International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Diagnosis and Prognosis of Lung Cancer & Lung Nodule Using Machine Learning Techniques

Sangeeta Devi[1], Pranjal Maurya[2], Munish Saran[3], Rajan Kumar[4], Upendra Nath Tripathi[5]

[1,2,3,4,5]Department of Computer Science,DDUGU, Gorakhpur, Uttar Pradesh, India

Abstract:
Lung cancer is a serious and challenging cancer to diagnose. It frequently results in death in both men and women; thus, prompt, precise nodule analysis is crucial to the course of treatment. Early cancer detection has been accomplished through a variety of techniques. This research compares machine learning techniques for lung cancer nodule detection. To find anomalies, we used machine learning techniques such as principal component analysis, K-nearest neighbors, support vector machines, Naïve Bayes, decision trees, and artificial neural networks. We examined every technique with and without preprocessing. According to the experimental results, decision trees produce the most accurate results with 93,24% effectiveness without image processing while artificial neural networks produce the finest results with 82,43% effectiveness after image processing.

Keyword: lung cancer, decision tree, artificial neural networks, Naïve Bayes, classification, machine learning, support vector machine and diagnosis.

I. Introduction:

One of the deadliest diseases a person can have is cancer. The late diagnosis can result in deaths worldwide. The primary function of the lungs is to supply oxygen to the human body and release carbon dioxide when performing essential bodily functions. The lungs' tissues and cells proliferate out of control, which leads to malignancy of the lungs. These masses have the potential to spread and harm nearby tissues when they develop out of control in the surrounding area. In terms of cancer-related deaths, lung cancer is the second most common in women and the first in men. An estimated 1.3 million individuals worldwide pass away from lung cancer each year. Every year in country of India, between 30,000 and 40,000 new cases of lung cancer are detected. Lung cancer symptoms could not cause serious problems until the illness is fairly advanced. The primary cause of lung cancer's extreme hazard is its ability to progress without showing any signs. About 25% of cancer patients experience no symptoms at all. The vast majority of people find out that lung X-rays from another illness cause lung cancer.

Lung cancer symptoms could not cause serious problems until the illness is fairly advanced. The primary cause of lung cancer's extreme hazard is its ability to progress without showing any signs. About 25% of cancer patients experience no signs or symptoms at all. The majority of people find out that lung X-rays from another illness or health problem cause lung cancer. Lung cancer identification at an early stage is

crucial. Because lung cancer frequently has the potential to spread quickly to the brain, liver, adrenal glands, and bones. However, the average quality of life and life expectancy have grown with the most recently treatment and methods for lung cancer. Thanks to developments in imaging methods like low-dose spiral computed tomography, lung cancer can now be identified early on.

For many years, lung cancer has been the primary cause of cancer-related mortality [1]. The incidence of lung cancer increased by 44% between 2009 and 2019 [2], and over 130,000 casualties in the US alone in 2022 were linked to the disease [1, 2]. Since early-stage lung cancer rarely exhibits symptoms, the majority of newly diagnosed patients are found to have an advanced-stage illness, which frequently carries a dismal prognosis (20.5% overall 5-year survival rate) [1, 3]. Computed tomography (CT) are the most widely used imaging modality for lung cancer screening and diagnosis due to its affordability, ease of use, and high degree of spatial accuracy of images [4, 5]. Even with the technique's inherent requirement for ionizing radiation, studies have demonstrated the accuracy of low-dose CT (LDCT) for lung cancer assessment, with a standard relevant dosage of approximately 1.5 mSv [5, 6].

Yet, there are two primary obstacles preventing lung cancer detection initiatives from being implemented more widely. The availability of people and technology is one issue since radiology capacity might not be enough to fulfill demand [8, 9]. Given the significance of strong and excellent training advised for those responsible for analyzing the pictures, the second potential flaw is associated with false positive instances and excessive diagnosis, which is closely associated with the first. According to earlier research, the harmless incidence of a diagnostic procedure after the discovery of a nodule could reach 40%. This underscores the significance of thorough nodule screening prior to more invasive treatments in order to reduce surgical risk, avoid needless challenges and problems, and avoid loss of lung capacity.

In Section I of this study, we describe lung cancer and the work associated with lung cancer detection. In Section II, we provide detailed details regarding the preprocessing, machine learning, and data set utilized to detect tumors. In Section III, we go over the findings based on the statistical evaluation.

II. METHODS:

A. Preprocessed Data:

The Japanese Society of Radiological Technology (JSRT) Standard Digital Image Database was used in this investigation [12]. The dataset in Fig. 1 contains 247 CT (computed tomography) pictures. The photos are labeled as having a lung nodule or not. Ninety-three photos are labeled as lacking a nodule, whereas 184 photographs are found to have one. 2048 x 2048 matrix data with two bytes for each byte is how images are preserved. For classification, the dataset is split into two parts: the first is the train set and the second is the test set. By selecting at random from the dataset, we use 70% of the data as the train set and 30% as the test set.

