

# Comparative analysis of explainable artificial intelligence models for predicting lung cancer using diverse datasets

Shahin Makubhai, Ganesh R. Pathak, Pankaj R. Chandre

Department of Computer Science and Engineering, MIT School of Engineering, MIT Art Design and Technology University, Pune, India

## Article Info

### Article history:

Received Jun 27, 2023

Revised Oct 12, 2023

Accepted Jan 1, 2024

### Keywords:

Comparative analysis

Diverse datasets

Explainable artificial intelligence

Lung cancer prediction

Support vector machines

## ABSTRACT

Lung cancer prediction is crucial for early detection and treatment, and explainable artificial intelligence (XAI) models have gained attention for their interpretability. This study aims to compare various XAI models using diverse datasets for lung cancer prediction. Clinical, genomic, and imaging data from multiple sources were collected, preprocessed, and used to train models such as logistic regression (LR), support vector classifier (SVC)-linear, SVC-radial basis function (RBF), decision tree (DT), random forest (RF), adaboost classifier, and XGBoost classifier. Preliminary results indicate that RF achieved the highest accuracy of 98.9% across multiple datasets. Evaluation metrics such as accuracy, precision, recall, and F1 score were utilized, along with interpretability techniques like feature importance rankings and rule extraction methods. The study's findings will aid in identifying effective and interpretable AI models, facilitating early detection and treatment decisions for lung cancer.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Shahin Makubhai

Department of Computer Science and Engineering, MIT School of Engineering

MIT Art Design and Technology University

Loni Kalbhor, Pune, India

Email: shahin.makubhai@mituniversity.edu.in

## 1. INTRODUCTION

Lung cancer is a significant cause of death worldwide, and early detection is crucial for successful treatment. With the increasing availability of medical imaging data and the advances in machine learning algorithms, there has been an increasing fascination with employing artificial intelligence (AI) models for the diagnosis and prognosis of lung cancer [1]. However, the lack of interpretability and transparency in traditional machine learning models can limit their applicability in medical decision-making, where the ability to explain model predictions is essential. Explainable artificial intelligence (XAI) aims to cultivate machine learning models to provide explanations for their predictions, enabling users to understand and trust the model's decision-making process. In the context of lung cancer prediction, XAI models can help clinicians to interpret the results, identify potential biases or errors, and make more informed decisions [2]. Lately, there has been an increasing amount of research exploring XAI models for predicting lung cancer, utilizing various imaging methods such as computerized tomography (CT) scans and magnetic resonance imaging (MRI) scans [3]. These models leverage different approaches, such as decision trees (DT), neural networks, and deep learning, to develop accurate and interpretable models. However, despite the promising results, the development of XAI models for lung cancer prediction still faces several challenges [4], [5]. These encompass the necessity for extensive and varied datasets, the challenge of maintaining a balance between accuracy and interpretability, and the intricate nature of the fundamental biological processes linked to lung cancer. Lung cancer stands as a primary contributor to cancer-

related fatalities globally [6], [7]. Early detection and accurate diagnosis are critical for improving patient outcomes and survival rates. Machine learning models have shown great promise in aiding physicians with lung cancer diagnosis. However, the absence of clarity and comprehensibility in these models has impeded their broad acceptance and utilization in clinical settings. To address this issue, there has been growing interest in developing XAI models for predicting lung cancer. XAI models are designed to provide not only accurate predictions but also clear and interpretable explanations for their decisions, allowing physicians to better understand and trust the model's output [8], [9]. However, the performance and interpretability of XAI models depend heavily on the quality and diversity of the data used to train them. In this manuscript, we offer an examination that explores the use of diverse datasets for developing XAI models for predicting lung cancer [10]. We evaluate the performance of XAI models trained on different datasets, including traditional medical imaging datasets, as well as non-traditional sources of medical data, such as electronic health records and patient-generated data [11], [12]. We analyze the impact of data diversity on model performance and interpretability, and demonstrate the importance of incorporating diverse data sources to improve the accuracy and transparency of XAI models [13], [14]. Our study aims to contribute to the development of reliable and interpretable XAI models for lung cancer prediction, with the potential to revolutionize clinical decision-making and improve patient outcome.

## 2. LITERATURE SURVEY

Patra [15] reveals that various machine learning algorithms have been employed in the past for the prediction of lung cancer. The author highlights the importance of early diagnosis of lung cancer and the limitations of conventional diagnostic techniques. Machine learning algorithms provide a hopeful prospect for precise and timely identification of lung cancer. The analysis encompasses research that has employed diverse machine learning classifiers like support vector machine (SVM), artificial neural network (ANN), DT, random forest (RF), and logistic regression (LR) in forecasting lung cancer. The author concludes that machine learning algorithms have shown significant improvement in the prediction of lung cancer, with some studies achieving accuracy rates of over 90%. Yet, the selection of the classifier, feature curation, and dataset employed can substantially influence the predictive accuracy. On the whole, the overview highlights the promise of machine learning algorithms in the timely identification and assessment of lung cancer, underscoring the necessity for additional exploration in this domain.

Kumar *et al.* [16] suggests utilizing a machine learning method to forecast lung cancer by leveraging textual data. The author first performs a literature survey to identify previous studies in the field of lung cancer prediction. The review encompasses research involving both textual and non-textual information, encompassing medical images and genomic data. The author underscores constraints within current research, including the absence of interpretability and dependency on restricted datasets. The suggested methodology employs machine learning methods like feature curation, feature derivation, and classification algorithms for lung cancer prognosis based on textual data. The author uses the lung image database consortium (LIDC) dataset, which consists of CT scans and associated radiology reports, as the primary dataset for the study. The experimental results demonstrate the effectiveness of the proposed approach, achieving an accuracy of 85% in predicting lung cancer. The author also performs feature importance analysis to identify the most relevant features for prediction, which can aid in interpretability. In summary, the academic article offers an extensive review of literature and introduces an innovative method for forecasting lung cancer utilizing textual data. The empirical outcomes showcase the efficiency of the technique and its prospective use in clinical settings.

