Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 4
April 2024

Comparison study on Lung Cancer Risk Assignment using Machine Learning and Deep Learning

Maragani Datta Pavan
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Cherukuri Bhavath Ram
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Dinesh Chowdary Vemulapalli
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Sri Teja Cheemakurthy
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

M Kavitha
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Vijaya Chandra Jadala
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Abstract
Lung tumour is the expansion of malignant cells in the lungs. The rising frequency of cancer has brought about a rise in the rate of death for both women and men. Uncontrolled cell multiplication in the lungs is a condition known as lung tumor. Although it's not possible to prevent lung cancer, but risk can be decreased. To keep lung issues from developing into long-term, serious illnesses, early diagnosis and treatment are essential. SVM, Logistic regression, KNN, and Random Forest Decision trees were among the techniques for classification used to assess the lung cancer risk prediction. further enhanced the technique's accuracy by using boosting classifiers like gradient, Ada, and XG classifiers. By assessing the effectiveness of classification algorithms, the primary goal of this study is to diagnose lung cancer early enough. In this research, we present an analysis comparing Machine learning methods with Deep Learning-based Computer Assisted Diagnosis methods. To tackle these issues, our study shows good results in lung disease identification even with a smaller dataset, offering a workable solution to the problems in the area of medical image processing. When compared to other machine learning algorithms, SVM has produced more precise results. Furthermore, we provide an outline of the benefits and drawbacks of the current set of lung cancer identification algorithms.

Keywords: Boosting Classifier, Logistic Regression, SVM, Machine learning, image classification, deep learning, computer aided diagnosis.

1. Introduction

Globally, lung tumour is a serious issue, and early detection and accurate prognosis are essential for the best possible care and outcomes. In nations with low or middle incomes, where availability of medical and preventative treatments is sometimes limited, the risk of lung cancer is especially high. Both machine learning and deep learning approaches have shown promise in a number of subjects, including medical diagnostics. Numerous tumours, such as breast cancer and melanoma, and colorectal cancer, have been successfully predicted and classified using these techniques. Computerized lung nodule identification is a promising study area that greatly improves lung nodule detection framework performance. Additionally, through the use of a variety of imaging data sources, they have demonstrated efficacy in the outlook and the identification of lung cancer [1].

The application of algorithms based on deep learning and machine learning for lung tumour studies has been the subject of a substantial amount of research in the past few years. Predicting lung tumour risk, occurrence, and patient survival rates with malignant tumours has shown promising results from these investigations. By leveraging large datasets and

sophisticated algorithms, machine learning can transform data from measurements into useful models for prediction. Numerous computer-aided systems have been thoroughly investigated for the purpose of recognizing and classifying lung cancer. In medical imaging, these systems do better at detecting cancer. Four phases are typically involved in the computer-based lung cancer identification system: image processing, extraction, selection of features, and classification. As this system depends on processing the images to extract consistent features, picking and classifying features are the most important processes in enhancing the CAD system's sensitivities and accuracy. Deep learning-based computer-aided design tools can significantly increase the effectiveness and precision of medical diagnostics, according to earlier research. Unlike traditional CAD systems, deep learning-based CAD systems can extract high-level characteristics from source pictures using multiple network frameworks [13].

Deep learning-based CAD tools can be time-consuming and have low sensitivity and high FP, among other drawbacks. For identifying lung cancer, a deep learning-based detection system that is quick, affordable, and extremely sensitive is thus essential. Research on machine learning-based lung imaging approaches focuses primarily on the identification, categorization, and separation of malignant tumours in the lungs. To improve the performance of machine learning models, researchers primarily work on creating new network architectures and loss function algorithms. Recent reviews of Deep learning techniques have been provided by several research groups. Data from earlier investigations has been included in this research [13].

The diagnostic and medical planning for patients were eventually boosted by these models' excellent accuracy in differentiating between both malignant and benign nodules in the lungs. Lung cancer risk assessment could undergo a revolution thanks to these developments in machine learning and deep learning methods. They can help medical practitioners identify those who are at high risk so that they can provide them with focused screening and treatments. This study analyses and compares classification algorithms for lung tumour identification in automated identification systems that depend on various deep learning frameworks. Overall performance based on classifying classes and the location features was computed using volumetric data and Based on the collected lung cancer dataset, this study yields distinct outcomes for every classification. Tumour nodule detection and classification are handled using classification techniques, which utilise the features extracted as training features. The trained model is then utilized to sort nodules using a network structure. The accuracy levels were acquired after classifiers like KNN, SVM, and Logistic Regression were put into practise. By using these neural networks, with little changes to the CT scans, tumours in the lungs can be categorised as malignant tumours [16].

2. Literature Review

TABLE I shows the Several research studies and publications use deep learning and machine learning to analyse lung cancer risk assignment. These articles provide a description of some of the well-known approaches and research conducted by different authors, as well as a range of methodology, datasets, and performance measures utilised for machine learning-based lung cancer prediction.

Author Name	Author Contribution	Data set	Parameters used	Observations
Mokhled S. AL-TARAWNEH et.al, 2012 [1]	Author compared proposed methodologies and found that pixel's proportion and mask-labelling are the major features that can be discovered for comprehensive picture comparison.	It begins by gathering a variety of images—both normal and abnormal from the IMBA Home through VIA ELCAP	Low performing pre-processing methods on the Gabor filter is used within the Gaussian kernel for the stage of picture quality analysis and augmentation.	When compared to other procedures, the outcomes from the suggested technique are quite encouraging. Pixel fraction and mask-labelling with high precision and consistent operation are the major features that can be recognised for

	The most important factors are image quality and precision.	Public Access.		accurate picture comparison.
C.Bhuvaneswari et.al, 2014 [2]	The author identified and categorised lung disorders by efficiently extracting features using instant invariants, selecting features using a genetic algorithm, and classifying the outcomes using Naive Bayes classification and decision trees.	Images from computed tomography and publically accessible datasets were employed in image processing algorithms.	Performance metrics for recall and precision are calculated.	The performance metric demonstrates that the decision tree classifier outperforms the naive bayes model in terms of producing more accurate output during training, testing, and classification.
Subrato Bharati et.al, 2020 [3]	The author suggested a Vds Net system that makes lung disease identification easier. Vds Net has a significantly shorter training period but a slightly lower validation precision.	In Kaggle, where there are 4143 lung or lung X-ray photos available, Openi was the most widely accessible.	There have been determined performance metrics. Calculations are made to determine the precision statistics as well as performance metrics for loss.	With a validation accuracy of 73%, VDSNet outperforms the majority of competing architecture frameworks. The validation precision value for VDSNet is 73%. The processing of the huge scale dataset provides certain difficulties. Small dataset can therefore offer good accuracy.
Ruchita Takade et.al, 2018 [4]	The author-proposed 3d multi-path VGG-like network framework built from datasets. Results are integrated using the UNet and a 3D VGG-like network framework. Malignancy level is determined, and the lung nodules are categorised.	This study utilises data from the TCIA repository, LIDC-IDRI, and the 2017 Kaggle Data Science Bowl.	The overall F1 score has also been computed for training and testing the data, and also used to calculate the accuracy, recall, and loss to assess the efficiency of the constructed model.	Using U-Net architecture, CT scans are separated with a precision of 95.60% to identify tumours in the lungs, classify them, and determine their level of malignancy.
Meet Diwan et.al, 2021 [5]	The author developed transfer learning-based lung illness categorization pipeline and tested it	The largest Contributions from the Kaggle and Scopus repositories.	The parameters used to verify the model's efficiency are accuracy and loss measures.	All five network framework structures can detect tuberculosis, pneumonia, COVID, and lung cancer, however Mobile Net has