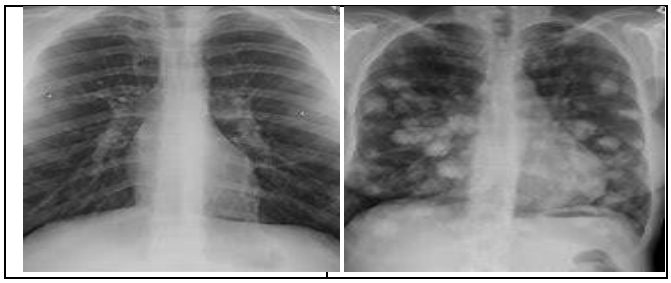

Fig. 1. Tomography images of dataset (nodule image at left side and non-nodule image at right side)

B. Principal Component Analysis- Eigen Vectors:

Principal component analysis (PCA) is a statistical method that is frequently used in face recognition and picture reduction applications to discover new patterns in high-dimensional data. The best performing feature has the most variance and diffusion, as recognized by PCA. There is lots of information in the feature with a huge variance because the entropy is too large for that feature. Another name for large eigenvectors is major components. We categorize the data using eigenvectors in order to find lung nodules. We identified the train set's most similar eigenvector for a fresh set of test data. The data's class is established by comparing it to the founded, comparable data from the test set.

Another method for reducing dimensionality is to employ Principal Component Analysis (PCA), an unsupervised linear conversion approach. With high-dimensional data, the PCA helps us determine maximum variances and reflecting on a new subspace that is equivalent to or smaller than the original. The direction of the largest variance is represented by the perpendicular axes of the new subspace.

Measurement Classification with Eigen Values

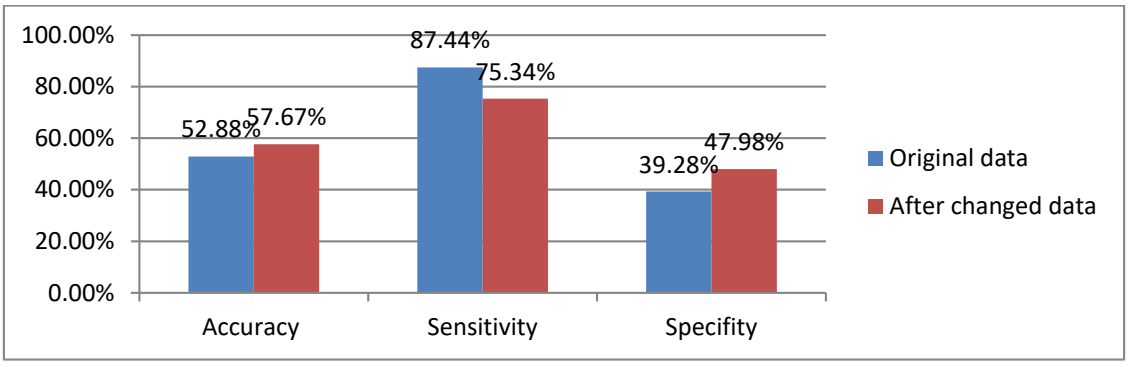

Fig 2: Measurement performance of Eigen value before PCA and after PCA

We also assessed the Eigen Vector's performance following the PCA-assisted 1/8 data dimension reduction. Figure 2 compares the outcome of eigen vector classification before and after PCA.

C. K-Nearest Neighbors:

The most well-known, traditional, straightforward, and successful pattern classification technique was put out by Thomas M. Cover and Peter E. Hart. The KNN algorithm locates the sample data point and uses the k value to decide which neighbor is closest. The neighbor number (k), the weighting technique, and the distance measure are the three crucial and effective parameters in the operation of the KNN algorithm.

In this research, we compared the method's performance using KNN with various k values. We determined that k equals 2, 3, and 5 in terms of neighbors. The performance metrics are listed in Table 2.

	k=4	k=5	k=7
Accuracy	0.73084	0.75435	0.73084
Sensitivity	0.95397	0.87475	0.85556
Specificity	0.54957	0.57778	0.56283