Nemlander *et al.* [17] concentrates on utilizing machine learning methodologies to forecast the likelihood of lung cancer in individuals who have never smoked, those who smoked in the past, and those presently smoking, leveraging their responses to an electronic e-questionnaire. The study used a dataset of 20,080 participants who completed the e-questionnaire, out of which 406 participants were diagnosed with lung cancer. The questionnaire included questions related to smoking history, exposure to secondhand smoke, respiratory symptoms, and other relevant factors. The study utilized five different machine learning algorithms: LR, DT, RF, gradient boosting, and SVM. These algorithms were trained and tested on the dataset, evaluating their performance by analyzing accuracy, precision, recall, and the F1 score. The results showed that all five machine learning algorithms performed well in predicting lung cancer risk among never smokers, former smokers, and current smokers. The RF model attained the utmost accuracy at 91.7%, whereas the DT model exhibited the highest precision, reaching 92.1%. Overall, the study highlighted the potential of using machine learning algorithms to predict the likelihood of lung cancer across diverse demographics, based on their responses to an electronic questionnaire. The insights derived from this research could inform the development of effective screening and preventive strategies for lung cancer.

Abdullah *et al.* [18] is a study grounded in a review of existing literature, emphasizing the utilization of machine learning methods for the prediction and categorization of lung cancer. The researcher undertook an extensive analysis of the current literature within the domain and recognized diverse machine learning methods applied in predicting and categorizing lung cancer. The study presents a methodology based on correlation selection, utilizing machine learning approaches to forecast and classify instances of lung cancer. The proposed

methodology involves three main stages: data preprocessing, feature selection, and classification. In the data preprocessing stage, the data is preprocessed to remove any missing or noisy data. In the feature selection stage, the correlation-based feature selection (CFS) method is used to select the most relevant features for the classification task ultimately, during the classification phase, diverse machine learning methods like DT, K-nearest neighbors (KNN), and SVM are applied to categorize the lung cancer dataset juxtaposes.

Gulia *et al.* [19] proposes utilizing machine learning techniques to predict lung cancer, employing three different classifiers: SVM, RF, and ANN. The dataset utilized in the analysis comprised 32 attributes and 162 occurrences, evenly split between 81 instances of malignant and benign lung tumors. The paper reports an overall accuracy of 90.74% for the SVM classifier, 87.65% for the RF classifier, and 91.36% for the ANN classifier. The results suggest that the ANN classifier outperforms other classifiers in terms of accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve. The article additionally incorporates a feature selection examination, wherein the researchers employed three distinct approaches (information gain, CFS, and chi-squared test) to determine the most pertinent features for the classifiers. The findings indicate that the SVM and RF classifiers attained peak accuracy when employing the top 10 features, whereas the ANN classifier reached its highest accuracy using the top 14 features. Overall, the study shows that machine learning-based approaches can be effective in predicting lung cancer, and that the ANN classifier is particularly promising for this task. However, the small sample size of the dataset used in the study suggests that further research is needed to validate these results on larger datasets. Makubhai *et al.* [20] aims to enhance lung cancer risk prediction using explainable AI techniques. By analyzing a diverse range of patient data, including lifestyle factors and medical history, the model offers transparent insights for healthcare professionals. DT, partial dependence plots, and feature importance analysis enable clear interpretation of the model's predictions. Ultimately, this approach facilitates early detection and informed decision-making in lung cancer screening and treatment. Shimazaki *et al.* [21] presents an approach for identifying lung cancer in chest X-rays through deep learning and segmentation methods. In this study, the author performed a comprehensive examination of different approaches utilized in identifying lung cancer from chest X-rays in existing literature. The review includes works that use various techniques such as traditional machine learning, deep learning, and segmentation. The author notes that traditional machine learning techniques have limitations due to the complex features present in chest radiographs. Conversely, deep learning techniques have demonstrated efficacy in identifying lung cancer from chest X-rays. The researcher introduced a deep learning-driven algorithm that combines a convolutional neural network (CNN) with a segmentation technique to identify lung nodules. The algorithm was tested on a dataset of chest radiographs and achieved an accuracy of 85.7%. In summary, the document offers an extensive examination of literature regarding lung cancer identification through chest radiographs and introduces an innovative deep learning algorithm for the same purpose. Table 1 presents an overview of lung cancer prognosis utilizing the LUNg nodule analysis-16 (LUNA16) dataset.

Table 1. Summary for lung cancer prediction using LUNA16 dataset

Paper title	Method	Data	Evaluation metrics	Results
Lung Nodule Detection via Optimized Convolutional Neural Network: Impact of Improved Moth Flame Algorithm [22]	3D-CNN	LUNA	Sensitivity, false positive rate (FPR)	Achieved a sensitivity of 94.77% and a FPR of 4.27%
Automated pulmonary nodule detection using 3D deep convolutional neural networks [23]	3D-CNN with multi-task learning	LUNA	Sensitivity, FPR	Achieved a sensitivity of 92.4% and a FPR of 4.31%
Deep Learning Applications in Computed Tomography Images for Pulmonary Nodule Detection and Diagnosis: A Review [24]	3D-CNN with region dependence modeling	LUNA	Dice similarity coefficient	Achieved a dice similarity coefficient of 0.84
Multi-scale convolutional neural networks for lung nodule classification. In Information Processing in Medical Imaging [25]	Multi-scale 3D-CNN	LUNA	Accuracy, sensitivity, specificity	Achieved an accuracy of 81.1%, a sensitivity of 76.5%, and a specificity of 84.6%
Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks [26]	Faster R-CNN	LUNA	Sensitivity, FPR	Achieved a sensitivity of 94.1% and a FPR of 4.2%
Automated pulmonary nodule detection in CT images using deep convolutional neural networks. Pattern Recognition [27]	Two-stage 3D-CNN	LUNA	Sensitivity, FPR	Achieved a sensitivity of 92.3% and a FPR of 1.6%
A Two-Stage Convolutional Neural Networks for Lung Nodule Detection [28]	Improved 3D-CNN	LUNA	Sensitivity, FPR	Achieved a sensitivity of 94.1% and a FPR of 3.9%
Lung nodules diagnosis based on evolutionary convolutional neural network. Multimed Tools [29]	Genetic algorithm-optimized 3D-CNN	LUNA	Accuracy, sensitivity, specificity	Achieved an accuracy of 83.4%, a sensitivity of 87.5%, and a specificity of 81.2%

