



An Optimized Extreme Gradient Boosting Regressor Approach Based Lung Cancer Detection

*Sasikala Dhamodaran¹, Shafqat Ul Ahsaan², Naheeda Zaib³ and Shraddha Jaiswal⁴

^{1, 2, 4}NIMS Institute of Engineering & Technology, NIMS University, Jaipur, India

³Department of Computer Science and Application, Vivekananda Global University, Jaipur, India

¹godnnature@gmail.com, ²mailforshafqat@gmail.com,

³naheedazaib93@gmail.com, ⁴meetjais04@gmail.com

Abstract. Presently, the pitiful part of this sphere is that the greatest of them, i.e. comprising a few children, youths on top of the ladies, are habituated to smoking. These vigorous smokers instigate nicotine pollution, thereby escalating the percentage of causing cancer to smokers and non-smokers in society. Lung cancer is the primary basis for demises around the globe and it ranks 4th foremost reason of decease in India. Apart from smoking cancer may be caused due to pollution, disclosure to radon and other infections, this rate of cancer rises gradually. When the lung cancer identification is initial then it is possible to cure it at the earliest. Then the amount of possibility of demise owing to lung cancer can be lessened as soon as possible. This forecast is performed these days stem from Artificial Intelligence (AI) related customs with deep learning (DL) or with machine learning (ML) incessant approaches and IoT centered image classifiers. Currently, this research work is application of varied ML algorithms have enormous applicability's to this arena as of their wise computational expertise for an appropriate forecast of such sicknesses with precise data in view of Lung Cancer (4/30/1992) UC Irvine ML Repository that is utilized here in this exploration.

In-depth investigation and a comprehensive literature analysis in this research work suggested a few algorithms that includes Extreme Gradient Boosting Regressor and continually the next one, i.e., an Optimized Extreme Gradient Boosting Regressor, also reflected as well suited for categorisation of lung cancer. Their experimental outcomes proves that the accuracy, AUC and sensitivity of XGBR and OXGBR are obtained as 91.73%, 0.89, & 77% for the former as well as 94.37%, 0.94 & 91% for the latter respectively. As these results are yet not the desirable ones. Still in future additional hyper-parameters are to be honed to obtain the best desirable outcomes with an intensive investigation put in further continually in this current methodology. For the reason that forecast and execute an additionally honed procedure for prompt identification and exposure of lung cancer besides affording the precise data for the appropriate well-chosen ML algorithm. Then the exact cure is to be tied up for healing this chronic illness, thereby preventing the death and reducing its rate, likewise prolonging the patient's lifespan cobble together their sustenance intended for their upper limit duration in a healthy way with care.

Keywords: Lung Cancer Detection, ML Techniques - XGB, XGBR, and OXGBR, Gamma and Lambda hyper-parameters.

1 Introduction

Presently, the pitiful part of this sphere is that the greatest of them, i.e. comprising a few children, youths on top of the ladies, are habituated to smoking. These vigorous smokers instigate nicotine pollution, thereby escalating the percentage of causing cancer to smokers and non-smokers in society. Lung cancer is the primary basis for demises around the globe and it ranks 4th foremost reason of decease in India [1, 2]. Apart from smoking cancer may be caused due to pollution, disclosure to radon and other infections, this rate of cancer rises gradually. When the lung cancer identification is initial then it is possible to cure it at the earliest. Then the amount of possibility of demise owing to lung cancer can be lessened as soon as possible. This forecast is performed these days stem from Artificial Intelligence (AI) related customs with deep learning (DL) or with machine learning (ML) incessant approaches [3 to 18] and IoT centered image classifiers [3]. These methods have the competence in ascertaining cancer at a preliminary stage and maintain the life stretch of the patients by associating an oncologists in perceiving and predict the fitness state of them. Then next that will infer cancer risk warnings and purge them. The practice of selection and attribute extraction by CT images implements a vibrant role in labelling the dynamic symptoms of cancer. Identification of such of these high-risk patients at the apt time and providing the essential treatment and therapies have the prospective to significantly progress lung cancer individuals' persistence rates.