	on small lung imaging data and used the U-net segment network, the InceptionV3 model classifier. The framework was compared with current models.	The information is taken from a variety of open-source datasets. LIDC-IDRI and Chest x-ray datasets were also used.		more precision than the others.
Inam Ullah Khan et.al, 2022 [6]	the author developed an MCCT model that is based on the CCT model and is contrasted with six learning transfer model including Vgg-19, Vgg-16, RESNET-152, RESNET-50, RESNet-50V2, and Mobile-Net.	The dataset was created using a freely accessible COVID-19 image dataset that was gathered from Kaggle that contains 21149 X-ray chest (CXR) images.	Predictions for every category are displayed in the proposed method's confusion metrics. The efficiency of the method is given by loss and accuracy.	Using 32 x 32-pixel input photos to train suggested technique, an accurate lung disorder classification framework was developed. Smaller size images may lead to quicker times and use less storage when training with large datasets. They also minimise the number of trainable parameters.
N. Sudhir Reddy et.al, 2022 [7]	The author presented a deep neural network and bio-optimized based Ldcc-Net approach to segment and various classes identification of lung tumours and analyses LDDC-Net efficiency with traditional methods.	CT scans with identified tumours are part of the Lung picture Database Consortium's (LIDC-IDRI) image library.	the use of various metrics, including peak signal to noise ratio, structural similarity score metric, mean square error, entropy, variance, the LDDC-Net is recommended with traditional models.	Author created a DLCNN model, used feature extraction in testing and training phases, and classified moderate and severe lung tumours. The model's outcomes demonstrate that, when compared to traditional approaches for lung disorder identification, the suggested LDDC-Net led to higher segmentation as well as classification performance.
Goram Mufarah M et.al, 2022 [8]	The author developed a pre-trained model called VGG19, then three blocks of convolutional neural network for extraction of features and a fully connected system for identification of TB, lung tumour, pneumonia and lung diseases.	From publicly accessible datasets, X-ray photos of respiratory infections, TB, lung cancer, pulmonary opacity, and the latest Covid-19 were accessed and gathered.	The author used precision, recall, loss and F-measure as parameters to analyses the proposed model.	According to the test results, the suggested VGG19 + CNN executed more effectively than preceding works with 96% accuracy, 93% recall, and 95% F1 score.
Lanjuan Li et.al, 2018 [9]	The author provided a new technique	The dataset was created	Performance measures such as	The author assesses and evaluates the accuracy

	called AECNN, which builds a deep neural structure model of AECNN framework, utilises and analyses the entire ROI image, and realises different identification of tuberculosis by combining the feature extraction process of CNN.	using a freely accessible tuberculosis CT images dataset that was gathered from a laboratory cooperating local hospital.	Accuracy, Recall, Loss, f1-score are evaluated, compared with different models	of the modified CNN model in identifying tuberculosis in comparison to machine learning algorithms such as SVM, KNN, Random forest classifier, decision tree, and finds that it performs better.
Chang Liu et.al, 2017 [10]	To cope with unstable X-ray images, the author presented a new technique using convolutional neural network. In categorising numerous TB symptoms, our technique significantly increases accuracy.	Dataset obtained from Peruvian partners at "Socios en Salud", Peru. This dataset consists of 4701 pictures, of which 453 are classified as normal and 4248 as abnormal.	Confusion metrics are used to validate performance. The overall F1 score, precision have also been computed for training and evaluating the data.	The author evaluates that to train the network, use shuffle sampling and cross-validation to obtain greater precision in comparison to other existing techniques.
Nidal Nasser et.al, 2021 [11]	The author developed a framework model based on Deep learning method for the Covid-19 detection using chest x-ray.	Covid-Chest x-ray dataset and Chex-Pert dataset are used.	The RESNET-50 model is tuned after 120 Epochs. The batch size 30 is needed.	The ResNet50 model's probability value determines whether the test's image is in the Covid-19 class or not.
A. Asuntha et.al, 2020 [12]	for the processing, the contrast of input scan images is enhanced by using the Histogram Equalization technique.	LIDC datasets are used.	two-dimension sizes were 10 and 30. The dimensions 10 and 30 in the FPSO corresponded to the numbers of iterations, which were set at 1000 and 2000.	For segmenting the lung tumour nodules, many techniques such as K-means, FCM and Ant Colony algorithms are applied.
Lulu Wang et.al, 2022 [13]	The author discussed new developments in deep learning-based nodule prediction and lung cancer. Lung nodule segmentation, detection, and image pre-processing are all included.	379 unduplicated lung nodule CT pictures and 50 LDCT images of the lungs are part of the Early Lung Cancer Action Programme.	anticipated regions and serves as a metric for a few semantic division and detection of objects issues.	Since nodules in the lungs can vary widely in size, shape, and appearance, and as a result can resemble other non-nodules like vessels and fibrosis, which lung nodule detection can be difficult.

