Journal of Science and Technology
ISSN: 2456-5660 Volume 7, Issue 01 (JAN-FEB 2022)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2022.v7.i01.pp188-198*

EFFICIENT MACHINE LEARNING MODEL TO IDENTIFY THE LUNG CANCER USING DYNAMIC FEATURE EXTRACTION

[1] ASIYA, [2] N. SUGITHA

[1] Research Scholar, CSE Dept, Noorul Islam University ,Noorul Islam Center for Higher Education,Kumarakovil, Thuckalay, Tamil Nadu

[2] Professor - ECE Dept, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu

To Cite this Article

ASIYA, N. SUGITHA**, EFFICIENT MACHINE LEARNING MODEL TO IDENTIFY THE LUNG CANCER USING DYNAMIC FEATURE EXTRACTION"***",Journal of Science and Technology, Vol. 07, Issue 01,-JAN-FEB 2022, pp188-198*

Article Info

Received: 04-02-2022 Revised: 15-02-2022 Accepted: 19-02-2022 Published: 27-02-2022

ABSTRACT

An estimated 1.2 million people were diagnosed with lung cancer in 2000, making it the most frequent disease worldwide (12.3% of all malignancies). Cigarette smokers are responsible for 80% to 90% of lung cancers. In both sexes, lung cancer continues to be the major cause of cancer-related death in the United States and elsewhere. Tobacco use and smoking are responsible for nearly all occurrences of lung cancer. Other causes of lung cancer include exposure to radon gas, asbestos, air pollution, and persistent infections. Further, many potential risk factors for developing lung cancer have been proposed, including both genetic and environmental factors. Small-cell lung carcinomas (SCLC) and non-small-cell lung carcinomas (NSCLC) are the two main histologic subtypes of lung cancer and exhibit distinct patterns of growth and metastasis (NSCLC). Surgery, radiation treatment, chemotherapy, and targeted therapy are all viable alternatives for treating lung cancer. Different characteristics, such as the nature and extent of the malignancy, inform suggestions for treatment approaches. A diagnosis of lung cancer at an early stage can save the lives of patients. Several machine learning algorithms were used to make lung cancer forecasts in this study.

Keywords: SCLC, NSCLC, Lung cancer, Machine Learning, Tobacco smoking

1. Introduction

Lung cancer treatment is intricate and constantly developing. Multidisciplinary assessment and management, as well as the use of many therapy modalities, necessitate a holistic and coordinated strategy for optimal patient outcomes. In the United States of America, lung cancer is the leading cause of cancer-related death in both men and women due to its aggressiveness, the speed with which it spreads, and overall prevalence (USA). According to recent estimates [1], a total of 224,210 new cases of lung cancer will be diagnosed in the

United States in 2014, with 159,260 deaths expected. It's responsible for more annual deaths in the United States than the next four major cancers combined (colon/rectal, breast, pancreas, and prostate). Both its prevalence and fatality rates are reliably linked to a history of smoking for 20 years or more.

Over a decade has passed since the first systematic analyses of the research into the link between lung cancer (LC) and ILD were published. These reviews (1,2) mostly focused on epidemiology and found an increased risk of LC in ILD. The last comprehensive review of LC was published in 2017; its subject matter was LC linked to IPF (3). In light of recent findings, we set out to expand on previous reviews of LC-ILD.

In order to distinguish IPF from other ILDs, it must be understood how it differs from LC, which shares risk factors such as smoking and chemical exposure (4,5). In contrast to the other forms of ILD, where inflammation and immunosuppression play major roles, IPF physiopathology is predicated largely on epithelial damage, repair problems, and epithelial mesenchymal transition, much like carcinogenesis. A better prognosis is shown by these latter cases compared to IPF.

There are a number of caveats to this study that we feel should be made clear at the outset. The majority of the information was gathered by looking back to Asian cohorts that had already been studied in the past. Very few research addressed all ILDs, while the vast majority concentrated on IPF. Figure 1 shows that 70% of individuals diagnosed with lung cancer have an advanced case of the illness.