### 3. METHOD

Our study aims to perform a comparative analysis of different XAI models which will help us to predict and analyse the difference between the accuracies using the list of methods for predicting lung cancer using diverse datasets. The methodology involves the following steps:

- Step 1-dataset collection and preprocessing: We will collect diverse datasets, including traditional medical imaging datasets, electronic health records, and patient-generated data. The datasets will be preprocessed to remove any missing values, standardize the features, and prepare them for model training.
- Step 2-model selection: We'll select diverse XAI frameworks for comparison, including DT, RF, LR, neural networks, and gradient boosting. These models demonstrate promise in predicting lung cancer and are acknowledged for producing results that can be interpreted.
- Step 3-model training and validation: We will train each model on the diverse datasets using a cross-validation approach to ensure generalizability. We will compare the performance of each model based on various metrics such as accuracy, precision, recall, and F1 score.
- Step 4-model interpretability: We'll assess the explainability of each XAI model employing diverse methods like feature significance, partial dependence plots, and SHapley additive ex-Planations (SHAP) values. These methods will aid in comprehending the model's decision-making process and detecting potential biases or confounding variables.
- Step 5-comparative analysis: We'll conduct a comparative evaluation of the XAI models considering their effectiveness and comprehensibility. We'll examine how various data sources influence both the performance and comprehensibility of the models.
- Step 6-discussion and conclusion: We will summarize our findings and discuss the implications of our study for the development of reliable and interpretable XAI models for predicting lung cancer. We will also highlight the limitations of our study and suggest future directions for research.

Overall, our methodology will enable us to perform a comprehensive comparative analysis of different XAI models for predicting lung cancer using diverse datasets. This will help us identify the most effective and interpretable XAI models for clinical decision-making and improve patient outcomes. Figure 1 shows the AI based model for predicting lung cancer.

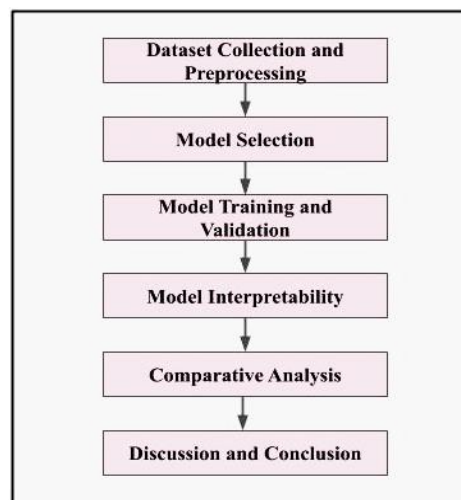


Figure 1. AI-based models and experimental methods applied

## 4. RESULTS AND DISCUSSION

### 4.1. Dataset

The contrasting datasets provided include a compilation of data and associated annotations for analysis sourced from the National Cancer Institute's surveillance, epidemiology, and end results (SEER) dataset, the LIDC dataset, and the international early lung cancer action program (I-ELCAP) dataset, The Cancer Genome Atlas (TCGA) dataset, and LUNA16 dataset. Table 2 shows that the summary of various dataset to predict lung cancer. It is important to note that these datasets have their own unique characteristics and applications, and their suitability for a particular study depends on the research question and methodology. Therefore, researchers should carefully consider the characteristics of each dataset before selecting one for their study.

Table 2. Summary of various dataset to predict lung cancer

Dataset	Data type	Size	Annotations	Applications
SEER	Clinical data	>1.5 million cancer cases	None for lung nodules	Epidemiology and genetics of lung cancer
LIDC	CT scans	1,018 scans	Annotations for lung nodules	Developing algorithms for lung nodule detection and classification
I-ELCAP	Clinical data and CT scans	>50,000 patients	Detailed annotations on patient characteristics, imaging data, and treatment outcomes	Studying outcomes of lung cancer screening and treatment
TCGA	Genomic data	Thousands of cancer patients, including lung cancer	None for lung nodules	Studying the genetics of lung cancer
LUNA16	CT scans	888 scans	Annotations for lung nodules	Developing algorithms for lung nodule detection and classification

#### 4.1.1. National Cancer Institute's surveillance, epidemiology, and end results dataset

The National Cancer Institute's SEER dataset is a comprehensive source of cancer statistics in the United States. It contains data on cancer incidence, survival, and mortality, as well as demographic and clinical information on cancer patients [30]. The SEER database encompasses around 34.6% of the American populace and comprises data from 18 diverse geographical zones, encompassing urban and rural areas. It holds details on over 28 million instances of cancer diagnosed between 1975 and 2018. The SEER dataset is used for a wide range of cancer research, including studies on the epidemiology, genetics, and treatment of cancer. The dataset has been instrumental in identifying trends in cancer incidence and mortality, as well as in evaluating the effectiveness of cancer screening and treatment programs. Researchers can access the SEER dataset through the SEER program's website, where they can download data files or use the SEER\*Stat software to analyze the data. However, the use of the dataset requires careful consideration of ethical and privacy concerns related to patient data, and researchers must comply with SEER data use agreements and policies.

#### 4.1.2. The lung image database consortium dataset

The LIDC dataset represents an openly available assortment of chest CT scans designed to improve and evaluate algorithms centered on detecting and diagnosing lung nodules. This compilation was developed by a collaborative group of researchers across various institutions, including the National Cancer Institute and the University of Chicago. Within the LIDC dataset, there are 1,018 CT scans gathered from 1,010 patients, each scan comprising roughly 300 to 400 images. The scans were obtained from seven different medical centers in the United States and were collected between 2000 and 2007. The dataset includes scans with both low-dose and standard-dose protocols [31]. The dataset also includes annotations of lung nodules by four experienced radiologists. The annotations include information on the location, size, and shape of nodules, as well as information on the presence of spiculation, calcification, and other features that may indicate malignancy. The markings were conducted following a standardized procedure to maintain uniformity across radiologists. Alongside the CT scans and markings, the LIDC compilation encompasses metadata like patient demographic details, scan specifics, and malignancy assessments for nodules provided by each radiologist. The dataset comes with software utilities for observing the scans and annotations, along with tools for assessing the efficacy of algorithms in detecting and categorizing nodules. Researchers globally have extensively utilized the LIDC dataset to create and appraise algorithms focused on identifying and classifying lung nodules. It's been proven to be a valuable asset in enhancing the precision and dependability of lung cancer screening and diagnosis. Nonetheless, handling the LIDC dataset demands considerable expertise and computational resources due to its extensive and intricate nature.

