A NOVEL APPROACH FOR LUNG CANCER DETECTION USING FILTER FEATURE SELECTION TECHNIQUE

[1]Sandeep Wadekar, [2]Dileep Kumar Singh

School of Engineering & Technology, Jagran Lakecity University, Bhopal, India

sandy12dec@gmail.com

School of Engineering & Technology, Jagran Lakecity University, Bhopal, India

dileep.singh@jlu.edu.in

Abstract – Over a million people die from lung illness each year worldwide. Since the bulk of the cells that create tumours are enclosed with one another and develop quickly, early diagnosis of lung illness can be challenging. Throughout the course of therapy, tumour detection handling systems, which are often used for the develop a cutting-edge method for lung tumour identification, a diagnosis of lung cancer is essential. Machine learning algorithms for cancer categorization and detection have recently gained popularity and acceptance. The effectiveness of cancer illness prediction using the suggested attribute selection measure is examined in this proposed study utilising a variety of supervised machine learning methods, including the support vector machine, Naive Bayes and Random Forest. Each model's effectiveness is contrasted in in an effort to identify the most effective, optimized algorithm. Experimental findings demonstrate the great computational efficiency of the suggested model in terms of accuracy. The multi-class Decision Tree classifier revealed a higher accuracy of 89.8%, 83%, and 93.1%, and each of the three datasets with 200, 500 and 1000 features extraction respectively.

Keyword- lung cancer detection, machine learning, feature selection, Healthcare.

1. Introduction

The leading cause of mortality has been determined to be lung cancer, which is extremely difficult to identify early because the majority of symptoms only manifest at the advanced stages. Compared to other cancers like breast, colon, or prostate cancers, lung cancer causes more incidences of mortality. Lung cancer can be found using a variety of methods, including sputum cytology, chest radiographs (x-rays), magnetic resonance imaging (MRI), and computed tomography (CT) [1]. However, the majority of these are costly and time-consuming to perform. Various methods can identify lung cancer in its various stages, but patients have a poor prognosis. Image processing is a quality technique that may be used to

enhance manual cancer analysis. Lung cancer can be detected early by a number of medical researchers using a study of sputum cells.

Cancer has a high death rate compared to other forms of cancer since the symptoms typically don't show up until the disease is well along. Using image processing techniques, lung scan histopathology pictures are utilized to categories lung cancer. Radiomics is a known technology that employs a quantitative, high-throughput picture feature for diagnosis. It takes the images as data and uses data mining to make predictions. The radiomic characteristics and their propensity for prediction in the detection of lung cancer have been described in several investigations.

1.1 Screening for Lung Cancer

Numerous studies have evaluated the efficacy of lung cancer screening with chest X-rays or CT scans, with or without other criteria such as sputum cytology, in asymptomatic high-risk individuals. Although early illness stage shift is an important result, death reduction is the standard for measuring screening effectiveness [2]. This is done to offset the effect of lead time bias, in which earlier cancer identification can enhance survival rate without affecting the disease's progression (i.e., if the time of death is the same). The ideal level of lung cancer screening is directed towards persons with the highest risk. In this way, lung cancer screening differs from screening for breast, colon, and cervical cancers, for which only individuals of a certain age or gender are eligible. The greatest obstacle is recruiting this so-called "difficult" population for screening studies.

1.2 Lung Cancer Identification Process

The entire work in this paper is separated into three phases:[3]

- **Image Enhancement:**

To enhance the image by removing noise, deterioration, or impedance. The following three tactics are employed for this purpose: Gabor channel produces the best results, whereas Auto improvement computation and FFT Fast Fourier Transform produce the worst results for image division.

- **Image Segmentation:**

To divide and part the enhanced images while computing their ROI.

- **Feature Extraction:**

To extract the general features of the enhanced image using the Binarization and Masking Approach.

The combination of many logical pedagogies, such as artificial intelligence, image processing, design recognition, etc., guarantees the reliability and effectiveness of Computer Aided Design Computer-aided design frameworks. Despite the fact that CAD frameworks demonstrated significant development, there is still more work to be done in lung division and other types of tumour detection. Computer-aided design frameworks continue to provide more false-positive results than an experienced radiologist and have not achieved 100% precision, affectability, and specificity, which are crucial framework metrics.[5] Figure 1 depicts the stages for the lung cancer image processing.