In this modern era there are enormous investigational works performed in this domain. There are a lot of datasets available for experimentations in this area. Currently, ML approaches have enormous applicability's to this field because of their wise computational expertise for an appropriate forecast of such sicknesses with precise data in view of Lung Cancer (4/30/1992) UC Irvine ML Repository [4]. Now, in this exploration, several ML classifiers customs to classify the existing lung cancer data as benign or malignant that are insisted. The CT (Computed Tomography) image diagnosing data imageries are pre-processed and restructured as binary depictions. These depictions are then categorized using a planned limited methods in Python programming language using Sklearn to classify the existing UCI dataset into cancerous and non-cancerous data, they are, K-nearest neighbours (KNN), naive Bayes (NB), SVM, decision tree (DT), LR, RF, LDA, gradient boosting (GB), extreme gradient boosting (XGB) [10] and extreme gradient boosting regressor (XGBR) are the prevailing ML algorithms. XGBR was the previous research work performed. This exploration gave an estimation system that revealed that the then put forward XGBoost Regressor (XGBR) that had led to an accuracy of 91.73%. But although this XGBR is a capable process for prompt lung cancer prediction, it has to be further enhanced for the same.

But then again as the progress was only very moderate in this recommended work too. Here is an Optimized XGBR recognized as OXGBR approach found by evolving

the XGBR predictive model of detecting lung cancer in patients. This present investigation work realized the accuracy outcome as 94.37% that is agreeable compared to the previous findings. On the other hand, for the medical data the consequences must to be the highest in terms of classification, so the investigation is still being processed i.e., to forge ahead up until demanded.

2 Literature Survey

This exploratory work [5] analyses the various techniques of feature selection for predicting the lung cancer incidence of histological subtypes by using ML methods. The goal of this exploration task [6] is to assess various ML approaches generally for learning and diagnosis of imminent enhancement of lung cancer recognition to examine for an inoperative lung cancer, also to examine for a void in the imminent enhancement. This ability is openly suitable for medical treatments specifically for finding and decision of lung cancer as they rely on multifarious measures of proteomics and genomics. This research [7], sketch and build up a risk forecast prototype to realize lung cancer illness patterns. Here Naïve Bayes (NB) and SVM (Support Vector Machine) classification processes were submitted to label the lung cancer information, and their accuracy were examined. The intention of this exploration [8] is to evaluate the proficiency of SVM and logistic regression (LR) methods to predict the survival rates of the sick lung cancer individuals. These practices support clinicians in imparting precise prognosis. The interface permits the user to state the essential factors and pick out if the individual is a cancer patient or not.

In this investigative task [9], data pre-processing is done first. Then next step is the feature selection with data filtering methods. Finally, using these ML types, for illustration, SVM, also random forest (RF) to categorize lung cancer patients from the normal individuals. The delayed diagnosis will go ahead in the speedy multiply of cancer to lungs and further to other organs in the physique turn out to be life endangering. The researchers based on their exhaustive literature analysis specifies that the delay spotted and initially detected persistence rate of lung cancer instances in five-years are merely 4%, and 16% respectively. The foremost reason that overlays for the delay is that the preliminary clinical indications are frequently ignored as irrelevant, also the global technique used for identification is radiomics that is very costly and the frequent withholding to it itself is too risky. In most circumstances, the diagnosis was at a later phase. Usage of clinical facts is a promising alternate for radiomics in quick identification of this disease. Here it was attempted to recognize the forecasters from clinical lung cancer facts by involving ML algorithms. Over and done with the secondary information analysis, the furthestmost powerful forecasters of this disease were recognized from the clinical facts. These predictive markers will support in the initial discovery of this disease, also when the patients that hit the identified predictive marker standards will be categorized as possibly risky group in cultivated lung cancer. At first, in this method, the clinical data is pre-processed and then next, the feature selection is utilized with the data filtering technique. At last, this pre-processed clinical data is employed in ML

prototypes for instance, SVM and RF to categorize the patients. This investigation [11] targets to adopt a proper procedure for recognizing lung cancer and its relationship with the lung cancer patient's way of life. This research work visualizes to associate twelve prospective ML algorithms across the clinical facts with eleven indications of lung cancer besides two main customs of patients to forecast a positive instance precisely. The end result includes the set up established on categorization and heat-map association. KNN, also Bernoulli NB prototypes are realized as the supreme noteworthy approaches for the prompt lung cancer forecasting. The exploration of this review of literature work [12] demands, imminent lung cancer therapies, so in view of that an ML methodological decision-accomplishing tool that must be a multi-modal method that ponders both imaging facts, also, omics data processing, and the blend of various data types. Last of all these complex tasks are anticipated to impact progression of investigations highlighting on the lung cancer therapies. In this investigative work [13], the LIDC-IDRI imagery information sets were employed and attained this multiresolution rigid registration (MRR) method, as well as, discrete wavelet transform (DWT), also principal component averaging for the recognition of lung cancer and for analysis ResNet-18 ascertains the exceptional operations of isolating lung cancer.