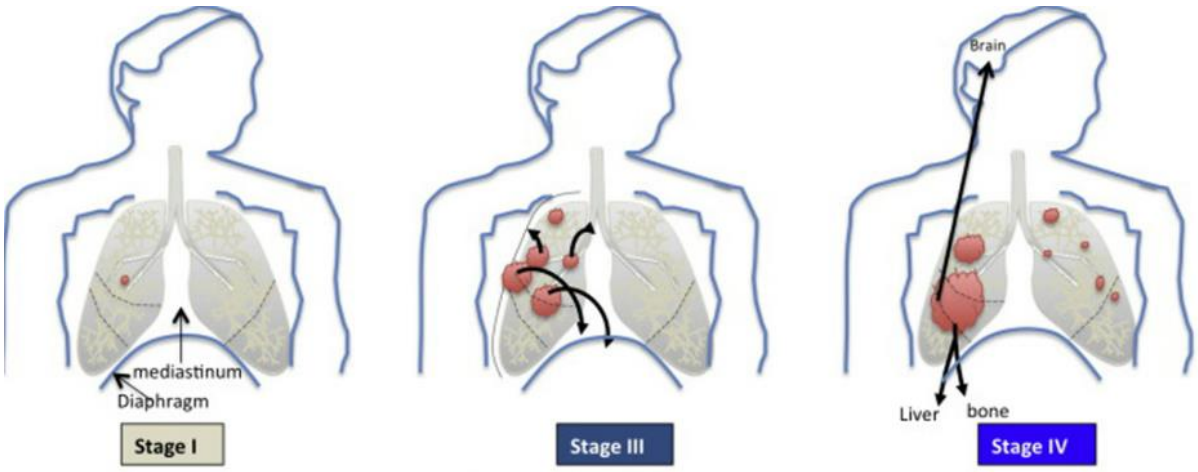


Figure.1. Different stages of Lung cancer

2. LITERATURE SURVEY

For both men and women, an estimated 1.4 million fatalities per year[2] can be attributed to lung cancer. Lung cancer accounted for around 13% of all cancer diagnoses in 2008, with an estimated 1.6 million new cases reported worldwide. 1 Twelve percent of the estimated 3.2 million cancer cases in Europe are caused by lung cancer[3]. Lung cancer was estimated to affect 221,130 people in the United States in 2011, making up 14% of all cancer diagnoses[4].

Lung cancer has a 15% 5-year survival rate, higher than the next three most common cancers combined (colon, breast, and prostate).

5 About 900,000 new multiple cancer cases (8% of 10.9 million cases) are detected in patients with a history of primary cancer.

The ageing of the population is largely to blame for the rise in the incidence of these malignancies; only 5-12 percent of cancer patients aged 50-64 had a history of cancer, compared to 12-26% of those aged > 80[8]. This rise in incidence is due in part to better diagnostic and therapeutic methods. As a result, this tendency places a significant strain on healthcare systems[9], and studies focusing on the prevalence of cancers with numerous source sites have gained considerable traction in epidemiological and clinical settings.

Small cell lung cancer (SCLC; or oat cell cancer) accounts for 20% of all lung cancer cases, while non-small cell lung cancer (NSCLC) accounts for 80% of cases.
[10].

Smoking and neuroendocrine factors are associated to SCLC, which is highly tumorigenic and metastatic in the primary and secondary bronchi.

NSCLC subtypes include: Lung cancer subtypes include squamous cell carcinoma (25%), adenocarcinoma (40%) and large cell lung carcinoma (15%). [11].

Computer aided diagnosis systems that employ the FPMC algorithm for segmentation have been shown to increase diagnostic precision, as stated by Gomathi et al. [2010] [11]. After cancer nodules have been segmented, a rule-based method is used to categorise them. Learning is accomplished with the assistance of an Extreme Learning Machine for improved categorization. Medical image analysis relies heavily on segmentation, as Patil et al. [2009] [4] emphasised. It's useful for determining whether or not a given image has signs of sickness. Estimating textural characteristics employs the Gray Level Co-occurrence Matrix (GLCM) method. As well as the TB database, it is used on small-cell and non-small-cell types of lung cancer images.

3. Proposed Methodology

Following the processes in Fig. 1, the proposed Lung Cancer Detection System can pinpoint the precise locations where cancer has spread. A pulmonary nodule shows up as a round lump on an X-ray of the lungs [13]. Adjacent anatomical formations can deform it. Size and spread in the lungs are unconstrained. Groups of pulmonary nodules are distinguished by the thin structure that links the nodule to the surrounding vessels [16]. Pre-diagnosis methods aid in identifying high-risk cases of lung cancer at an early stage [9].