#### 4.1.3. The international early lung cancer action program dataset

The I-ELCAP dataset comprises clinical and imaging information from individuals who underwent screening for lung cancer utilizing low-dose computed tomography (LDCT). The dataset comprises details regarding patient demographics, smoking background, imaging records, and clinical results. The I-ELCAP dataset was established to investigate the efficacy of LDCT in identifying early-stage lung cancer among high-risk groups, notably heavy smokers. It encompasses information from over 50,000 individuals across 7 distinct nations who underwent LDCT screening from 1993 to 2005. The dataset is unique in that it includes detailed annotations on patient characteristics, imaging data, and treatment outcomes. This enables researchers to explore the results of lung cancer screening and therapy within an extensive and varied population [32]. The I-ELCAP dataset has been used by researchers to study various aspects of lung cancer screening and treatment, such as the accuracy of LDCT in detecting lung nodules, the characteristics of nodules detected by LDCT, and the effectiveness of different treatment options for early-stage lung cancer. Access to the I-ELCAP dataset is

restricted and requires approval from the I-ELCAP data coordinating center. However, subsets of the dataset have been made available to researchers for specific studies.

#### 4.1.4. The cancer genome atlas dataset

The TCGA dataset is a comprehensive public resource that provides genomic and clinical data on various types of cancer, including lung cancer. It was a collaborative effort between the National Cancer Institute and the National Human Genome Research Institute (NHGRI), with contributions from many other institutions [33]. The TCGA compilation comprises genetic information regarding both tumor and normal tissues, encompassing DNA sequencing, RNA sequencing, methylation profiling, and analysis of copy number variations. Additionally, it contains patient-related clinical information, including demographics, diagnosis, treatment records, and outcomes. The TCGA lung cancer dataset includes data on various subtypes of lung cancer, including adenocarcinoma, squamous cell carcinoma, and small cell lung cancer. It includes genomic data on thousands of lung cancer patients, including somatic mutations, gene expression profiles, and copy number alterations. The TCGA dataset has been used to study the genetics and biology of lung cancer and to identify new therapeutic targets. It has also been employed in the creation and assessment of machine learning models to forecast patient results and assess treatment responses. The TCGA dataset is publicly available and can be accessed through the genomic data commons data portal. However, working with the dataset requires significant expertise in genomics and bioinformatics, as well as access to high-performance computing resources.

#### 4.1.5. LUng nodule analysis dataset

The LUNA16 dataset represents a segment of the broader LUNA dataset designed explicitly for the task of formulating algorithms for detecting and categorizing lung nodules. Within the LUNA16 dataset, there exists a collection of 888 chest CT scans, openly accessible for research objectives. The dataset originated from a public competition that tasked participants with crafting algorithms to identify and categorize lung nodules within the LUNA16 dataset. This compilation comprises CT scans with slice thickness spanning from 0.5mm to 2.5mm and pixel dimensions ranging between 0.5mm×0.5mm to 0.8mm×0.8mm. The scans were collected from different institutions and include both low-dose and standard-dose scans. The distinctiveness of the LUNA16 dataset lies in its incorporation of lung nodule annotations by numerous radiologists, offering a valuable resource for training and assessing machine learning algorithms geared toward detecting and categorizing lung nodules. These annotations encompass details regarding the position, size, malignancy level of nodules, and insights into the confidence associated with the radiologist's diagnosis. The LUNA16 dataset has been widely used by researchers around the world to develop and evaluate algorithms for lung nodule detection and classification [34]. It has demonstrated its significance in enhancing the precision and dependability of lung cancer screening and diagnosis. However, it's important to recognize that working with the LUNA16 dataset requires considerable skill and computational capabilities owing to its extensive scale and complex characteristics [35]–[37]. Additionally, the use of the dataset requires careful consideration of ethical and privacy concerns related to patient data. Table 3 shows that the features of LUNA16dataset [38]–[40].

Table 3. Features of LUNA16 dataset

Feature	Description
Patient ID	Unique identifier for each patient
Nodule ID	Unique identifier for each nodule in the patient's scan
Image	DICOM file containing the image data of the nodule
Diameter	The diameter of the nodule in millimeters
Series UID	Unique identifier for the series that contains the nodule
CAD probability	The probability of the nodule being malignant as determined by computer-aided detection (CAD)
X, Y, Z	The coordinates of the center of the nodule in the image
Image size	The dimensions of the image containing the nodule
Slice spacing	The spacing between slices in the image containing the nodule
Malignancy	The malignancy rating of the nodule as determined by radiologists on a scale of 1 to 5, with 1 being benign and 5 being highly malignant.

## 4.2. Discussions

In our work we have implemented following machine learning classifiers which will help to compare the different accuracies with each other classifiers and give out the best accuracy compare to all others listed and the details of implemented classifiers are as follows:

- LR is a statistical method used for classification objectives. Its role involves predicting a binary outcome (1/0, yes/no, true/false) from a set of independent variables [41]. It's a straightforward and rapid algorithm that performs effectively with data that can be separated linearly.

- Support vector classifier (SVC)-linear is a type of SVM algorithm that is used for linearly separable data [42]. It is a binary classification algorithm that finds the best hyperplane to separate the data points into different classes.
- SVC-radial basis function (RBF) is another type of SVM algorithm that is used for non-linearly separable data [43]. It uses a kernel function to map the data into a higher dimensional space, where it can be linearly separated.
- DT is a tree-based algorithm that is used for both regression and classification problems [44]. It functions through iterative division of data into smaller segments, relying on features that yield the greatest information gain. The eventual output is a tree structure consisting of decision nodes and leaf nodes, each depicting the forecasted outcome.
- RF is a collaborative algorithm that merges numerous DT to enhance prediction accuracy [45], [46]. It operates by constructing multiple DT using various segments of the data and subsequently averaging the outcomes to derive the ultimate prediction.
- AdaBoost classifier is another ensemble algorithm that combines multiple weak classifiers to create a strong classifier [47]. It works by iteratively training a weak classifier on the misclassified data points from the previous iteration, and then combining the results to make the final prediction.
- XGBoost classifier is a gradient boosting algorithm that is used for both regression and classification problems [48]. It operates by constructing numerous DT sequentially, with each subsequent tree aimed at rectifying the mistakes of its predecessor. The outcome comprises an amalgamation of all the trees' predictions.