This exploration [14] acquires the customary cutting-edge imageries with ML methods for implementation of lung cancer determinations putting an end to it. This investigation specifies a precise categorization and forecasting of lung cancer by means of ML as well as image processing aided technology. To start with, photographs are essential to be collected. In the experimental analysis of this research work, real-time 70 different patients with their 83 CT scans were employed as the dataset. Throughout the pre-processing of these images, the geometric mean filter was applied, as a result, the quality of those 83 CT scan images were enriched. The K-means procedure then segmented those imageries. The fragments of them are set up by means of this segmentation. At that moment, categorization approaches via ML algorithms are carried out, for instance, ANN, KNN, and RF were executed. In addition to those, the ANN prototype is yielding more precise outcomes for forecasting lung cancer promptly. A profound analysis of diverse ML procedures' execution is specified in this exploratory function [15] of lung cancer labelling. In research [16] hyper-parameter tuning was performed based on the Gamma and C parameters with value 10 in Kaggle information set and these impacts the kernel width as well as regularization strength, one-to-one. Therefore, amplified lung cancer recognition forecasting. Forecasting lung cancer [17] by using Lung Imaging Database Consortium (LIDC) database, with an effectual CT categorization applying hyper parameter tuned five-fold cross validation function was used. Especially when integrating GLCM and SVM, it executed comparatively well. This investigation work [18] predicted the lung cancer well by hyper-parameter tuning for the CNN models where these further heightened the performance of the lung cancer detection. Five-fold cross validation function with Grid search was utilized and also various values of Batch Sizes 4,8,16,32, Epochs 5, 10,15,20, and Learning Rates 0.1,0.01,0.001 were the hyper-parameters tuned. A 'softmax' activation function with is compiled with 'adam' optimizer with Batch Size of 32, 20 Epochs with 0.001 Learning Rate are the optimum factors, set with the Grid Search Cross-Validation (GSCV)

procedure by utilizing five-fold cross validation, were then put into operation in creating this CNN prototype.

3 Prevailing Methodologies

Exploration works [3-18] ascertain that ML categorizers are exceptionally suitable for prompt lung cancer identification and forecasting. At that moment the relevant treatment is made certain to heal these lung cancer patients. Maximum research works utilize, KNN, NB, and SVM are the conventional ML approaches employed and considered for their mandatory measures, that includes accuracy, AUC and sensitivity, next subsequently, RF, LR, GB, XGB[10]. Then a few more categorizers including the current topic of interested research works XGBR and OXGBR were the advancement of XGB are also involved. The discussion about these XGB, XGBR and OXGBR are available subsequently.