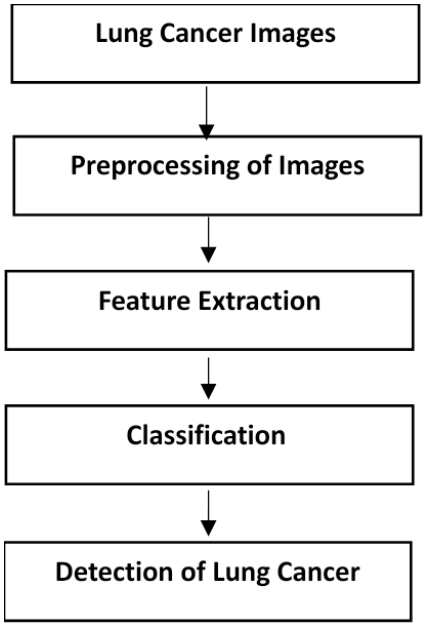

Figure 1: Stages Involved to detect Lung Cancer

1.3 Detection using Machine Learning

Medical researchers are increasingly using machine learning techniques to categories medical pictures automatically. As feature selection is an NP-hard task, it is suggested that heuristic approaches be used to determine the best subset of features. These methods may accurately forecast the future outcomes of a cancer type while discovering and identifying patterns and their interrelationships in large datasets [5],[6]. It is observed that feature selection increases the categorization of pictures; more research was conducted to refine the classifier.

There are several varieties of machine learning algorithms, however they may be grouped into four groups according to their function.

- **Supervised learning**

In this type of learning, a function is inferred from labelled training data that maps a new input to an output based on the learned function from a series of training instances.

- **Unsupervised learning**

This sort of learning does not require prior training due to the unlabeled dataset. The system is unable to automatically categories unsorted data based on similarities, differences, and hidden patterns.

- **Semi-supervised learning**

This type of learning sits between supervised and unsupervised learning since the input data is only partially labelled.

- **Reinforcement Learning**

In the process of reinforcement learning, a machine is educated by trial and error. The computer learns from its prior experiences until the exploration of all potential states, at which point it decides the optimal performance-maximizing behavior. It is typically employed for robotics, gaming, and navigation.

1.4 Challenges in Images for Lung Cancer Diagnosis

The module for image analysis is a vital component of the computer-aided detection and diagnostic system. The major objective of a module for image analysis is to identify cancer histological pictures. [10]In order to conduct the categorization work effectively, it is necessary to handle the following aspects of importance:

- **Less Availability of Dataset**

It has been observed in the literature that the majority of research on Lung Cancer histological pictures is conducted using limited datasets that are not even made available to the public.

- **Feature extraction without segmentation**

Due to the complexity of histopathology pictures, segmenting such images is a difficult process. Nevertheless, segmentation is the most important phase in an algorithm since the performance of a system depends heavily on its success.

- **Variation in picture appearance with image resolution:**

In an automated classification of histology images, the system must determine the categorization of images collected at various magnification levels [7]. In addition to increasing in complexity as magnification increases, the look of a picture changes when noise is introduced at greater magnification. Therefore, categorization gets harder to do as magnification increases.

1.5 Motivation and Objective

Lung cancer is one of the leading causes of cancer-related fatalities worldwide. The patient's medical history and histological categorization in terms of lung cancer have offered crucial information regarding the features and anatomical placements of tissues. Numerous research have portrayed the radiomic characteristics and their predictive potential in the identification of lung cancer. However, its quantitative magnitude in terms of data is vast, posing significant hurdles for categorization algorithms. To circumvent this, a symbolic method to data analysis employing vastly distinct quantitative data is offered.

Symbolic data or radiomic characteristics are used to examine further ways of feature selection for predicting the histologic subtypes of lung cancer. Once the Z score has been normalised, these features have been retrieved using Principal Component Analysis (PCA) and the fusion performed via concatenation.

2. Literature Survey

This chapter analyses the available lung disease detection techniques currently in use. Diverse approaches to addressing and enhancing the viability of tumour growth site identification have been proposed.