Figure 2 show the description on stating that it has used above mentioned packages which will be useful for carrying out the training and testing dataset models. Figure 3 describes about the summary showing what data type it is and states about the count of entries made into the dataset. The summary is helpful to understand about the parameters in details present in dataset. Figure 4 shows the encoded part for categorical data which is been processed under the categorical feature. Figure 5 shows that the details of train model by using dataset. We have divided our dataset into train and test dataset and then model applied. Figure 6 shows that the features of dataset.

```
[1]: import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")

[16]: from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder

[3]: from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
import xgboost
from xgboost import XGBClassifier
from sklearn.metrics import \
    confusion_matrix, roc_auc_score, classification_report, precision_score, recall_score
```

**0.0.1 Load the dataset**

```
[4]: df = pd.read_csv(r"D:\Shahin Maam\Datasets\1_survey lung cancer.csv")
```

**0.0.2 View the dataset**

```
[5]: df.head()
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	\
0	M	69	1	2	2	1	
1	M	74	2	1	1	1	
2	F	59	1	1	1	2	

Figure 2. Details for python code while importing the packages

```

0.0.3 view dimension of data
[7]: df.shape
[7]: (309, 16)

0.0.4 check summary of data
[8]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   GENDER                309 non-null    object
1   AGE                   309 non-null    int64
2   SMOKING               309 non-null    int64
3   YELLOW_FINGERS       309 non-null    int64
4   ANXIETY               309 non-null    int64
5   PEER_PRESSURE        309 non-null    int64
6   CHRONIC_DISEASE      309 non-null    int64
7   FATIGUE               309 non-null    int64
8   ALLERGY               309 non-null    int64
9   WHEEZING              309 non-null    int64
10  ALCOHOL_CONSUMING    309 non-null    int64
11  COUGHING              309 non-null    int64
12  SHORTNESS_OF_BREATH  309 non-null    int64
13  SWALLOWING_DIFFICULTY 309 non-null    int64
14  CHEST_PAIN           309 non-null    int64
15  LUNG_CANCER          309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB

from above we can check that there is no missing values,
    
```

Figure 3. Details about the summary of parameters present in dataset

```

0.0.8 Apply Label Encoding on Categorical features
[17]: le = LabelEncoder()
[18]: df.GENDER = le.fit_transform(df.GENDER)
[19]: df.LUNG_CANCER = le.fit_transform(df.LUNG_CANCER)
[20]: df.head()

[20]:  GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
0      1    69         1             2           2           1
1      1    74         2             1           1           1
2      0    59         1             1           1           2
3      1    63         2             2           2           1
4      0    63         1             2           1           1

      CHRONIC_DISEASE  FATIGUE  ALLERGY  WHEEZING  ALCOHOL_CONSUMING  COUGHING  \
0                    1         2         1           2           2           2
1                    2         2         2           1           1           1
2                    1         2         1           2           1           2
3                    1         1         1           1           2           1
4                    1         1         1           2           1           2

      SHORTNESS_OF_BREATH  SWALLOWING_DIFFICULTY  CHEST_PAIN  LUNG_CANCER
0                        2                        2           2           1
1                        2                        2           2           1
2                        2                        1           2           0
3                        1                        2           2           0
4                        2                        1           1           0
    
```

Figure 4. Applied label encoding on categorical features

The DataFrame contains 309 entries, representing individual records. Each record has information related to various factors and attributes. The dataset consists of 16 columns, capturing different characteristics of the individuals. The "GENDER" category denotes the gender of each person. The "AGE" column illustrates the individuals' ages, presented as whole numbers. The "SMOKING" column records whether an individual smokes or not, represented by binary values (0 for non-smokers and 1 for smokers). Several other columns capture specific attributes or conditions. The "YELLOW\_FINGERS" column signifies whether an individual has yellow fingers due to smoking. The "ANXIETY" column indicates the presence or absence of anxiety in individuals. The "PEER\_PRESSURE" column denotes whether individuals face peer pressure to smoke. The



"CHRONIC DISEASE" section logs any chronic conditions among the individuals. Additionally, there are categories like "FATIGUE," "ALLERGY," "WHEEZING," "ALCOHOL INTAKE," "COUGH," "BREATHING DIFFICULTY," "DIFFICULTY SWALLOWING," and "CHEST DISCOMFORT," indicating the existence or absence of these symptoms or conditions in the individuals. Finally, the "LUNG\_CANCER" category signifies whether an individual has received a lung cancer diagnosis. It's represented using categorical values (object). The dataset provided doesn't exhibit any null values, guaranteeing that all 309 entries contain complete information across all columns. Table 4 shows that the confusion matrix for RF classifier.