3.1 Extreme Gradient Boost Classifier

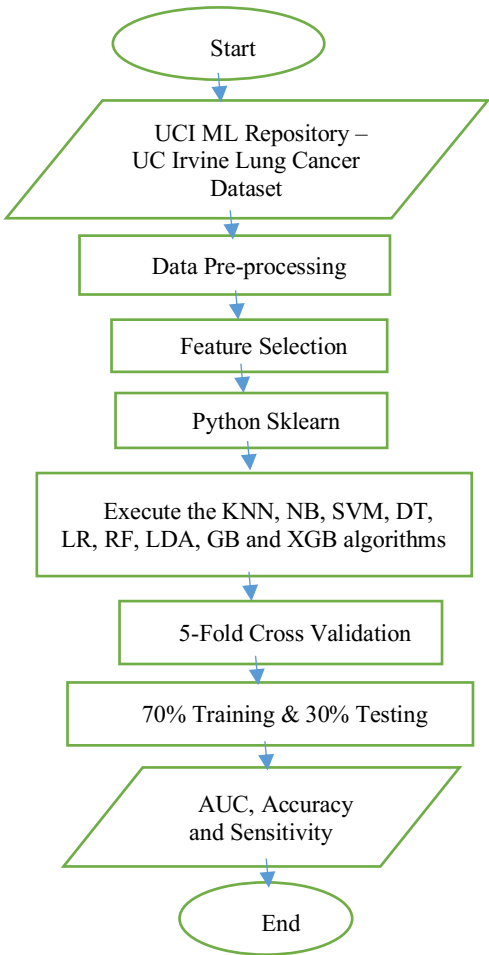


Fig. 1. Flowchart of ML algorithms for lung cancer prediction

After realizing various experimentations after an exhaustive review of literature, [10] Extreme Gradient Boosting (XGB) was well-thought-out for categorization of lung cancer. This was the first step of research, where real-time data was used for predicting lung cancer in this particular exploration. Therefore this XGB algorithm and the traditional KNN, NB, SVM, DT, LR, RF, LDA, and GB algorithms were implemented with the dataset used in research work [4].

Fig.1. depicts the flowchart of ML algorithms chosen for lung cancer prediction. This research investigation provided an assessment system that discovered that the prevailing XGB had stemmed with an accuracy of 73.24% and sensitivity as 72% that are to be highly developed, but AUC is 0.80, then again it also should be enriched. This exploration proves that restrictions of these prevailing ML approaches including XGB contains the succeeding specifics:

1. Complex optimality of the infinite Kernels, creating it tough to choose the optimal one with its generality.
2. Computationally in-depth and multifaceted.
3. Possibly will over fit the sample.

3.2 Extreme Gradient Boost Regressor

After carrying out these testing's, as the outcomes obtained were undesirable an intensive literature analysis was performed, [19] Extreme Gradient Boosting Regressor (XGBR) was ingenious for labelling of lung cancer. This was the second stage of this research work, where with the same dataset used in research work [4], this XGBR algorithm was executed.

xgbregressor source code is available in github. Here XGBR is modelled with hyper-parameters set as the following, `n_estimators=100`, `max_depth=5`, `eta=0.1`, 5 fold cross validation function is set and the other hyper-parameters are set to their default values. Then their metrics are got by the following Packages Functions, `accuracy_score`, `roc_auc_score` and `recall_score`.

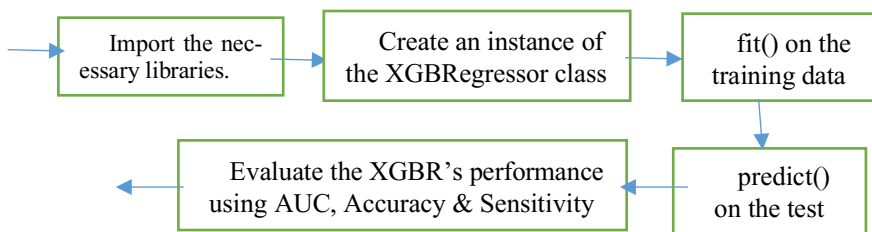
**Fig. 2.** Flowchart of XGBR algorithm for lung cancer prediction

Fig.2. give a picture of the flowchart of XGBR algorithm for lung cancer prediction. This research exploration gave an evaluation system that revealed that the prevailing XGBR had made available with an accuracy of 91.73% and sensitivity as 77% that are to be still more amplified, but AUC is 0.89, then yet again it too must be enhanced. This investigation ascertains that the drawbacks of this mentioned XGBR method are found below as: 1. Ability in handling the omitted data. 2. Massive information sets are held powerfully. 3. Plentiful hyper parameters existing are to be tuned and adjusted values have to rise this XGBR implementation.

Subsequently implementing these test's, as the results achieved were unfavourable a severe literature exploration was done, [13] Extreme Gradient Boosting Regressor (XGBR) was resourceful for classification of lung cancer and that was the third phase of this investigation work, where with the same dataset used in exploration [4], this Optimized XGBR algorithm (OXGBR) was implemented. That is first this XGB procedure had a trivial number of confines that were remedied by handling an Extreme Gradient Boosting Regressor (XGBR) second and then again in its place of an Optimized Extreme Gradient Boosting Regressor (OXGBR) third.