Radhika P R et.al, 2019 [14]	The SVM learning approach was employed by the author to examine information for the purpose of classification analysis. SVM is more appropriate since it lowers the rate of misclassification for datasets that are not linearly separable.	The dataset in this study are taken from the uci machine learning Repository and data.world.in.	Logistic regression parameters are estimated by maximize the logarithmic probability function using training dataset.	The data was classified using a high dimension using the SVM algorithm. Its efficiency is therefore the finest. This method can be used to identify lung cancer more accurately.
Debnath battacharya et.al, 2020 [15]	It is critical to accurately segment suspected nodule candidates, as they may represent a vessel or a nodule. Comparing the ensemble classifier to various other machine learning methods, the author discovered that it performed better.	Brats Dataset, Oasis Dataset, Nbtr Dataset are taken from machine learning repository.	with the characteristics of precision, specificity, sensitivity, ROC, and mean deviation taken into account. Following a thorough examination, we concluded that none of the classifiers achieves accuracy levels near to 100%.	The author used variety of machine learning algorithms and compared to find which one achieved the highest accuracy to classify the lung cancer.
Sushmita das et.al, 2020 [16]	The author addressed several benefits and drawbacks of the current algorithms for the identification of lung cancer. They also presented a comparison between traditional CAD systems and approaches based on deep learning methodology.	The datasets from the RIDER, SPIE challenge and LIDC-IDRI are used.	The author analysed the model's efficiency utilising accuracy, specificity, sensitivity, and area as parameters.	Due to the complicated framework of lung's structure and the unpredictability of nodule features, traditional CAD systems are limited in their capacity to accurately detect pulmonary nodules.
Syed saba raoof et.al, 2020 [17]	The author discusses the deep learning methods to enhance the precision of lung cancer prediction and gives an overview of numerous studies and methods of machine learning-based lung cancer prediction.	The Standard CT database, Lung1 & TCIA, Chest X-ray 14 & JSRT, and TCIA are the datasets that were used.	Cross validation, f1 score, precision and recall, area Under the Characteristic Curve are parameters.	This study offers an overview of lung cancer, covering its causes, signs, and death rate as well as the use of methods to cancer diagnosis.
Shigao Huang et.al, 2023 [18]	The author compares deep learning and classic machine learning	A dataset of 10,001 lung cancer patients was	Accuracy rates, ROC curves, and cost-benefit analyses were	The study's conclusions indicate that a 5 layer neural network (DNN) using deep learning

	algorithms for predicting the survival period of patients with lung cancer based on 12 clinical and demographic data.	utilised in the study, and it contained clinical and demographic information about the patients' age.	used to compare the models' performances.	techniques can be helpful in precisely forecasting how long lung cancer patients will live.
Gur Amrit Pal Singh et.al, [19]	The author presented an efficient technique for recognising and categorising CT scans of lung cancer is presented. Seven distinct classifiers are used, along with image methods of processing and extracting features. of all the classifying, the multi-layer perceptron classifier has better accuracy, coming in at 88.55%.	The dataset that the authors used included 15,750 clinical photos relating to lung cancer, both malignant and benign.	The MLP classifier attained better values for few other performances. Parameters that were assessed in the paper, including F1 score, precision, and recall.	Among all the classifiers, the MLP classifier have the better accuracy, at 88.55%.
Preeti Katiyar et.al, 2020 [20]	The study compares different techniques for identifying lung cancer and classification in CT scans, with a special focus on early-stage lung cancer detection.	The dataset was created using a freely accessible CT scans images from cancer image archive.	the model is analysed using a range of measures, including computing time, efficiency, true positive rate, true negative rate, as well as the rate of false-positives.	In CT scan image analysis, several types of techniques, including SVM, DWT, ANN, watershed algorithm, and CNN, are applied to achieve high accuracy and sensitivity. Numerous methods determined by thresholding and texture have demonstrated encouraging outcomes in the detection and classification of lung tumours.

3. Methodology

Lung illness affects individuals of any gender or age and is one of the main causes of death. Recent evidence indicates that it ranks among the most common causes of death. To identify the illness early and potentially save lives, it is imperative to diagnose it as soon as possible. Although independent algorithms for deep neural learning and machine learning architectures are available, to use methods of ensemble learning in this study to deal with the given challenge. This choice draws from the understanding that ensemble learning provides unique benefits over single models, including increased adaptability and efficacy in prediction tasks.

Our goal in implementing ensemble learning is to improve the ability to predict our models and generate more dependable and accurate outcomes in the field of study. Scientific data and earlier research demonstrating the effectiveness of ensemble learning across a range of tasks and domains improve this decision. We believe that by using this strategy, we will be able to make forecasts that are more

reliable and precise, which will help the health sector advance.

A. Experimental Setup:

The computer has 8 GB of RAM and runs the operating system Microsoft Windows 10. Jupyter Notebook v 6.3.0 was utilised as a developing environment for this research. Several crucial packages were used to carry out the techniques, including Pandas, Numpy, Matplotlib, and Scikit-Learn. Pandas is employed for efficient data preparation, whilst Scikit-Learn is applied to FE scaling, and classification. The efficiency of the model is measured using multiple metrics such as the confusion matrix, f1-score, accuracy, precision, and recall. To develop simple and relevant representations, Matplotlib is utilised, while the Standard Scaler was critical in data scaling methods.

B. Dataset Collection:

We used datasets usually include related clinical data as well as medical images, such as CT scans, X-rays, and other imaging data. The following frequently used datasets are IDRI-LIDC includes CT scans annotated for NIH and lung nodules. Images from chest X-rays that are connected to different lung ailments make up the Chest X-ray Dataset. The number, variety of lung diseases represented, annotation quality, and condition-specific focus of these databases differ. These datasets are used by researchers to train machine learning models for the identification of lung diseases. This allows for the development of algorithms that can recognise different lung pathologies and support early detection and diagnosis. The datasets contain various types of the diseases data which contains various type of diseases distribution diagram is shown in Fig.1.

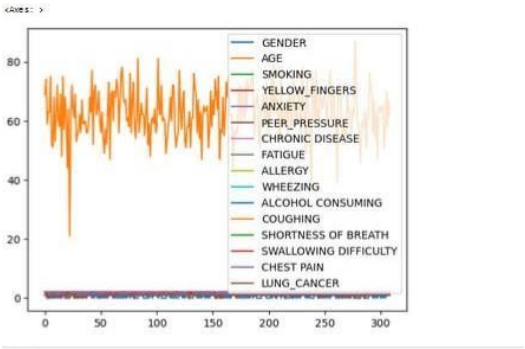

Fig.1- shows distribution of dataset.

C. Dataset preprocessing:

Preprocessing is an important stage in any kind of machine learning or deep learning assignment, particularly when working with medical datasets such as lung cancer risk prediction. It entails preparing and cleaning data before feeding it into a training model for preprocessing a set of data. Managing values that are missing is a necessary step for machine learning to be efficient. Handling outliers is an important process of preprocessing that provides precise and reliable results. If any outliers are detected in this research, they will be removed to protect the data's integrity. Resampling is the procedure for modifying the distribution of classes of a dataset to obtain a balanced set of data.

D. Data Splitting :

To construct a model, it is crucial to divide the data into sets for testing and training. Insufficient data splitting could lead to biassed outcomes due to issues with underestimating or overfitting the model. For the model, it must be able to generalise successfully to new data, which means that partitioning the set of data properly is critical for fine-tuning parameters and accurately assessing the model's performance. To ensure the final model functions properly, the test and the training sets of data are compared. To guarantee a precise evaluation of our models, we followed a standard process to divide the information into test and train datasets, 20% of the information is for testing, while the remaining 80% is set aside for training.