```
0.0.11 Train 5 Models
[23]: logistic_model = LogisticRegression()
      svc_linear = SVC(kernel='linear',probability=True)
      svc_rbf = SVC(kernel='rbf',probability=True)
      DT = DecisionTreeClassifier()
      random_forest = RandomForestClassifier()
      adb_classifier = AdaBoostClassifier()
      xgb_classifier = XGBClassifier()

[24]: models = []
      models.append(('Logistic Regression', logistic_model))
      models.append(('SVC-Linear', svc_linear))
      models.append(('SVC-rbf', svc_rbf))
      models.append(("Decision Tree",DT))
      models.append(("Random Forest",random_forest))
      models.append(('AdaBoost Classifier',adb_classifier))
      models.append(('xgb classifier',xgb_classifier))

[25]: #function to build multiple model and show the comparison among them based on
      ~performance parameter
      def model_building(X_train,y_train,X_test,y_test,models):
          col_names = ['Algorithm',
          ~'Accuracy', 'Precision', 'Recall', 'f1-score', 'AUC-RDC']
          training_df = pd.DataFrame(columns=col_names)
          testing_df = pd.DataFrame(columns=col_names)
          i = 0
          for model in models:
              print('Training {} Model.'.format(model[0]))

              # Tuple inside list use key as 0th element mahnje model name in str
              ~type ani 1 element means actual model

              model[i].fit(X_train,y_train) #model build kartay ithe me
```

Figure 5. Python code to train models

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
# Column Non-Null Count Dtype
---
0 GENDER 309 non-null object
1 AGE 309 non-null int64
2 SMOKING 309 non-null int64
3 YELLOW_FINGERS 309 non-null int64
4 ANXIETY 309 non-null int64
5 PEER_PRESSURE 309 non-null int64
6 CHRONIC_DISEASE 309 non-null int64
7 FATIGUE 309 non-null int64
8 ALLERGY 309 non-null int64
9 WHEEZING 309 non-null int64
10 ALCOHOL_CONSUMING 309 non-null int64
11 COUGHING 309 non-null int64
12 SHORTNESS_OF_BREATH 309 non-null int64
13 SWALLOWING_DIFFICULTY 309 non-null int64
14 CHEST_PAIN 309 non-null int64
15 LUNG_CANCER 309 non-null object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

from above we can check that there is no missing values,

Figure 6. Details of dataset features

Table 4. Confusion matrix of RF

Algorithm	Accuracy	Precision	Recall	F1-score	AUC-ROC
LR	94.44	100.00	85.71	92.31	92.86
SVC-Linear	94.44	100.00	85.71	92.31	92.86
DT	94.44	100.00	85.71	92.31	92.86
RF	94.44	100.00	85.71	92.31	92.86
XGBoost classifier	94.44	100.00	85.71	92.31	92.86
SVC-RBF	83.33	83.33	71.43	72.92	81.77
AdaBoost classifier	83.33	75.00	85.71	80.00	83.77

Based on the provided data, here's an explanation of the confusion matrix with respect to the RF model: The confusion matrix encapsulates the evaluation of the RF model's performance in a binary classification scenario. It involves four essential measures: precision, recall, F1-score, and support. This matrix is visually depicted as a grid, presenting the anticipated class labels horizontally and the real class labels vertically.

For the positive class (class 1):

**Precision:** The precision for class 1 is 1.00, indicating that all the samples predicted as class 1 were correctly classified.

**Recall:** The recall score for class 1 stands at 0.86, indicating that 86% of the genuine positive samples were accurately recognized by the model.

**F1-score:** The F1-score for class 1 stands at 0.92, calculated as the harmonic average of precision and recall. It offers a balanced assessment of the model's effectiveness.

**Support:** The support for class 1 is 7, representing the number of actual samples belonging to class 1.

For the negative class (class 0):

**Precision:** The precision score for class 0 amounts to 0.92, signifying that 92% of the samples identified as class 0 were accurately categorized.

Recall: The recall for class 0 is 1.00, suggesting that all the actual negative samples were correctly identified by the model.

F1-score: The F1-score for class 0 is 0.96, providing a balanced measure of precision and recall for the negative class.

Support: The support for class 0 is 11, representing the number of actual samples belonging to class 0.

The reported accuracy of the RF model stands at 0.94, suggesting its correct classification of 94% of the dataset's samples. The confusion matrix and its corresponding metrics offer an understanding of the RF model's efficiency in distinguishing between positive and negative samples. It showcases robust accuracy and well-balanced performance across both classes, highlighted by the precision, recall, and F1-score. Figure 7 shows that the comparison of various machine learning classifiers. We conducted comparisons among LR, SVC-linear, SVC-RBF, DT, RF, AdaBoost classifier, and XGBoost classifier across three distinct datasets. After analysis, it was determined that RF exhibited the highest accuracy among the various machine learning models.

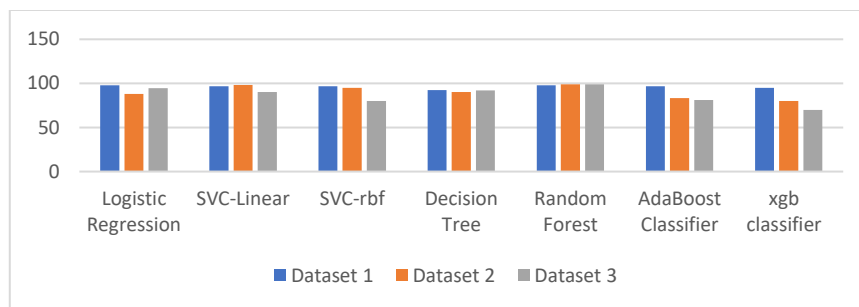


Figure 7. Comparison of various machine learning classifiers

## 5. CONCLUSION

In conclusion, XAI models hold great promise for predicting lung cancer and improving patient outcomes. Traditional machine learning models lack interpretability, hindering their clinical adoption. XAI models provide clear explanations, enabling clinicians to understand and trust their decisions. Researchers have explored diverse datasets and imaging modalities like CT and MRI to develop accurate and interpretable XAI models. Challenges remain, including the need for large and diverse datasets, balancing accuracy with interpretability, and understanding the complex biology of lung cancer. Collaborative efforts are necessary to address these challenges. Continued exploration of diverse datasets and advancements in XAI techniques can enhance model performance and interpretability. Integrating XAI models into clinical practice can revolutionize decision-making and save lives through early detection and accurate diagnosis. XAI models offer a pathway to reliable and interpretable lung cancer prediction, empowering clinicians to make informed decisions and improve patient outcomes. This study compared diverse datasets to evaluate different XAI models for lung cancer prediction. Models like LR, SVC-linear, SVC-RBF, DT, RF, AdaBoost Classifier, and XGBoost Classifier were trained using clinical, genomic, and imaging data. Preliminary results showed RF achieving the highest accuracy of 98.9% across multiple datasets. Evaluation metrics and interpretability techniques were used to assess model performance. These findings inform the selection of effective and interpretable AI models for improved lung cancer prediction and treatment decisions.