4 Recommended Methodology: Optimized Extreme Gradient Boosting Regressor

In-depth investigation and a comprehensive literature analysis, [13] Optimized Extreme Gradient Boosting Regressor is reflected as well suited for categorisation of lung cancer. As this procedure have a lot of returns that were mended by this Optimized Extreme Gradient Boosting Regressor instead of an Extreme Gradient Boosting Regressor. This research examination delivered an estimation scheme that uncovered that the suggested OXGBR had resulted with a great accuracy of 94.37%. So this OXGBR is a proficient process for initial forecast of lung cancer. In OXGBR, hyper-parameter tuning is performed for $cv=5$, Gamma and Lambda parameters and those values are set as 10.

$$XGBRegressor(\text{gamma: } 10, \text{lambda: } 10, \text{cv}=5)$$

The benefits of this suggested procedure OXGBR are discussed as listed below:

1. Gamma parameter when set to value 10,
 - (i) Regularization in XGBR supports to ease this.
 - (ii) Dropping the Prototype Complexity: By reprimanding larger factors, regularization streamlines the prototype.
 - (iii) Cultivating Simplification: Guarantees that the prototype implements sound on fresh data.
 - (iv) Put a stop to Overfitting: Retains the prototype suitable for excessively custom-made in the training data.
2. Lambda parameter when set to value 10,
 - (i) Regulates bias-variance interchange.

- (ii) Deal severely with upper values.
- (iii) Precludes overfitting.
- (iv) Lessens prototype complexity steering to a fewer-overfit-and-more-general prototype.

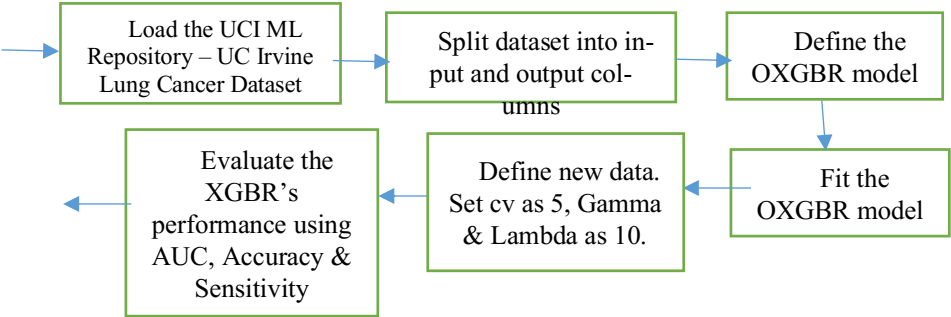


Fig. 3. Flowchart of OXGBR algorithm for lung cancer prediction

Table 1. Implementation of ML algorithms for lung cancer prediction

S.No	Methodologies	AUC	Accuracy	Sensitivity
	Prevailing Methodologies:			
1.	KNN	0.79	93.6%	92%
2.	NB	0.76	91.18%	91%
3.	SVM	0.78	91.24%	92%
4.	DT	0.76	86.3%	84%
5.	LR	0.72	89.2%	86%
6.	RF	0.78	90.7%	87%
7.	LDA	0.77	90.6%	88%
8.	GB	0.79	72.88%	71%
9.	XGB	0.80	73.24%	72%
10.	XGBR	0.89	91.73%	77%
11.	Recommended Methodology: OXGBR	0.94	94.37%	91%

Table 1. Illustrates that the accuracy OXGBR is upper when associated to the other ML algorithms for lung cancer detection. All the same that Accuracy, AUC and Sensitivity measures are supreme in this recommended OXGBR. SVM too gives good results identical to OXGBR in terms of Sensitivity measure.