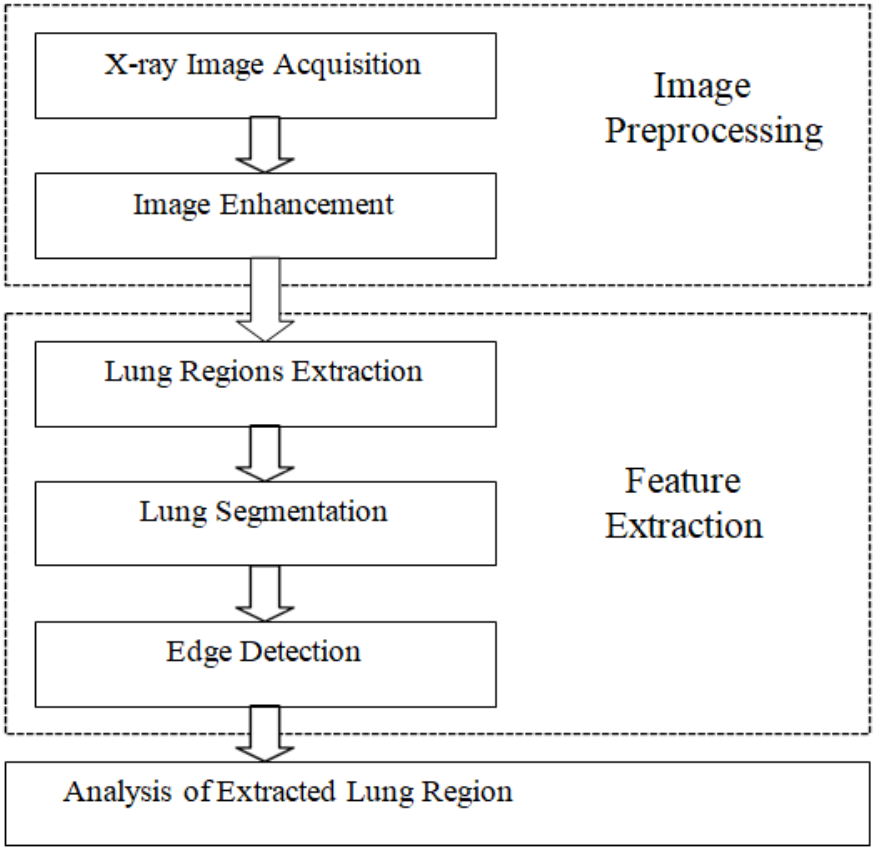


Figure2: Basic structure of proposed model

The author undertook the process of denoising, also known as enhancing the image's structure and contrast, in order to create a more high-quality X-ray image. Denoising with Median, Laplacian, and Gaussian filters is followed by an adjustment to sharpen the image's edges, remove unsharp areas, and boost contrast through histogram equalisation.

In image processing, white noise is the most frequent issue. A filter's primary recommendation is to determine the relative importance of individual pixels. Alternatively, the median filter is a common nonlinear improvement digital filtering technique that can clean up a picture without sacrificing detail [11]. See also Figs. 3(b) and 3(c)[11].

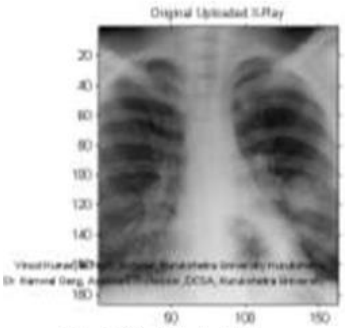


Fig. 3(a): Original X-ray

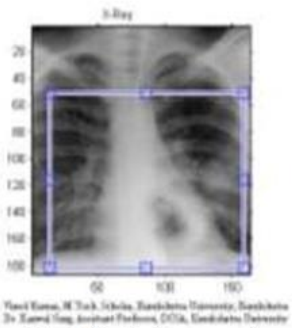


Fig. 3(b): To be cropped

Fig. 3(c): Cropped X-ray

3.1 FEATURES EXTRACTION

According to the author's feature extraction, the X-ray images of malignant nodules have a low contrast, while the surrounding healthy tissue is in between. Where T1 and T2 are two threshold values with limits T1 = 120 and T2 = 170, the author uses a multi-level threshold to categorise each point (x,y) in the image f(x,y) as belonging to object class if T1 f(x,y) T2, to the other object if f(x,y) > T2, and to the background if f(x,y) T1.