E. Classification :

classification is the procedure of grouping individuals or cases based on the probability or existence of lung tumour. Deep learning techniques are used in the classification process to divide the information into classes. The photos that are trained by disease classes are classified by it. The acquired features are fed into the classification model, and a dataset with labels is needed to convert the information into distinct groups.multi-class and binary classification algorithms forecast a patient's class or stage.

F. Machine Learning Techniques:

1. Logistic regression:

Logistic regression is a core method for developing models that predict and can be used to compare the efficiency of more complicated algorithms used in deep learning as well as machine learning techniques. The model is used binary classification and estimates the possibility of a person belongs to the positive group according to characteristics provided. The program provides prediction as specified by a set threshold of often 0.5 identified as having lung tumour. logistic regression model's efficiency is evaluated using measures including as accuracy, specificity, sensitivity, and ROC curve. Depending on a variety of parameters, logistic regression models can assist predict the likelihood of acquiring lung cancer.

2. Support Vector Machine (SVM):

Support Vector Machines (SVM) is an advanced algorithm based on supervised learning that is employed in classifying issues such as lung cancer diagnosis. SVM is a categorization method which can be utilised for binary and multi-class classification. The algorithm finds the hyperplane that optimizes the space between categories while reducing classification mistakes. The margin-maximizing characteristic aids SVM generalization and avoids overfitting. SVMs are well-known for their capacity to manage complicated decision boundaries and perform well in settings with non-linear connections among variables and outcomes. The algorithm can determine whether a patient is at risk of developing tumours in their lungs or not by calculating where the patient's vector of features lies on the decision boundary. Measures that include sensitivity, specificity, and precision are used to evaluate the model's performance in lung cancer diagnosis.

3. Random Forest Classifier:

Random Forest Classifier is a form of ensemble learning method that includes a number of selection trees to improve prediction accuracy. It creates a set of decision trees, each of which is trained on a random portion of the data utilised for training and a random subset feature of the input. A model of prediction capable of reliably categorizing individuals into two groups: those who are likely to be classified with lung tumour (positive class) and those who are unlikely to develop lung cancer (negative class).It works well for regression and classification problems. Each tree contributes for a class in binary classification in lung cancer detection. The Random Forest technique trains numerous decision trees on subsets of data and attributes that are chosen at random. It prevents overfitting and increases the generalizability of the model. Like the previous models, we evaluated the Random Forest Classifier using precision, recall, accuracy, F1-score, and confusion matrices.

4. KNN (K Nearest Neighbors):

K Nearest Neighbors is a basic categorization method that groups data points according to how near they are to other data points. In order to function, K-Nearest Neighbors stores the training datasets and categorizes new data points according to how close they are to preexisting data points. The number of closest neighbors taken into account for categorization is represented by the "K" in KNN. The K nearest neighbor class that the latest data point belongs to is determined by KNN. The algorithm indicates that the latest data point is likely to have lung cancer if the vast majority of its K neighbors also have the disease. In contrast to other algorithms, KNN does not require intentional training. Rather, in order to identify the closest neighbors, the algorithm computes the distances among the new and old data points during the prediction stage. Metrics like precision, specificity, sensitivity and ROC (AUC) curve are used to assess the KNN model's performance in detecting lung cancer. These parameters help determine how effectively the framework can accurately categorize people as having or not having lung cancer.

5. Gradient boosting classifier:

Gradient Boosting Classifier is used for classification tasks and an ensemble learning technique also known for its high predictive accuracy. Gradient Boosting creates a group of trees of decisions in a stepwise manner, with each new tree concentrating on fixing the mistakes caused by the older ones. The total prediction error is progressively decreased by training each new tree to estimate the residuals from the prior results. By continuously reducing the loss function and going in the path

that minimizes the loss the most and it updates the model. Each tree's prediction is weighted according to its performance as well as its contribution to the entire model, and predictions are formed by adding the predicted results of all the individual trees in the ensemble. The performance of Gradient Boosting Classifier in the detection of lung tumour is assessed through application of measures such as area behind ROC (AUC) curve, f1-score, accuracy, sensitivity, and recall. Gradient Boosting is a robust algorithm that may effectively categorize patients according to several characteristics or risk factors.

6. Ada Boost:

AdaBoost, which stands for Adaptive Boosting, is a technique for ensemble learning that enhances machine learning model performance, especially in classification applications. Although It is also suitable for multiclass classification, its original purpose was binary classification. AdaBoost's main concept is to build a strong learner by combining the predictions of several weak learners, which are typically straightforward and marginally more accurate than guesswork. The Ada Boost gives the high accuracy than other models.

7. XG Boost:

XGBoost, short for eXtreme Gradient Boosting, is a powerful and efficient machine learning algorithm that belongs to the family of gradient boosting methods. It has gained widespread popularity and has been employed to win numerous Kaggle competitions because of its exceptional performance on a variety of tasks.

G. Deep Learning neural networks:

Since deep learning neural networks are capable of extracting complicated patterns and representation from healthcare imaging data, including CT images, a chest X-rays, they demonstrated significant potential in the detection and treatment of lung cancer. Convolutional neural layers, a part of several layers in these networks, are responsible for learning a hierarchical structure of characteristics from picture inputs. They gain the capability to recognize patterns like masses, tumours, or other anomalies that point to lung cancer. deep Lung cancer pictures with truth annotations identifying the presence or the absence of the disease are employed to train neural network frameworks. Transfer learning is used when there is a few labelled health imaging information since it allows pre-trained models to be improved on smaller datasets. The efficiency of the developed neural network model is evaluated by testing and validating it on different datasets. The model's capacity to accurately classify cases of lung cancer is measured using evaluation measures like precision, specificity, sensitivity, recall and f1-score.

4. Results

This section aims to present the outcomes of the conducted experiments and compare them with the findings of relevant prior studies. After extracting the datasets from an open repository, the data is pre-processed and used to evaluate Five standalone ML models and One Deep Learning Models.

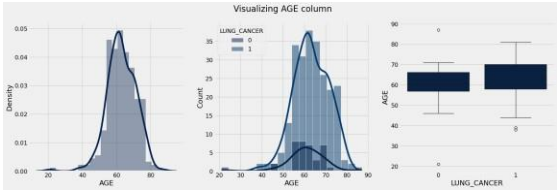

Figure 2: Visualizing AGE column.

In this Figure 2, It Shows Working on visualizing the 'AGE' column in a dataset that presumably contains both continuous and categorical features. To hold the names of the continuous and categorical columns, respectively, we first create two lists, con_col and cat_col.You only include the 'AGE' column in con_col for continuous columns.When adding names to the cat_col list, you iterate over each column in the DataFrame (df) that is categorical, leaving out the 'AGE' column.

Creating Subplots for Visualization:

• We create a 1x3 grid of subplots using 'plt.subplots(1,3,figsize=(20,6))'. This grid will contain three plots arranged horizontally.