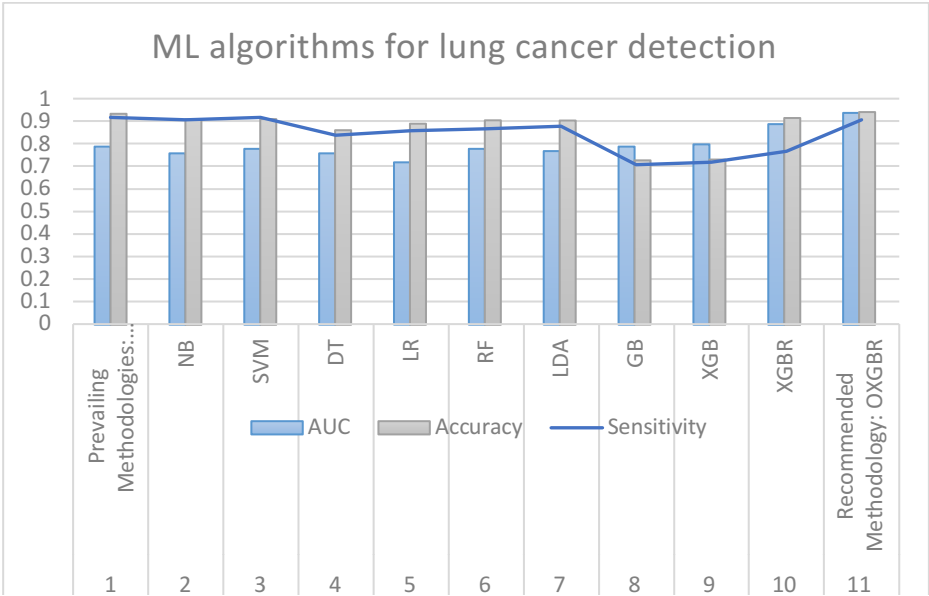


Fig. 4. Comparison of ML algorithms for lung cancer prediction

Fig.3. provide a representation of the flowchart of OXGBR algorithm for lung cancer prediction. Fig. 4 displays that the OXGBR’s sensitivity = 91% though AUC is superior in SVM, better than other techniques. For that reason, it is needed to accomplish further exploration to enhance the sensitivity measure in the OXGBR algorithm. It is essential that the other ML algorithms too have to be further enriched in terms of the same sensitivity measure. There is also a necessity that the other measures accuracy and AUC must be improved for all the ML algorithms. For that reason, it is the need of the hour for advancing immensity in this investigation, also it is being embraced continually for discovery and subsequently curing this deadly disease prolonging the lifespan of these patients with best health.

5 Conclusion And Future Scope

From the investigation of the research work it is exposed that the new Optimized Extreme Gradient Boosting Regressor is improved with measures AUC and Accuracy when associated to the prevalent ML procedures comprising of Extreme Gradient Boosting Regressor, whereas SVM is similarly just as participating with its functioning excluding for its sensitivity measure. SVM is maximum related to the sensitivity measure. So, explore the other hyper-parameters for elevation of OXGBR in terms of sensitivity too along with exploring some additional innovative procedures to upsurge the execution in relation to the OXGBR process efficacy. Here for the OXGBR the accuracy acquired is 94.37%, but still as it is the medical data, the accuracy should be

as maximum as possible. Also its AUC is 0.94 and sensitivity is 91% respectively. Consequently, it is desirable to rise these measures in addition to recuperate the quality of categorization and precisely diagnose this fatal disease lung cancer for appropriate therapies, healing and prolonging their healthy lifespan sustainable to a maximum extent possible. Hence forecast and execute an additionally honed procedure for prompt identification and exposure of lung cancer besides affording the precise data for the appropriate well-chosen ML algorithm. Then the exact cure is to be tied up for healing this chronic illness, thereby preventing the death and reducing its rate, likewise prolonging the patient's lifespan cobble together their sustenance intended for their upper limit duration in a healthy way with care.

Acknowledgments. Grateful for NIMS University and VGU University Management and for all those who have directly and indirectly helped in our research work.