3.2 DATA SET

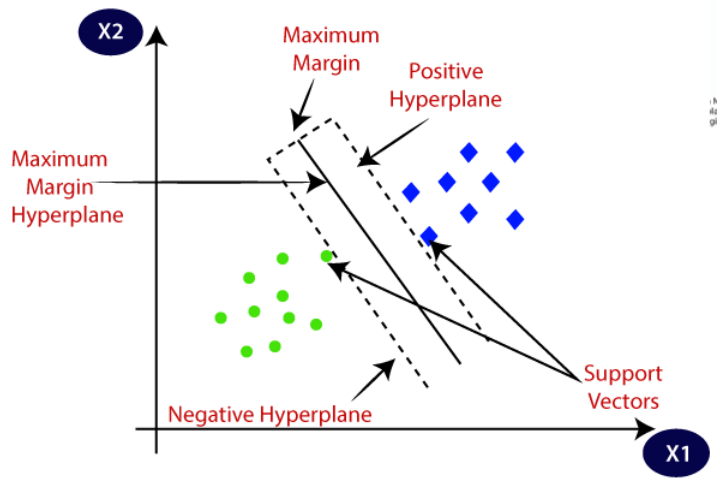
Over a thousand low-dose CT scans from high-risk patients are included in LUNA16, which is a dataset of DICOM CT scans. Among the LIDC-IDRI datasets, the LUNA16 dataset [12] stands out since it filters out photos with different types based on a variety of parameters. Annotations classed as non-nodules and nodules 3 mm are disregarded in this dataset since they are not relevant to lung cancer screening techniques. All patients are assigned a "ground value," which is essentially a binary number representing whether they have cancer (1) or not (0%).

4. PROPOSED MODELS:

4.1 Support Vector Machine (SVM)

Common for both Classification and Regression tasks, SVM is a popular Supervised Learning technique. However, its primary application is in Machine Learning for Classification issues. To classify fresh data points efficiently in the future, the SVM algorithm seeks to find the optimal line or decision boundary that divides the space into n distinct classes. The term "hyperplane" is used to describe this optimal decision-making boundary.

A hyperplane can be constructed with the help of SVM, which selects the most extreme points and vectors. The name "Support Vector Machine" refers to the algorithm's use of "support vectors," which are the most extreme situations. Take a look at the diagram below, which uses a decision boundary (or hyperplane) to classify items into two groups:



The SVM algorithm has several applications, including face identification, picture classification, text classification, etc.

The SVM can be either a If a dataset can be split into two groups along a straight line, we say that it is linearly separable, and we employ a classifier called a Linear Support Vector Machine (SVM) for such data. Non-linear support vector machines (SVMs) are used to classify data that cannot be reliably divided along a straight line.

Hyperplane: In n-dimensional space, there may be numerous lines or decision borders that can separate the classes. It is required to select the most effective decision border for classifying the data points. SVM hyperplane, optimum boundary.

Support Vectors: Support Vectors are data points closest to the hyperplane that impact its position. These vectors support the hyperplane.

4.2 DECISION TREE

Even while the supervised learning method known as the "Decision Tree" can be used to solve both "Classification" and "Regression" problems, it is more commonly employed for the former. It is a classifier in the form of a tree, with internal nodes standing in for the

features of a dataset, branches for the rules used to make decisions, and leaf nodes for the final results.

There are two types of nodes in a Decision tree: the Decision Node and the Leaf Node. Leaf nodes represent the results of previous decisions and do not contain any additional branches, whereas Decision nodes are used to make any decisions and have numerous branches. The characteristics of the provided dataset are used to make the determinations or conduct the tests. It's a visual aid for discovering every conceivable answer to a problem or decision under specific constraints. The name "decision tree" comes from its resemblance to a tree's structure, in which a central node grows into a network of branches. The CART algorithm, which stands for "Classification and Regression Tree algorithm," is used to create a tree. A decision tree consists of a series of questions, each of which is followed by a branching structure determined by the answer (Yes/No).