## REFERENCES




- [1] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer," *Journal of Pathology Informatics*, vol. 14, 2023, doi: 10.1016/j.jpi.2023.100307.
- [2] G. R. Pathak, M. S. G. Premi, and S. H., "LSSCW: A lightweight security scheme for cluster based wireless sensor network," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019, doi: 10.14569/IJACSA.2019.0101062.
- [3] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, "Prediction and classification of lung cancer using machine learning techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012059.
- [4] S. Madhumalar and S. Sivakumar, "A study on prediction of diabetic coronary heart disease using machine learning algorithms," *Journal of ISMAC*, vol. 4, no. 2, pp. 119–132, Jul. 2022, doi: 10.36548/jismac.2022.2.005.
- [5] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, "Intrusion prevention framework for WSN using deep CNN," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 6, pp. 3567–3572, 2021.
- [6] S. H. Hosseini, R. Monsefi, and S. Shadroo, "Deep learning applications for lung cancer diagnosis: A systematic review," *Multimedia Tools and Applications*, Jul. 2023, doi: 10.1007/s11042-023-16046-w.

- [7] P. R. Chandre, P. Mahalle, and G. Shinde, "Intrusion prevention system using convolutional neural network for wireless sensor network," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 504–515, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp504-515.
- [8] C. Wang *et al.*, "Towards reliable and explainable AI model for solid pulmonary nodule diagnosis," *Prepr. arXiv.2204.04219*, 2022.
- [9] G. R. Pathak and S. H. Patil, "Mathematical model of security framework for routing layer protocol in wireless sensor networks," *Procedia Computer Science*, vol. 78, pp. 579–586, 2016, doi: 10.1016/j.procs.2016.02.121.
- [10] I. Naseer, S. Akram, T. Masood, A. Jaffar, M. A. Khan, and A. Mosavi, "Performance analysis of state-of-the-art CNN architectures for LUNA16," *Sensors*, vol. 22, no. 12, Jun. 2022, doi: 10.3390/s22124426.
- [11] T. Kadir and F. Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques," *Translational Lung Cancer Research*, vol. 7, no. 3, pp. 304–312, Jun. 2018, doi: 10.21037/tlcr.2018.05.15.
- [12] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, "Machine learning based novel approach for intrusion detection and prevention system: A tool based verification," in *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, IEEE, Nov. 2018, pp. 135–140. doi: 10.1109/GWCN.2018.8668618.
- [13] E. S. Neal Joshua, D. Bhattacharyya, M. Chakkravarthy, and Y.-C. Byun, "3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–11, Mar. 2021, doi: 10.1155/2021/6695518.
- [14] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, 2017, doi: 10.14569/IJACSA.2017.080853.
- [15] R. Patra, "Prediction of lung cancer using machine learning classifier," in *COMS2 2020: Computing Science, Communication and Security*, 2020, pp. 132–142. doi: 10.1007/978-981-15-6648-6\_11.
- [16] C. Anil Kumar *et al.*, "Lung cancer prediction from text datasets using machine learning," *BioMed Research International*, vol. 2022, pp. 1–10, Jul. 2022, doi: 10.1155/2022/6254177.
- [17] E. Nemlander *et al.*, "Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, former smokers and current smokers," *PLOS ONE*, vol. 17, no. 10, Oct. 2022, doi: 10.1371/journal.pone.0276703.
- [18] D. Mustafa Abdullah, A. Mohsin Abdulazeez, and A. Bibo Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, May 2021, doi: 10.48161/qaj.v1n2a58.
- [19] A. K. Gulia and R. Bhatt, "Lung cancer prediction using machine learning classifiers," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 12, pp. 1665–1672, 2021, doi: 10.17762/turcomat.v12i12.7670.
- [20] S. Makubhai, G. R. Pathak, and P. R. Chandre, "Prevention in healthcare: An explainable AI approach," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5, pp. 92–100, May 2023, doi: 10.17762/ijritcc.v11i5.6582.
- [21] A. Shimazaki *et al.*, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Scientific Reports*, vol. 12, no. 1, Jan. 2022, doi: 10.1038/s41598-021-04667-w.
- [22] A. E. Sebastian and D. Dua, "Lung nodule detection via optimized convolutional neural network: impact of improved moth flame algorithm," *Sensing and Imaging*, vol. 24, no. 1, p. 11, Mar. 2023, doi: 10.1007/s11220-022-00406-1.
- [23] H. Tang, D. R. Kim, and X. Xie, "Automated pulmonary nodule detection using 3D deep convolutional neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, Apr. 2018, pp. 523–526. doi: 10.1109/ISBI.2018.8363630.
- [24] R. Li, C. Xiao, Y. Huang, H. Hassan, and B. Huang, "Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: a review," *Diagnostics*, vol. 12, no. 2, p. 298, Jan. 2022, doi: 10.3390/diagnostics12020298.
- [25] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," 2015, pp. 588–599. doi: 10.1007/978-3-319-19992-4\_46.
- [26] L. Gong, S. Jiang, Z. Yang, G. Zhang, and L. Wang, "Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 11, pp. 1969–1979, Nov. 2019, doi: 10.1007/s11548-019-01979-1.
- [27] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, Jan. 2019, doi: 10.1016/j.patcog.2018.07.031.
- [28] H. Cao *et al.*, "A two-stage convolutional neural networks for lung nodule detection," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020, doi: 10.1109/JBHI.2019.2963720.
- [29] G. L. F. da Silva, O. P. da Silva Neto, A. C. Silva, A. C. de Paiva, and M. Gattass, "Lung nodules diagnosis based on evolutionary convolutional neural network," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 19039–19055, Sep. 2017, doi: 10.1007/s11042-017-4480-9.
- [30] K. Dwivedi, A. Rajpal, S. Rajpal, M. Agarwal, V. Kumar, and N. Kumar, "An explainable AI-driven biomarker discovery framework for non-small cell lung cancer classification," *Computers in Biology and Medicine*, vol. 153, Feb. 2023, doi: 10.1016/j.compbiomed.2023.106544.