Plotting the Distribution of 'AGE':

• In the first subplot (‘ax[0]’),you use ‘sns.distplot(df['AGE'], ax=ax[0])’ to plot the distribution of the 'AGE' column. This provides insights into the overall distribution of ages in the dataset.

Plotting a Histogram with KDE by Lung Cancer Status:

• In the second subplot (‘ax[1]’), you use ‘sns.histplot(data=df, x='AGE', ax=ax[1], hue='LUNG_CANCER', kde=True)’ to create a histogram of ages with kernel density estimates (KDE) displayed. The histogram is differentiated by the 'LUNG_CANCER' status, providing a visual comparison of age distributions between individuals with and without lung cancer.

Creating a Boxplot by Lung Cancer Status:

• In the third subplot (‘ax[2]’), you use ‘sns.boxplot(x=df['LUNG_CANCER'], y=df['AGE'], ax=ax[2])’ to generate a boxplot. This plot allows for a comparison of the central tendency, spread, and potential outliers in age between individuals with and without lung cancer.

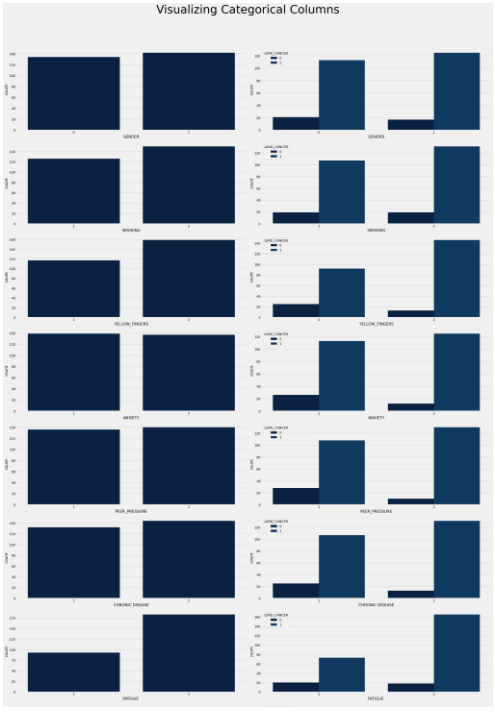

Fig 3: Visualizing Categorical Columns.

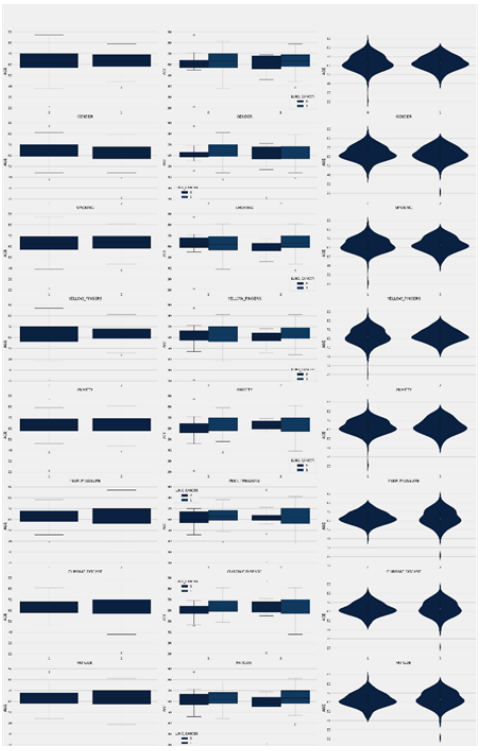

Fig 4: Visualizing AGE vs Categorical Columns

The heap map (Fig :5) that you forwarded is a representation of a correlation map demonstrating how several variables are correlated. From -1 to 1, indicates the degree of relationship between two variables. A perfect negative correlation, or one where a different factor perfectly drops as the other one increases, is shown by a correlation coefficient of -1. A perfect positive correlation, or one in which both variables perfectly increase as one increases, is indicated by a correlation value of 1. There is no correlation, when the correlation coefficient is 0. The greater the association between the two factors in the heap map, the darker the hue. As an illustration, the correlation between the variables "Gender" and "Smoking" is 0.04, which is extremely low and suggests that there is hardly any association at all. The moderately high correlation of 0.43 between the elements "Anxiety" and "Wheezing" displays that these two factors are positively correlated. "Alcohol Consuming" and "Wheezing" have a somewhat negative correlation (-0.27), meaning that there is a bias within the two variables.

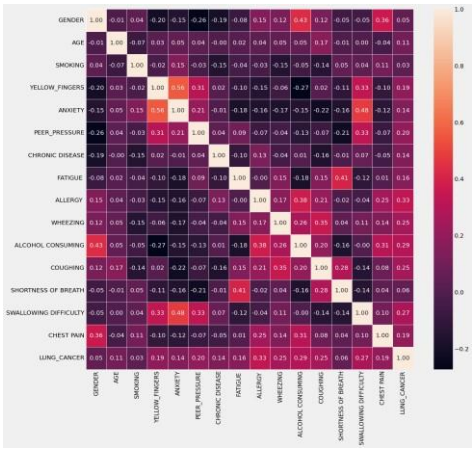

Fig 5: Correlation Heatmap of DataFrameFeatures.

Additional observations from the heap map are as follows:

• "Age" and "Chronic Disease" have a very strong positive association (0.56). The terms "Anxiety" and "Wheezing" have a significant positive association (0.43). This is probably because of the fact that anxiety can make asthmatics wheeze.

• "Alcohol Consuming" and "Wheezing" have a somewhat negative connection (-0.27). This is probably alcohol can aggravate wheezing by irritating the airways. "Coughing" and "Wheezing" have a somewhat favorable association (0.35) due to wheezing and coughing are signs of respiratory issues.

• "Chest Pain" and "Lung Cancer" have a very strong positive connection (0.36) because lung cancer can cause chest pain.

It Shows the (fig:6) implementation of the k-Nearest Neighbors (k-NN) algorithm to find the optimal value of the hyperparameter 'k' through cross-validation. The code begins by importing essential libraries for machine learning, including scikit-learn modules for SVM, KNN, Logistic Regression, Random Forest, and tools for hyperparameter tuning.The main focus is on the k-NN algorithm(fig:6). A loop is implemented to iterate through values of 'k' ranging from 1 to 19. The main focus is on the k-NN algorithm. A loop is implemented to iterate through values of 'k' ranging from 1 to 19. The resulting scores are stored in the 'knn_scores' list. The resulting scores are stored in the knn_scores list. The x-axis represents the values of 'k', and the y-axis represents the mean cross-validation scores. The plot helps visualize how the model's efficiency changes with distinct values of 'k'. The grid lines enhance readability. The code demonstrates how to adjust the k-NN algorithm's(fig:6) hyperparameters via cross-validation. It emphasizes the trade-off between performance, as indicated by the cross-validation scores, and model complexity (regulated by 'k'). Plotting makes it possible to determine the ideal 'k' value that maximizes the accuracy of the model. To make predictions on unseen data, the next stages would usually be to fit the final k-NN model(fig:6) on the whole training set and choose the 'k' value that maximizes performance.