Diciotti et al. [6] devised a computer approach that groups images according to their similarity. This approach uses Histogram Equalization for image preprocessing, a feature extraction procedure, and a neural system classifier to determine if the initial state of a patient is normal or abnormal. Then, we predict the patient's survival rate based on the deleted characteristics. This technique employs the Neural Network algorithm because characteristics are selected at random and neural networks increase system performance. The suggested technique cannot account for low-quality photos since the system cannot predict tumour cells.

Zang et al. [15] presented a method called one-class kernel principle component analysis (KPCA), in which individual features were extracted from all class images and then a KPCA model was trained on each feature separately. After the same procedure was carried out on all of the images included in the remaining classes, the trained KPCA models were integrated to draw a conclusion. The KPCA method attained 92% accuracy on histopathology images, which was a major breakthrough.

Prabukumar et al. [11] Computer Aided Diagnostic System (CAD) aids in the early identification of lung cancer nodules in Computed Tomography (CT) pictures of the chest. Support Vector Machine and Extreme Learning Machine (ELM) are used for categorization. The dataset yields 2,474 slices, from which the most appropriate slice is selected for further

processing. The dataset contains 21 malignant nodules, eight of which are less than 2mm in diameter. The suggested method accurately identifies 14 cancer nodules when using the SVM classifier and 19 nodules when using the ELM classifier. SVM detects a false positive area of 123, whereas ELM detects 134.

Lung cancer is a condition that is regularly misdiagnosed. In the medical industry, Artificial Neural Network (ANN) plays a crucial role in resolving a variety of health issues, such as acute and even moderate disorders. In their paper [10], the authors present a feed forward neural network and a back propagation neural network for early lung cancer detection. A misdiagnosis is one of the most prevalent types of medical misconduct worldwide. Due to the correction of error detection, system performance is enhanced.

Based on the results of the survey, the following research needs have been identified:

- Traditional feature selection approaches are time-consuming and cost-effective, but feature selection methods play a crucial role in tumour prediction.
- Traditional approaches are complicated, and the feature selection method is also inaccurate. The feature extraction technique on the dataset is difficult and the cancer cell extraction process is time-consuming.
- The conventional approach fails to detect cancers in their earliest stages and smallest sizes. The classification of tumour cells using traditional approaches is erroneous.
- Existing approaches for tumour detection rely on data mining techniques, which are time-consuming and have a poor rate of accuracy in recognizing and classifying tumours.

3. Methodology

3.1 Dataset Collection

Standard Value Dataset (SVD) was used in the present work to propose a new filter-based feature selection algorithm. The SVD benefits from agents' direct communication while determining optimal subsets of features.

For classification, Decision Tree, the Naive Bayes technique, and the Random Forest Approach's performance were used. The experimental results demonstrate that the proposed approach achieved superior performance compared to current strategies.

The utility of the proposed Standard Value Dataset is evaluated utilising data collected using 200 descriptor features in the first experiment, 500 descriptor features in the second

experiment, and 1000 descriptor features in the third experiment by feature selection method. This is done in order to test the usefulness of the proposed Standard Value Dataset. 729 different values are gathered in an average Standard value Dataset. In the parts that follow, we will go through the functioning mechanism in more detail.

3.3. Working Mechanism

The following is a description of the work mechanism:

- Take images for processing as input

Output: Obtain the output score and forecast for both interacting and non-interacting indications.

Step 1: Gathers images and stores them as data objects.

Step 2: Transform the information to a stastical description.

Step 3: Initialize matrix to compute the pairwise descriptor score. By calculating BLOSUM, no score less than 62% is obtained for a size.

Step 4: Select and reduce features using a filter-based strategy.

Step-5: Introduce three kinds of dataset PCA-Best 200, PCA-Best 200 and PCA-Best 200 from Standard Value Dataset.

Step 6: Create a model using a nonlinear machine learning model.

Step 7: Conduct out the categorization and prediction.

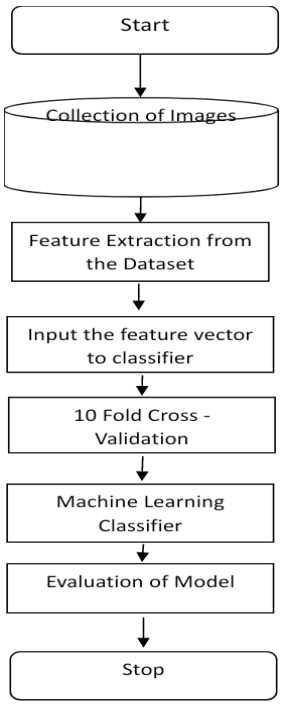

Figure 2 Proposed work flow diagram

4 Result and Discussion

Before selecting performance indicators including accuracy, sensitivity, specificity, precision, F-1 score, and mean AUROC values for analyzing simulation results, we looked at the

properties of these datasets in Section 3. The confusion matrix displays preliminary simulation results for the aforementioned methods and the suggested Standard Value Dataset. The results of the suggested Standard value dataset are shown in Table 1.