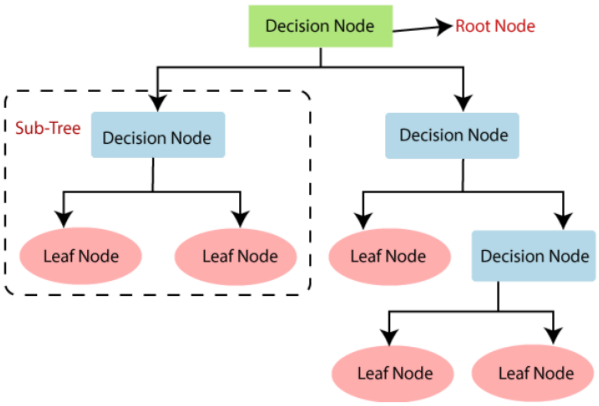


Figure 4: Basic structure of DT.

4.3 K-NEAREST NEIGHBOUR

KNN is an acronym for "K-Nearest Neighbor," which describes a type of neighbouring relationship between two nodes. An example of an algorithm for machine learning in which the data is governed by human oversight. Both classification and regression problem statements can be solved with the approach. Assigning a value of 'K' to the number of nearest neighbours of a new unknown variable that needs to be predicted or categorised.

This is also the basis for the KNN algorithm. In order to determine what category a new data point falls under, it seeks out all of its nearest neighbours. That method relies on the physical

separation between objects. We have a new data point x1, and we want to know if it belongs in the first category, Category A, or the second, Category B. A K-NN algorithm is required for this kind of problem. K-NN is useful for quickly determining the dataset's class or category. Take into consideration the illustration below:

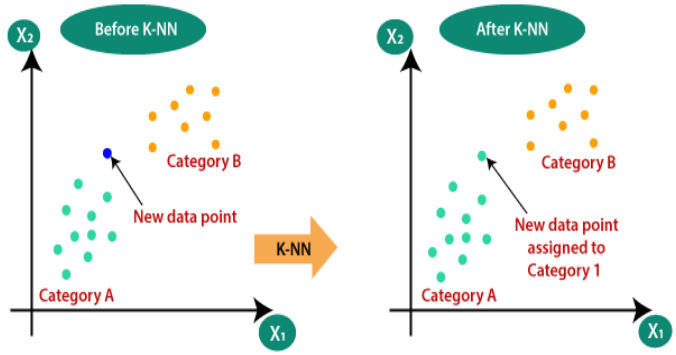


Figure 5: KNN model

4.4 Logistic Regression

In the first half of the twentieth century, scientists in the life sciences began employing the statistical method of Logistic Regression. Many subsequent applications in the social sciences made use of it. When the target variable is a categorical dependent variable, logistic regression is utilised.

Model

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

Sigmoid Function

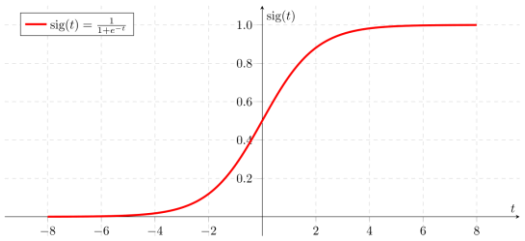


Figure 2: Sigmoid Activation Function

If 'Z' goes to infinity, Y(predicted) will become 1 and if 'Z' goes to negative infinity, Y(predicted) will become 0.

5. RESULTS

Models from the field of machine learning were applied to the LUNA16 dataset for this study. Below is a table comparing the efficacy of several models. 1

Table.1 Accuracies of different machine learning models

S.NO	Models	Accuracy
1	SVM	92.5
2	DECISION TREE	90.1
3	LR	93.2
4	KNN	88.9

One can see that Logistic Regression has better accuracy than the rest of the models in the table. Figure 6 depicts the reliability of these models.