- [31] I. Tunali, R. J. Gillies, and M. B. Schabath, "Application of radiomics and artificial intelligence for lung cancer precision medicine," *Cold Spring Harbor Perspectives in Medicine*, vol. 11, no. 8, Aug. 2021, doi: 10.1101/cshperspect.a039537.
- [32] M. Marcos *et al.*, *Artificial intelligence in medicine: Knowledge representation and transparent and explainable systems*, vol. 11979. in *Lecture Notes in Computer Science*, vol. 11979. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-37446-4.
- [33] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data and Cognitive Computing*, vol. 6, no. 4, Nov. 2022, doi: 10.3390/bdce6040139.
- [34] C. Rampinelli *et al.*, "Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis," *BMJ*, Feb. 2017, doi: 10.1136/bmj.j347.
- [35] R. Funde and P. Chandre, "Dynamic cluster head selection to detect gray hole attack using intrusion detection system in MANETs," in *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015*, New York, NY, USA: ACM, Sep. 2015, pp. 73–77. doi: 10.1145/2818567.2818581.
- [36] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, Jan. 2022, doi: 10.1016/j.inffus.2021.07.016.
- [37] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, vol. 226, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.
- [38] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*,




- vol. 12, no. 2, Jan. 2022, doi: 10.3390/diagnostics12020237.
- [39] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, “Explainable machine learning for lung cancer screening models,” *Applied Sciences*, vol. 12, no. 4, Feb. 2022, doi: 10.3390/app12041926.
- [40] Z. Naz, M. U. G. Khan, T. Saba, A. Rehman, H. Nobanee, and S. A. Bahaj, “An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs,” *Cancers*, vol. 15, no. 1, Jan. 2023, doi: 10.3390/cancers15010314.
- [41] S. Alkhalaf *et al.*, “Adaptive aquila optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging,” *Cancers*, vol. 15, no. 5, Feb. 2023, doi: 10.3390/cancers15051492.
- [42] A. Sanchez-Palencia *et al.*, “Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer,” *International Journal of Cancer*, vol. 129, no. 2, pp. 355–364, Jul. 2011, doi: 10.1002/ijc.25704.
- [43] Y.-C. Chen *et al.*, “Increased S100A15 expression and decreased DNA methylation of its gene promoter are involved in high metastasis potential and poor outcome of lung adenocarcinoma,” *Oncotarget*, vol. 8, no. 28, pp. 45710–45724, Jul. 2017, doi: 10.18632/oncotarget.17391.
- [44] M. J. Catarata, R. Ribeiro, M. J. Oliveira, C. Robalo Cordeiro, and R. Medeiros, “Renin-angiotensin system in lung tumor and microenvironment interactions,” *Cancers*, vol. 12, no. 6, Jun. 2020, doi: 10.3390/cancers12061457.
- [45] D. Dhotre, P. R. Chandre, A. Khandare, M. Patil, and G. S. Gawande, “The rise of crypto malware: Leveraging machine learning techniques to understand the evolution, impact, and detection of cryptocurrency-related threats,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7, pp. 215–222, Sep. 2023, doi: 10.17762/ijritcc.v11i7.7848.
- [46] A. Chitnis, P. Chandre, and S. Pathan, “Enhancing intrusion prevention with explainable convolutional neural networks: Recent developments and promising applications,” in *2023 8th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Jun. 2023, pp. 912–917. doi: 10.1109/ICCES57224.2023.10192764.
- [47] P. R. Chandre *et al.*, “Explainable AI for intrusion prevention: A review of techniques and applications,” in *ICTIS 2023: ICT with Intelligent Applications*, 2023, pp. 339–350. doi: 10.1007/978-981-99-3758-5\_31.
- [48] S. Sahu, R. Kumar, and P. M. Shafi, “Movie rating prediction and viewers’ sentiment trend analysis using YouTube trailer comments,” in *Micro-Electronics and Telecommunication Engineering*, 2023, pp. 127–142. doi: 10.1007/978-981-19-9512-5\_12.

## BIOGRAPHIES OF AUTHORS






**Shahin Shoukat Makubhai**    is a research scholar in the Department of Computer Science and Engineering at MIT School of Computing, MIT ADT, and Pune, India. She received her B.E. degree in Computer Science and Engineering from DKTE Society’s Textile & Engineering Institute (an autonomous institute), Ichalkaranji, India, and her M.Tech. degree in Computer Engineering with specialization in Cloud Computing from Vellore Institute of Technology - VIT Chennai, India in 2020. She is currently pursuing her Ph.D. in Computer Science at MIT-ADT University, Pune, India. She has earned several global certifications and has also contributed to research through patents and copyrights. Her research interests include artificial intelligence and cloud computing. She can be contacted at email: shahin.makubhai@mituniversity.edu.in.



**Ganesh R. Pathak**    received Bachelor of Engineering (B.E. -Computer) from Walchand Institute of Technology, Maharashtra, India, Master of Engineering (M.E. -CSEIT) from Savitribai Phule Pune University (formerly University of Pune), Maharashtra, India and Ph.D. degree in Computer Science and Engineering at Sathyabama Institute of Science and Technology (deemed to be University), Chennai, Tamil Nadu, India. He is presently working as professor in the Department of Computer Science and Engineering, School of Computing, MIT Art, Design and Technology University, Pune. His research interests include artificial intelligence, machine learning, data science, computer networks, wireless communication, and wireless sensor network, especially security in wireless sensor network. His teaching areas include mobile computing, pervasive computing information assurance, and security and usability engineering. He is a member of IEEE and CSI. He can be contacted at email: ganesh.pathak@mituniversity.edu.in.



**Pankaj R. Chandre**    has obtained his B.E degree in Information Technology from Sant Gadge Baba Amravati University, Amravati, India; M.E. degree in Computer Engineering from from Mumbai University Maharashtra, India in the year 2011; and Ph.D. in Computer Engineering from Savitribai Phule Pune University, Pune, India in the year 2021. He is currently working as an associate professor in Department of Computer Science and Engineering, MIT School of Computing, MIT ADT, Pune, India. He has published 60 plus papers at international journals and conferences. He has guided more than 30 plus undergraduate students and 20 plus postgraduate students for projects. His research interests are network security and information security. He can be contacted at email: pankaj.chandre@mituniversity.edu.in.