Table 1: Comparative analysis of proposed system for (a) Decision Tree, (b) Naïve Bayes and (c) Random Forest

Model	Parameter (%)	1000	500	200
Decision Tree Approach	Accuracy	0.931	0.83	0.898
	Sensitivity	0.951	0.913	0.928
	Specificity	0.859	0.741	0.789
	Precision	0.956	0.923	0.936
	F1 – Score	0.953	0.918	0.932
	AUROC	0.953	0.914	0.94
Naive Bayes approach	Accuracy	0.933	0.889	0.901
	Sensitivity	0.908	0.917	0.929
	Specificity	0.856	0.763	0.821
	Precision	0.951	0.932	0.941
	F1 – Score	0.91	0.921	0.937
	AUROC	0.958	0.922	0.934
Random Forest	Accuracy	0.953	0.912	0.936
	Sensitivity	0.971	0.921	0.931
	Specificity	0.881	0.814	0.852
	Precision	0.954	0.948	0.935
	F1 – Score	0.952	0.939	0.944
	AUROC	0.975	0.939	0.959

All nonlinear machine learning techniques are acceptable for our suggested model, although the Decision Tree performs better than the other nonlinear machine learning algorithms described. As we can see, the outcome improved as the number of features increased. All suggested systems using nonlinear machine learning algorithms produced optimum results for 1000 features.