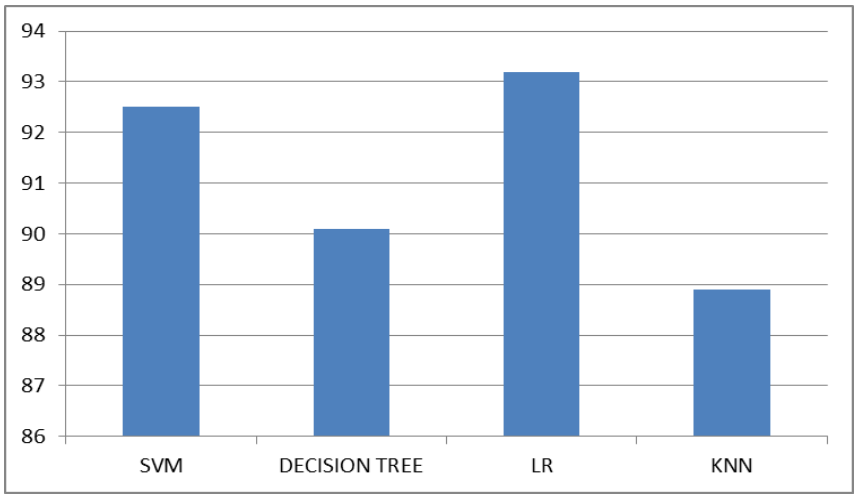


Figure.6. Performance of models

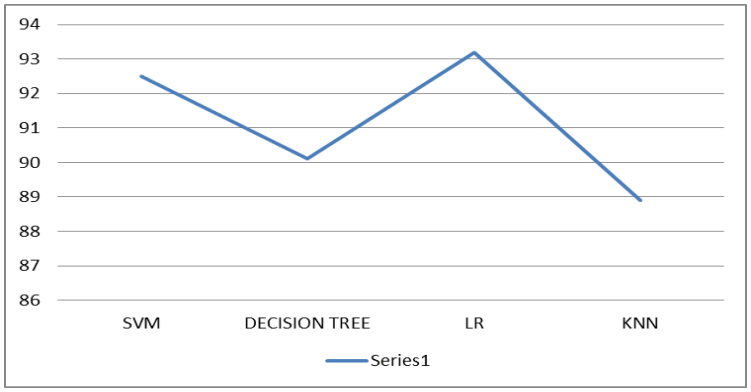


Figure.7. LR model performance

6. Conclusion

Here, we use several machine learning models—including support vector machine (SVM), logistic regression, k-nearest neighbour, and decision tree—to determine whether or not a given individual has lung cancer. Logistic regression, compared to the other models, offers the best results. We used the LUNA16 dataset for this identification. These days, all fields are abandoning machine learning models in favour of Depp learning models due to their superior accuracy. A variety of deep learning models will be used in the near future to improve performance and accuracy in detecting lung cancer.

REFERENCES

1. American-Cancer-Society. Cancer facts & figures 2014. *Atlanta: American Cancer Society.* 2014.

2. Travis WD, Brambilla E, Riely GJ. New pathologic classification of lung cancer: Relevance for clinical practice and clinical trials. *J Clin Oncol.* 2013;31(8):992–1001.

3. Rubin P, Hansen JT. *Tnm staging atlas with oncoanatomy*. Lippincott Williams and Wilkins; 2012

4. Murray N, Coy P, Pater JL, Hodson I, Arnold A, Zee BC, Payne D, Kostashuk EC, Evans WK, Dixon P, et al. Importance of timing for thoracic irradiation in the combined modality treatment of limited-stage small-cell lung cancer. The national cancer institute of canada clinical trials group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 1993;11(2):336–344.

5. Siegel R, Ward E, Brawley O, et al. Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *CA Cancer J Clin.* 2011;61(4):212–236.

6. Jemal A, Bray F, Center MM, et al. Global cancer statistics. *CA Cancer J Clin.* 2011;61(2):69–90.

7. Howlader N, Noone AM, Krapcho M, et al., editors. *SEER Cancer Statistics Review, 1975–2008.* Bethesda (MD): National Cancer Institute; 2010. Available at: http://seer.cancer.gov/csr/1975_2008/, based on November 2010 SEER data submission, posted to the SEER web site, 2011.

8. Kohler B, Ward E, McCarthy B, et al. Annual report to the nation on the status of cancer, 1975–2007, featuring tumors of the brain and other nervous system. *J Natl Cancer Inst.* 2011;103:1–23.

9. Herbst RS, Heymach JV, Lippman SM. Lung cancer. *N Engl J Med.* 2008;359(13):1367–1380.

10. Satla, S.P., Sadanandam, M., Suvarna, B. (2020). Dangerous prediction in roads by using machine learning models. Ingénierie des Systèmes d'Information, Vol. 25, No. 5, pp. 637-644. https://doi.org/10.18280/isi.250511