KNN Performance with k different value

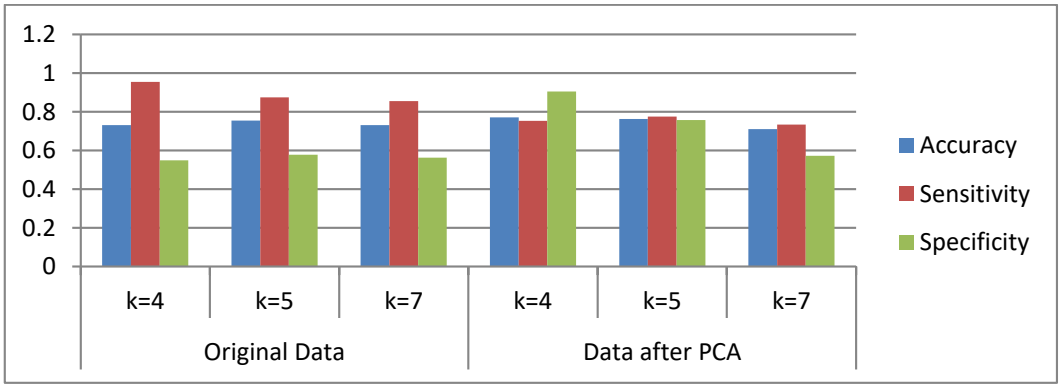

Fig.3 Measurement the KNN Performance with different k value before PCA and after PCA

As seen in Fig. 3, we also calculated the f-score of the approaches that provide more broad details on the performance. As illustrated in Fig. 4, dimension reduction of the data reveals some significant changes, in contrast.

D. Support Vector Machine:

A non-parametric classification technique based on statistical learning theory is called Support Vector Machines (SVM) [13]. This approach for supervised machine learning can be applied to issues involving regression or classification. SVM was created for binary classifications, and with a little amount of sample data, precise classification results can be achieved. It is a technique where the feature space boundary between classes is found using the best algorithm.

One of the easiest and most effective methods for detecting lung cancer is SVM. As with earlier approaches in this work, Fig. 5 provides information on SVM with both original and modified data.

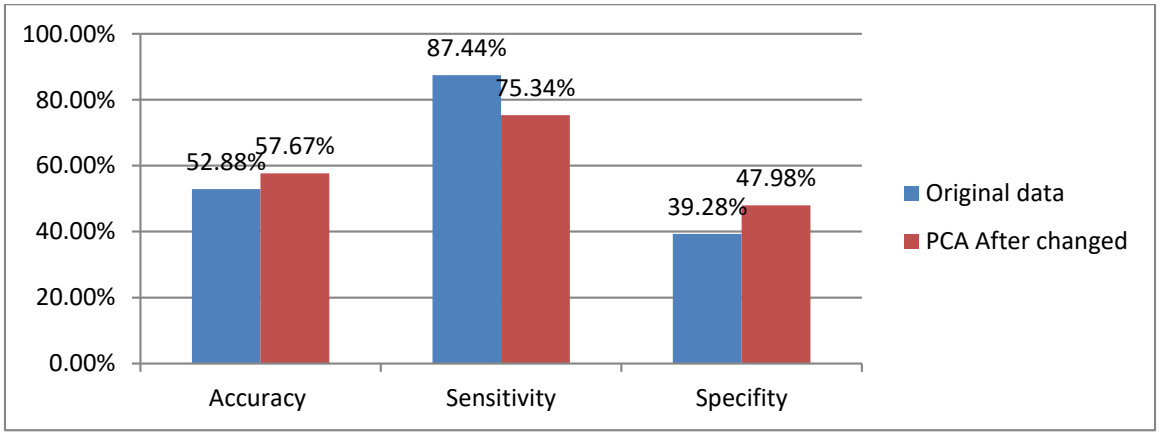

Fig 4. Measurement performance of SVM data before PCA and after PCA

E. Decision Tree:

A supervised machine learning technique called Decision Trees,[15] where data repeatedly divides data based on a given parameter. It consists of two entities: the leaves and the decision node. Leaves stand for choices and outcomes. Decision nodes show the division of data. Flowcharts and decision trees are comparable. A node represents each attribute. The structural components of the tree are its branches and leaves.

Although it is a simple procedure, it is highly effective in detecting lung cancer. We attempted to use dimension reduction using source data and after PCA to examine its effects on decision trees. Figure 6 makes the result of reductions quite evident.