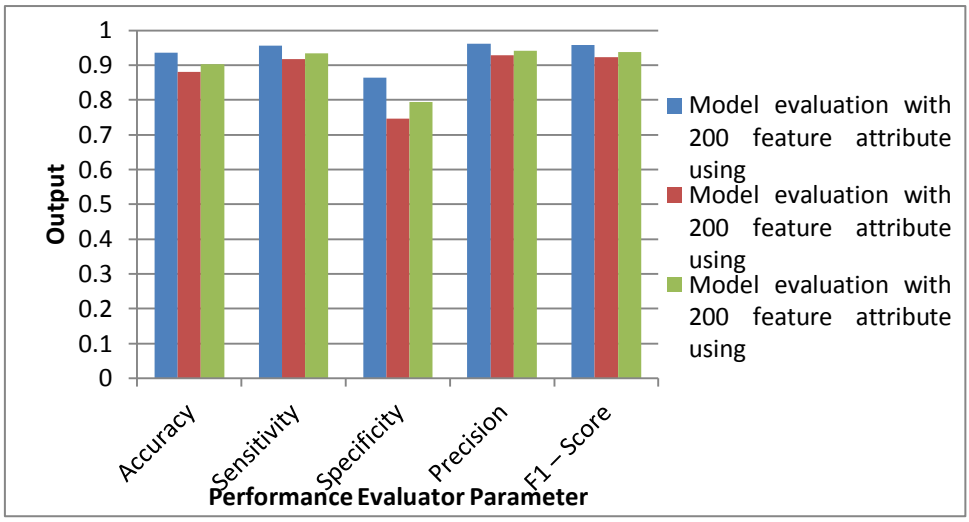

Figure 3: Comparative analysis of result for proposed Dataset

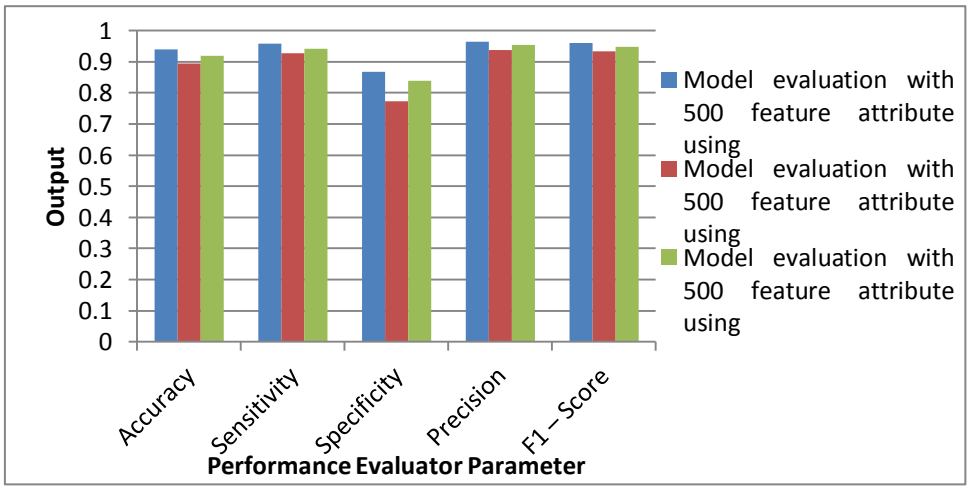

Figure 4: Comparative analysis of result for proposed Datset

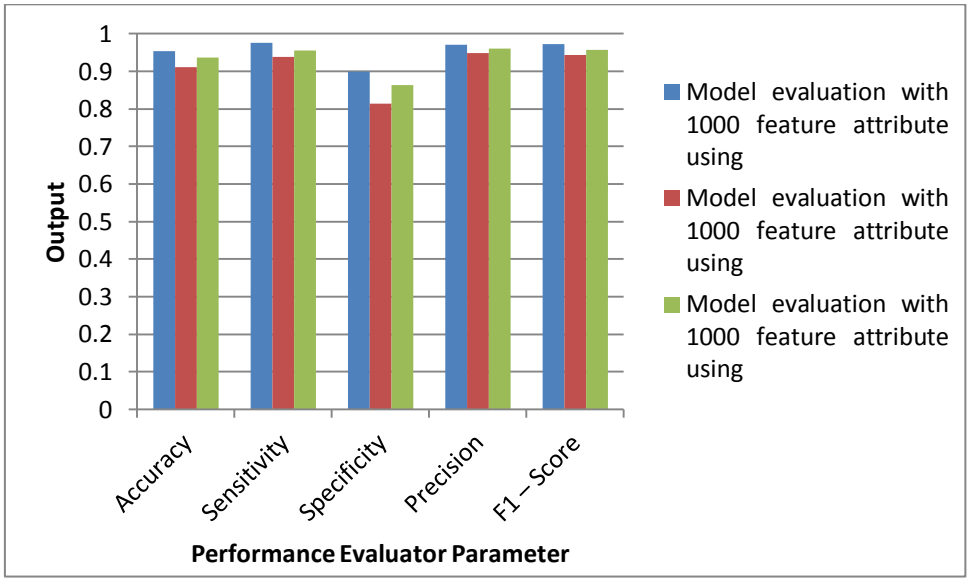

Figure 5: Comparative analysis of result for proposed Dataset

The optimality of the classifier may be determined by looking at the mean AUROC value. When compared to the other two nonlinear Machine Learning algorithms, as shown in figure 6, the Decision Tree produces the most desirable outcomes. When compared to Naive Bayes, the Decision Tree method likewise produces decent results, but less accurate predictions than Decision Tree.