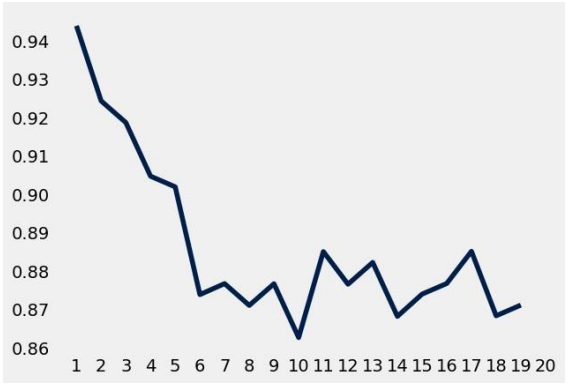

Fig.6: KNN efficiency for unique Values of k.

This Confusion Matrix Describes the It is a table used in machine learning and statistics to assess the performance of a classification model. It summarizes the results of classification by showing the counts of true positive, true negative, false positive, and false negative predictions.

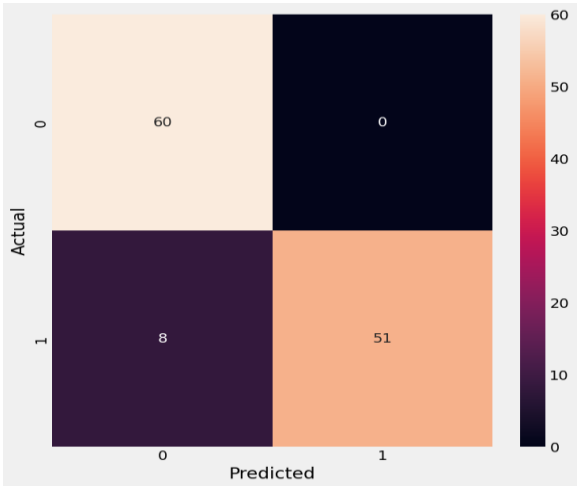

Fig 7: Confusion Matrices of K-Nearest Neighbors Classifier (k=1).

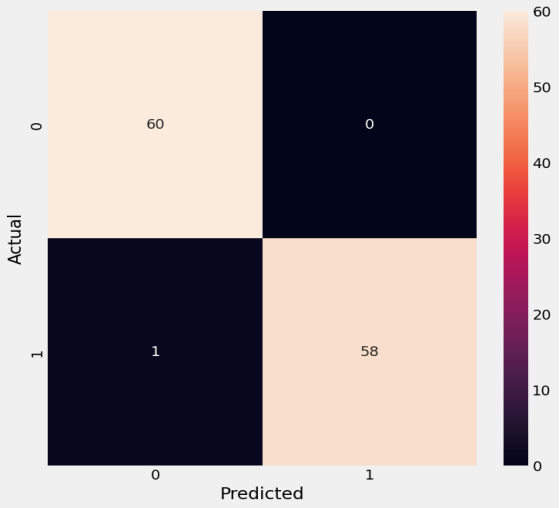

Fig 8: Confusion Matrices of SVM Classifierwith search CV.

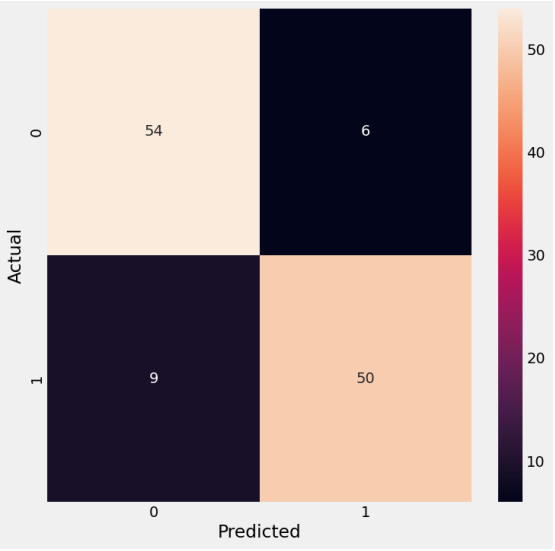

Fig 9: Confusion Matrices of Logistic Regression with search CV.

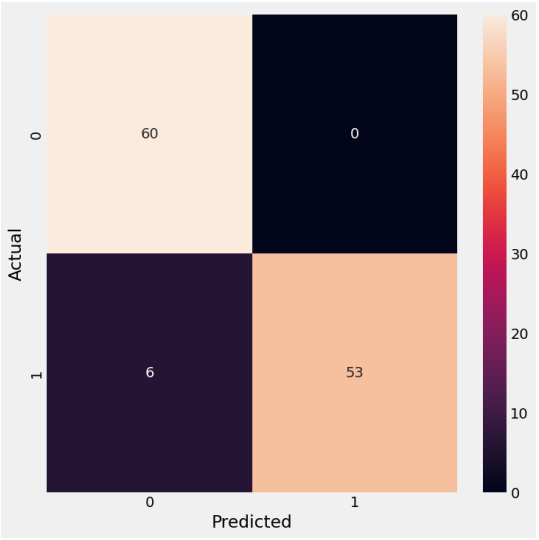

Fig 10: Confusion Matrix of Random ForestClassifier with search CV.

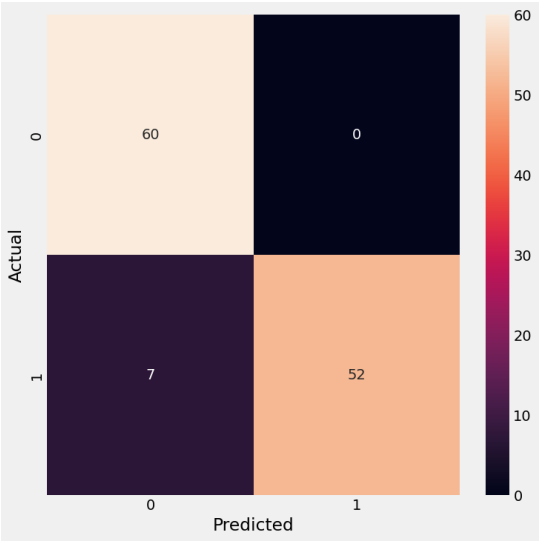

Fig 11: Confusion Matrix of Gradient BoostingClassifier with search CV.

The graph(fig:12) of the ROC and AUC curves is what you submitted. A graph that displays a classification model's performance across all classification criteria is called the ROC curve. Plotting is done with the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. A classification model's total performance is gauged by its AUC. By calculating the ROC curve, it is determined. A model with a greater AUC is said to perform

better. The image(fig:12) you supplied has an extremely high AUC score of 0.9831. This suggests that the model performs an excellent job of differentiating between the two classes.

