2498-z
19. K. Karthikayani ,T. Selvakumar , K. Ananthajothi. 2021. "Design of Convolutional Neural Network for Lung Cancer Diagnosis".Annals of the Romanian Society for Cell Biology, April, 7630 -. http://annalsofrscb.ro/index.php/journal/article/view/3417.
20. Ananthajothi.K & Subramaniam.M , 'Efficient Classification of Medical data and Disease Prediction using Multi Attribute Disease Probability Measure', Applied Mathematics & Information Sciences, ISSN 1935-0090, E.ISSN 2325-0399, Vol. 13, no. 5, pp. 783-789 (2019). https://doi:10.18576/amis/130511