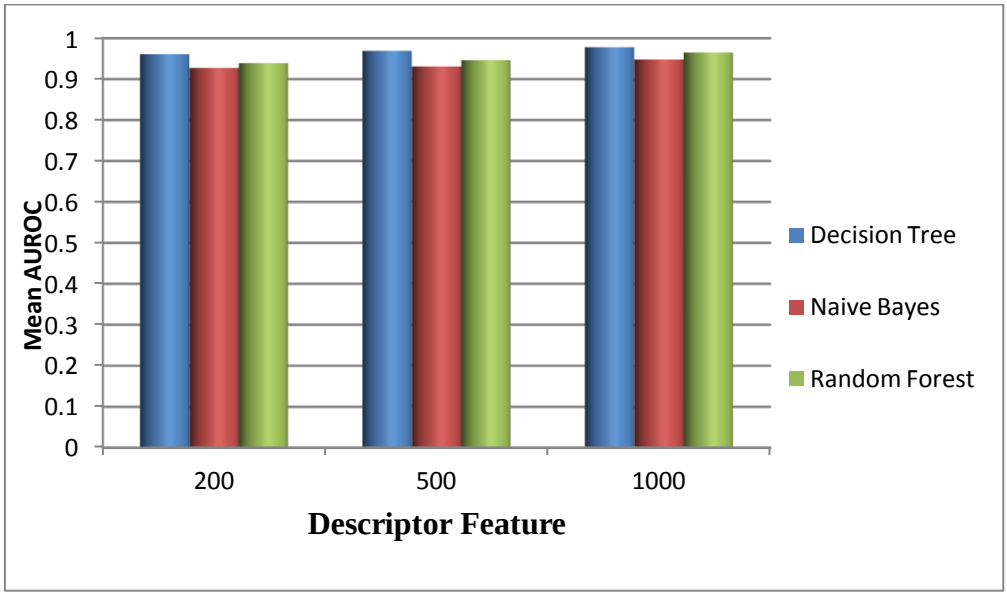

Figure 6: Comparative analysis of optimality of classifier for proposed Dataset

5. Conclusion and Future Scope

To categorise lung cancer detection based on the feature extraction approach, we have presented the multiclass nonlinear Machine learning algorithms - Decision Tree, Naive Bayes, and Random Forest algorithm in this study. Three standard datasets are used for training and testing, with feature extraction ranging from 200 to 1000.

Our suggested model is compatible with any of the aforementioned nonlinear ML techniques, however we found that Decision Tree performed very well. It turns out that the more features we used, the better the outcome. For 1000 features, we found that every suggested system yielded best results using a unique nonlinear machine learning technique.

Accuracy levels of 92.8%, 93.9%, and 95.3% were discovered by the multi-class Decision Tree classifier when applying it to datasets containing 200, 500, and 1000 features extracted, respectively. After concluding that Decision Tree yields the best results, we plan to use experimental results from deep learning models in our future work.

1. Avanzo, M., Stancanello, J., Pirrone, G., & Sartor, G. (2020). Radiomics and deep learning in lung cancer. Strahlentherapie und Onkologie, 196(10), 879-887.
2. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.
3. Burki, T. K. (2016). Predicting lung cancer prognosis using machine learning. The Lancet Oncology, 17(10), e421.
4. Araújo, et al., "Classification of breast cancer histology images using convolutional neural networks," PloS One, vol. 12, pp. 1-28, 2017.
5. Cai, Z., Xu, D., Zhang, Q., Zhang, J., Ngai, S. M., & Shao, J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molecular BioSystems, 11(3), 791-800.
6. Diciotti, S, Yagis, E., Citi, L., Marzi, C., Atnafu, S. W., & De Herrera, A. G. S. (2020, July). 3d convolutional neural networks for diagnosis of alzheimer's disease via structural mri. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 65-70). IEEE.
7. Hosni M., Abnane, I., Idri, A., de Gea, J. M. C., & Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. Computer methods and programs in biomedicine, 177, 89-112.
8. Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., & Tsuneki, M. (2020). Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific reports, 10(1), 1-11.
9. Nasser, I. M., & Abu-Naser, S. S. (2019). Lung cancer detection using artificial neural network. International Journal of Engineering and Information Systems (IJEAIS), 3(3), 17-23.
10. Prabukumar, M., Agilandeeswari, L., & Ganesan, K. (2019). An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. Journal of ambient intelligence and humanized computing, 10(1), 267-293.
11. Srivastava, D., Kumar, P., & Ghildiyal, S. (2022). A novel approach of drug repurposing in immuno-oncology therapeutic agents using machine learning algorithm. International Journal of Health Sciences, 6(S2), 9688–9702. https://doi.org/10.53730/ijhs.v6nS2.7523.

12. Srivastava, D., Soni, D., Sharma, V., Kumar, P., & Singh, A. K. (2022). An Artificial Intelligence Based Recommender System to analyze Drug Target Indication for Drug Repurposing using Linear Machine Learning Algorithm. Journal of Algebraic Statistics, 13(3), 790-797.
13. Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Translational lung cancer research, 7(3), 304.
14. Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational oncology, 14(1), 100907.
15. Zhang , Bardou, D., K., & Ahmad, S. M. (2018). Classification of breast cancer based on histology images using convolutional neural networks. Ieee Access, 6, 24680-24693.