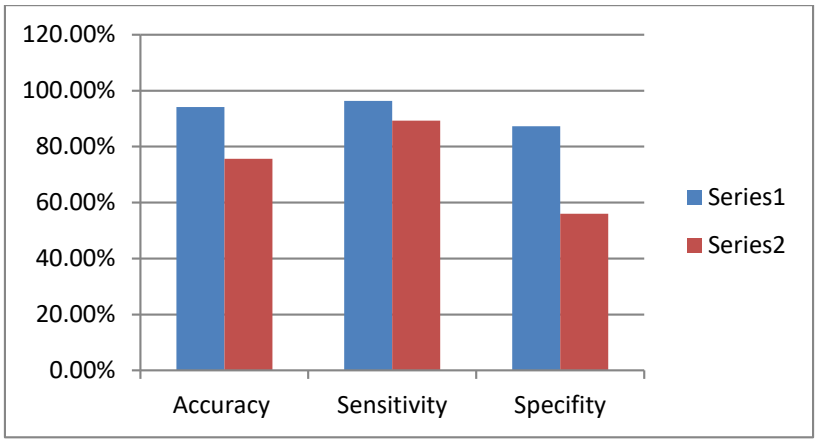

Fig 5. Measurement performance of DT data before PCA and after PCA

F. Artificial Neural Networks:

A popular supervised machine learning technique for picture categorization issues is the artificial neural network (ANN) [16]. The input, hidden, and output layers make up an ANN, in that order.

We employed a feed forward neural network with ten hidden layers in our investigation. Rather than utilizing the full-size photograph, we decreased its dimensions to a ratio of 1/32. ANN statistical measurement values are displayed in Fig. 6.

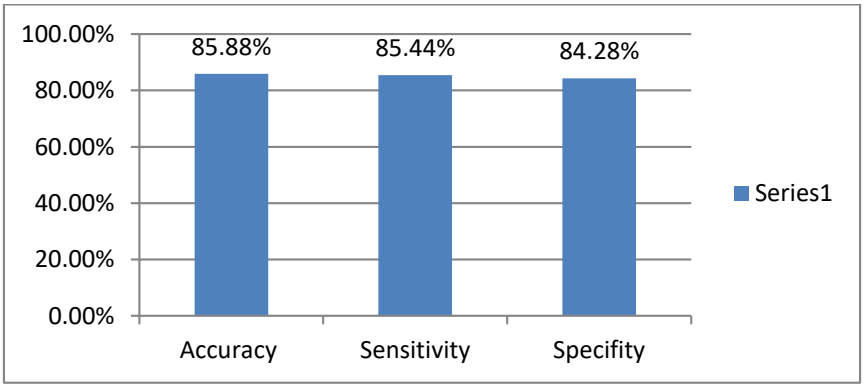

Fig 6. Measurement performance of ANN

III. ANALYSIS, AND PERFORMANCES OF ALL EXPERIMENT

The confusion matrix is the most often used technique in machine learning for assessing the performance of models created from data sets including predefined target values. A True Positive (TP) indicates that the test data's class and the model's class for that data are the same. A False Negative (FN) indicates that the test data's class and the model's class for that data are different. A False Positive (FP)

occurs when a positively classified real value is actually negative. When a value is classed as True Negative (TN), it indicates that its genuine value is negative. The ratio of properly diagnosed nodules to total nodules is known as accuracy.

The ratio of correctly recognized nodules to affirmative nodules is known as accuracy. The ratio of correctly classified to positively classified cases is known as recall. The harmonic average of recall as well as accuracy is known as the F-measure.

In order to compare machine learning strategies, we created confusion matrices for each method and assessed recalled, precision, reliability, and the f-measure value, as indicated in Table IV.

	KNN	SVM	DT	ANN	Eigen Vectors
Accuracy	75.55%	54.00%	96.34%	--	88.57%
Accuracy (after PCA)	76.88%	57.54%	78.98%	84.44%	80.13%
Precision	75.56%	36.46%	95.23%	--	92.23%
Precision (after PCA)	83.56%	46.69%	72.45%	92.43%	84.13%
Recall	87.46%	86.00%	98.11%	--	88.32%
Recall (after PCA)	79.34%	74.45%	88.45%	84.56%	88.45%
f-measure	82.00%	47.76%	96.23%	--	91.21%
f-measure (after PCA)	82.00%	57.10%	79.23%	87.70%	85.64%

CONCLUSION:

Several machine learning techniques were used in this work to identify lung cancer from chest radiographs. Additionally, we employed PCA to minimize chest radiograph dimensions by a factor of 1/8. Loss of features would result from reducing dimension. It did not result in a significant loss of information in our circumstance. Additionally, it improved SVM and KNN accuracy (where k=2 and k=3). The accuracy results have minor effects that might be disregarded in order to minimize storage space and cut down on processing time[12].

Due to the large amount of data, we were unable to apply Naïve Bayes and a 10-layer Feed Forward Neural Network to our photos without PCA. Despite the fact that neural nets achieve comparatively higher precision than other methods used in machine learning. In the initially collected data, Decision Tree yielded the best results across all performance metrics.

The image set was used before any noise reduction techniques were used. To achieve better accuracy, different noise reduction techniques may have been used in further study. A full chest radiograph was also utilized in this investigation as input information for machine learning techniques. Only the lung portion of radiographs could be removed to use as input using imaging segmentation techniques.