References

- [1]. V. Noronha, R. Pinninti, V. M. Patil, A. Joshi, and K. Prabhash, "Lung cancer in the Indian subcontinent," *South Asian J. Cancer*, vol. 5, no. 3, pp. 95–103, Jul.-Sep. 2016, doi: 10.4103/2278-330X.187571.
- [2]. V. Gumdal, "Lung cancer: The deadliest cancer," *Times of India*, Mediawire, News, Nov. 10, 2023, 12:15 IST.
- [3]. A. K. Dutta, "Detecting lung cancer using machine learning techniques," *Intelligent Automation & Soft Computing*, Tech Science Press, vol. 2021, received Apr. 25, 2021; accepted Jun. 24, 2021, doi: 10.32604/iasc.2022.019778.
- [4]. R. Patra, "Prediction of lung cancer using machine learning classifier," in *COMS2 2020*, N. Chaubey et al., Eds., vol. 1235, Springer Nature Singapore, 2020, pp. 132–142.
- [5]. S. S., "A survey on non-small cell lung cancer prediction using machine learning methods," in *2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, Springer Nature, 1 Oct. 2020, pp. 255–266, India.
- [6]. A. Jaweed and F. Siddiqui, "Implementation of machine learning in lung cancer prediction and prognosis: A review," in *Cyber Intelligence and Information Retrieval*, Springer Nature, 29 Sep. 2021, pp. 225–231, Singapore.
- [7]. V. Rani, S. Das, and T. K. Kundu, "Risk prediction model for lung cancer disease using machine learning techniques," in *Innovations in Computer Science and Engineering*, Springer Nature, 26 Mar. 2022, pp. 417–425, Singapore.
- [8]. S. D. Kavila, L. S. V. S. S. S., R. Bandaru, S. Riyaz, and N. Sai Venkata Rushikesh, "Lung cancer disease prediction using machine learning techniques," in *Machine Learning and Big Data Analytics (ICMLBDA 2022)*, Springer Nature, 7 Jun. 2023, pp. 501–514, India.
- [9]. A. M. Sebastian and D. Peter, "Identifying the predictors from lung cancer data using machine learning," in *Sentiment Analysis and Deep Learning Conference*, 1 Jan. 2023, pp. 691–701, Thailand.
- [10]. X. Guan, Y. Du, R. Ma, N. Teng, S. Ou, H. Zhao, and X. Li, "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices," *BMC Med. Informatics Decis. Making*, vol. 23, no. 107, pp. 1–16, 2023.
- [11]. S. P. Maurya, P. S. Sisodia, R. Mishra, and D. P. Singh, "Performance of machine learning algorithms for lung cancer prediction: A comparative approach," *Springer Nature Sci. Rep.*, vol. 14, no. 18562, pp. 1–12, 2024.

- [12]. Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine learning for lung cancer diagnosis, treatment, and prognosis," *Genomics Proteomics Bioinformatics*, vol. 20, no. 5, pp. 850–866, Dec. 2022, doi: 10.1016/j.gpb.2022.11.003.
- [13]. I. Nazir, I. ul Haq, S. A. AlQahtani, M. M. Jadoon, and M. Dahshan, "Machine learning-based lung cancer detection using multiview image registration and fusion," *J. Sensors*, Wiley Online Library, 16 Aug. 2023, pp. 1–19, doi: 10.1155/2023/6683438.
- [14]. S. Nageswaran, A. G. A. K. Bisht, S. Mewada, S. K. J. N. V. R., M. Jawarneh, and E. Asenso, "Lung cancer classification and prediction using machine learning and image processing," *Biomed. Res. Int.*, vol. 2022, pp. 1755460, 22 Aug. 2022, doi: 10.1155/2022/1755460.
- [15]. S. Dodia and A. B., "Chapter 7 - Machine learning-based detection and classification of lung cancer," in *Computational Intelligence in Cancer Diagnosis, Progress and Challenges*, Academic Press, 2023, pp. 149–174.
- [16]. S. M. Nabeel, S. U. Bazai, N. Alasbali, Y. Liu, M. I. Ghafoor, R. Khan, C. S. Ku, J. Yang, S. Shahab, and L. Y. Por, "Optimizing lung cancer classification through hyperparameter tuning," *Digital Health*, vol. 10, pp. 1–15, 2024, doi: 10.1177/20552076241249661.
- [17]. P. R. Urs M and S. B. G. N. K., "Optimized hyperparameter approach for lung cancer prediction and classification of types using GLCM," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4, pp. 1873–1878, 2024.
- [18]. S. Bhattacharyya, "Lung cancer detection using hyperparameter tuned convolution neural network," *Int. J. Novel Res. Dev.*, vol. 8, no. 12, pp. 867–873, 2023.
- [19]. F. Sarder, S. Akter, and S. Akter, "Predicting dengue outbreak from climate data using machine learning algorithms," in *2022 IEEE Int. Conf. Data Sci. Inf. Syst. (ICDSIS)*, India, 2022, doi: 10.1109/ICDSIS55133.2022.9915862.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