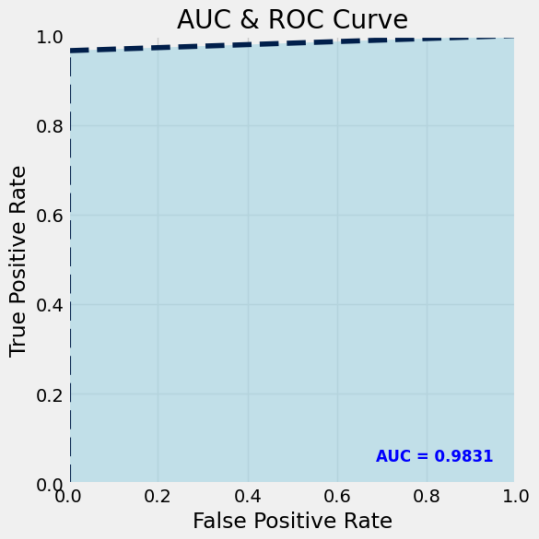

Fig 12: AUC & ROC Curve of SVM Classifier (SVC).

Several types of metrics for performance measurement are used to evaluate how well algorithms detect and categorise lung cancer. These are a few popular assessment measures:

- Precision: The quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

- Recall: It is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive.

$$\text{Recall} = \frac{TP}{TP+FN}$$

- F1-Score: F1 score is a measure of the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

TABLE II displays the categorization task efficiency of various deep learning as well as machine learning models. The table's rows display the various models, while the columns display the models' accuracy, weighted average, support, recall, F1-score, precision, and Marco average score. The models with the best accuracy for this task are the SVM classifier and Random Forest Classifier.

TABLE II: Metrics for measuring the efficiency of machine learning techniques.

Models	Precision	Recall	F1-Score	Support	Marco AVG	Weighted Avg	Accuracy
KNN	0.88	1.00	0.94	119	0.94	0.93	93
SVM	0.99	1.00	0.99	119	0.99	0.99	99
Logistic Regression	0.89	0.90	0.88	119	0.88	0.87	85
Random Forest Classifier	1.00	1.00	0.95	119	0.95	0.95	88
Gradient Boosting Classifier	1.00	0.88	0.94	119	0.95	0.94	88
Ada Boost	0.97	0.99	0.98	114	0.97	0.97	97
XG Boost	0.96	0.97	0.97	114	0.96	0.96	95
Deep Neural Network	1.0	0.77	0.74	112	0.72	0.82	76

Deep Learning Model:

This (Fig:13) histogram showing a collection of individuals' ages. The age distribution is plotted on the x-axis, while the number of individuals in each age group is plotted on the y-axis. There are more persons in their 20s and 30s than in any other age group, according to the histogram. Not many people in their 50s, 60s,

or 70s, although there are a handful in their teens and 40s.The graphic is titled "Age Histogram," and "Density" is written on the y-axis. This shows that rather than displaying the total number of individuals in each age group, the histogram is displaying the density of the age distribution. A distribution's density is a measurement of its degree of dispersion. It appears that the distribution is centered around the middle because there is a high density there and low densities on the sides. The histogram in the picture you supplied me shows low densities on the sides and a high density in the 20s and 30s, which indicates that the group's age distribution.

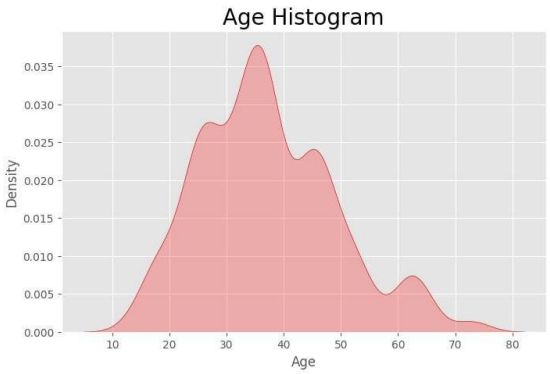

Fig.13: age histogram of neural networks

A line graph displaying the number of levels in a building may be seen in the image you submitted. Level is showed on the X-axis, and count is showed on the Y-axis. Fig (14) "Count of Levels" is the graph's title. According to the graph, there are 303 low-rise buildings compared to 332 medium-rise and 365 high-rise structures. It also explains that there are no significant gaps or spikes in the distribution of values, which is rather uniform. All things considered; the graph offers a succinct summary of the dataset's building level distribution.

Here's a succinct explanation:

- Building level numbers on a line graph.
- Most structures have a low height (303). Less frequently found are high-rise structures (365) and medium-rise buildings(332).The levels are distributed in a quite regular manner.

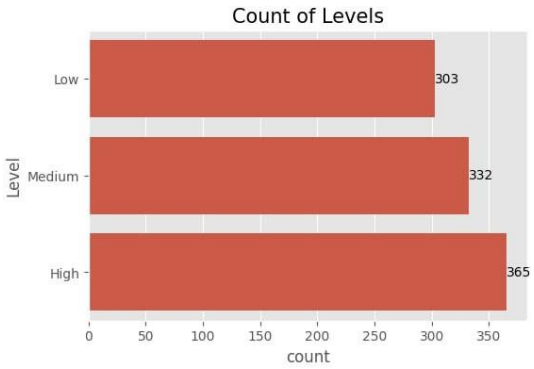

Fig.14: Count of levels of neural networks.

A kernel density estimate (KDE) plot of the Age column in the pandas dataframe df is the picture you supplied. The Age column distribution is displayed as a smooth, continuous curve in the KDE plot. The density of the Age values is shown by the red colouring. The fact that the plot is titled "Age Histogram" suggests that it displays the Age values' distribution as a histogram. The distribution is marginally skewed to the right, as indicated by the Age column's skewness of 0.419. The distribution is marginally more platykurtic than a normal distribution, as indicated by the Age column's kurtosis of 2. 199.The Age column is normally distributed overall, according to the KDE plot, with a small rightward bias.

Here are a few more observations on the KDE plot:

The age range of 20 to 35 represents the majority of the Age values.

A few people stand out, including one who is sixty years old.

There are no noticeable gaps or spikes in the Age values' distribution, which is generally smooth.

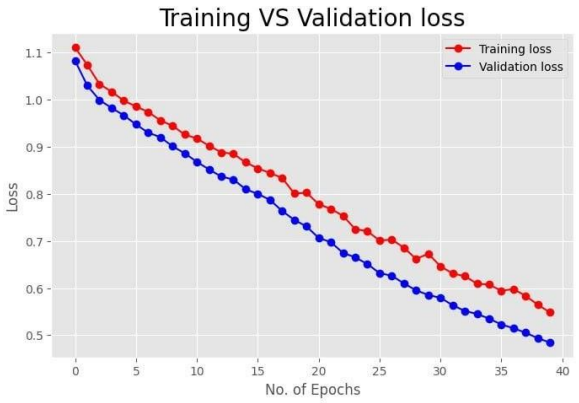

Fig.15: Training VS Validation loss of neural networks.

A machine learning model's training and validation accuracy are displayed on the line graph in the image you submitted. The precision is displayed on the y-axis, while the x-axis displays the number of epochs. The graph is titled "Training VS Validation Accuracy". At every epoch, the graph demonstrates that the training accuracy is greater than the validation accuracy. This suggests that the training data are being overfitted by the model. When a machine learning model learns the training data too well and is not able to generalize to new information, this is known as overfitting. Put another way, instead of learning the underlying patterns in the training data, the model is learning the noise in the data. To deal with overfitting, you can take the following actions including decrease the model's intricacy. This can be achieved by employing regularization techniques, utilizing a smaller model, or lowering the Several variables in the model. Expand the training data set in size. Fig(16) By doing this, the model will be better able to identify the true patterns in the data as opposed to noise. Apply approaches for augmenting data. The current training data must be randomly transformed to produce a new training data. This may facilitate the model's ability to generalize to fresh data. It is crucial to remember that bias and variance are always trade-offs. Unable to recognise the underlying patterns in the data, bias is the inaccuracy that results. Error that arises when a model is overly susceptible to noise is called variance.

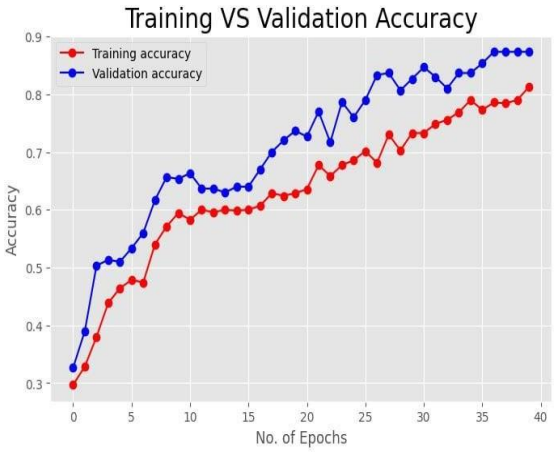

Fig.16: Training vs validation of neuralnetwork.

5. Conclusion

Initial lung disease identification and diagnosis has demonstrated encouraging outcomes for both methods of deep learning and machine learning. Lung cancer and other medical diseases can now be detected early because of the superiority of the algorithms in autonomously obtaining data from medical photos without the need for human intervention. The specifics of the problem, the type of data, and the differences between accuracy, interpretation, computational power, and other aspects all have an impact on identifying lung tumours. The outcomes of five independent machine learning algorithms are examined in this study. Moreover, the application of deep learning and machine learning algorithms to the identification of lung cancer has resulted in enhanced accuracy when compared to current techniques. The models under evaluation demonstrated high accuracy scores, and all the models' Precision, Recall, F1-score, and Macro Avg values ranged from 0.88 to 1.00, indicating consistent good performance. SVM classifier has achieved 99% accuracy. As a result, SVM performed better than other methods of machine learning in lung cancer prediction with respect to accuracy. The method has just been evaluated on one collection of datasets, That's one of its drawbacks.

6. References

[1]. Al-Tarawneh, Mokhled S. "Lung cancer detection using image processing techniques." Leonardo electronic journal of practices and technologies 11.21 (2012): 147-58.

[2]. Bhuvaneswari, C., P. Aruna, and D. Loganathan. "Classification of lung diseases by image processing techniques using computed tomography images." International Journal of Advanced Computer Research 4.1 (2014): 87.

[3]. Bharati, Subrato, Prajoy Podder, and M. Rubaiyat Hossain Mondal. "Hybrid deep learning for detecting lung diseases from X-ray images." Informatics in Medicine Unlocked 20 (2020): 100391.

[4]. Tekade, Ruchita, and K. Rajeswari. "Lung cancer detection and classification using deep learning." 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018.

[5]. Diwan, Meet, Bhargav Patel, and Jaykumar Shah. "Classification of Lungs Diseases Using Machine Learning Technique." International Research Journal of Engineering and Technology (IRJET) 9 (2021).

[6]. Khan, Inam Ullah, et al. "An effective approach to address processing time and computational complexity employing modified CCT for lung disease classification." Intelligent Systems with Applications 16 (2022): 200147.

[7]. Reddy, N. Sudhir, and V. Khanaa. "LDDC-Net: Deep Learning Convolutional Neural Network-based lung disease detection and classification." Journal of Algebraic Statistics 13.1 (2022): 526-542.

[8]. Alshmrani, Goram Mufarah M., et al. "A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images." Alexandria Engineering Journal 64 (2023): 923-935.

[9]. Li, Lanjuan, Haiyang Huang, and Xinyu Jin. "AE-CNN classification of pulmonary tuberculosis based on CT images." 2018 9th international conference on information technology in medicine and education (ITME). IEEE, 2018.

[10]. Liu, Chang, et al. "TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network." 2017 IEEE international conference on image processing (ICIP). IEEE, 2017.

[11]. Nasser, Nidal, et al. "A smart healthcare framework for detection and monitoring of COVID-19 using IoT and cloud computing." Neural Computing and Applications (2021): 1-15.

[12]. Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." Multimedia Tools and Applications 79 (2020): 7731-7762.

[13]. Wang, Lulu. "Deep learning techniques to diagnose lung cancer." Cancers 14.22 (2022): 5569.

[14]. Radhika, P. R., Rakhi AS Nair, and G. Veena. "A comparative study of lung cancer detection using machine learning algorithms." 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT). IEEE, 2019.

[15]. Joshua, Eali Stephen Neal, Midhun Chakkravarthy, and Debnath Bhattacharyya. "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study." Revue d'Intelligence Artificielle 34.3 (2020).

[16]. Das, Susmita, and Swanirbhar Majumder. "Lung cancer detection using deep learning network: A comparative analysis." 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEE, 2020.

[17]. Raoof, Syed Saba, M. A. Jabbar, and Syed Aley Fathima. "Lung Cancer prediction using machine learning: A comprehensive approach." 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020.

[18]. Huang, Shigao, et al. "A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability." Multimedia Tools and Applications 82.22 (2023): 34183-34198.

[19]. Singh, Gur Amrit Pal, and P. K. Gupta. "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans." Neural Computing and Applications 31 (2019): 6863-6877.

[20]. Katiyar, Preeti, and Krishna Singh. "A Comparative study of Lung Cancer Detection and Classification approaches in CT images." 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2020.