International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 09 | September - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Machine Learning Approaches in Early Lung Cancer Prediction: A Comprehensive Review

R.Shanthi Krishna [1], Dr.T.Vijaya Kumar [2]

[1] Assistant Professor, Department of Computer Applications, SSMRV College, Karnataka, India
[2] Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India

Abstract - *Diagnosing cancer at an early stage helps the doctors for the successful treatment of the disease. Lung Cancer, a leading cause of cancer-related deaths globally, has emphasized the importance of early detection to enhance patient survival outcomes. This paper offers a comprehensive review of various machine learning algorithms employed in the prediction and early detection of lung cancer. Through a diligent survey of recent literature, we evaluated the effectiveness of techniques ranging from ensemble learning methods to regression and classification algorithms. Several studies exhibited promising results, with some algorithms achieving accuracies upwards of 99%. Particularly, SVM, Logistic Regression and ensemble methods consistently demonstrated high prediction accuracies across multiple datasets. However, there remains substantial potential to expand on these findings, especially in the domain of hybrid model development, real time predictions and seamless clinical integrations. This review underscores the transformative role of ML in Lung Cancer diagnostics and charts a course for future exploration in this critical medical domain.*

Keywords: Machine Learning, Lung Cancer, Classification Algorithm, SVM, Random Forest, XGBoost

I Introduction

Cancer are a group of diseases associated with abnormal growth of cells which are called as Malignant Tumour. Malignant Tumour can arise anywhere in the body and can affect people from all age-groups, socio-economic strata and race. As per the International Agency for Research on Cancer's GLOBOCAN 2020 database, there are 36 cancer types in 185 countries worldwide.

As per WHO, the most common causes of cancer death in 2020 were 1.80 million due to Lung cancer, 916000 due to Colon and rectum, 830000 due to Liver cancer 769000 due to Stomach cancer and 685000 due to Breast cancer [1].

A Cancer that starts in lung is Primary Lung Cancer whereas which starts in lung and spread to other parts of the body is Secondary Lung Cancer. An Early Stage Cancer is a small cancer that is diagnosed in lung. Lung Cancer is seen on Chest radiographs and Computed Tomography (CT) scan and the lung cancer biopsy is done by bronchoscopy. The mortality rate due to lung cancer is increasing day by day in humans irrespective of the age. Thereby it becomes highly important to detect lung cancer at an early stage so that necessary treatments are taken immediately.

In the era of big data, Machine Learning (ML) provides tools for harnessing healthcare data to predict lung cancer with greater accuracy and speed than traditional methods. The objective of this survey paper is to examine and evaluate the effectiveness of various machine learning algorithm in predicting lung cancer based on clinical, radiological and patient data. Early detection of lung cancer is essential for patient survival and machine learning-based prediction models have potential use in predicting lung cancer.

II Literature Survey

Researchers have been working on various algorithms in detecting lung cancer at early stage. The Past research work done on detecting lung cancer at an early stage is specified below:

Muntasir Mamun et al. used Ensemble Learning Techniques which are a subgroup of machine learning algorithms to

evaluate lung cancer prediction. They have applied four types of ensemble learning techniques such as XGBoost, AdaBoost, Bagging and Light GBM classifiers to predict lung cancer. Among these techniques, XGBoost achieved the highest accuracy of 94.42% for predicting lung cancer [2].Radhika P R etal. used classification algorithms under machine learning to evaluate the lung cancer prediction. They used 2 different datasets for their study – UCI Machine Learning Repository and Data World. The authors used the following classification algorithms such as SVM, Logistic Regression, Naïve Bayes and Decision Tree. From UCI Machine Learning Repository dataset, logistic regression algorithm got an accuracy rate of 96.9% in detecting the lung cancer. From Data World dataset, Support Vector Machine algorithm got an accuracy rate of 99.2% among all other algorithms in detecting lung cancer [3]. Dr.K.Sivanagireddy et al. used the Kaggle dataset to identify the association between Lung Cancer and various symptoms of the same and they have used various Regression Algorithms. Among the various regression algorithms they have implemented, it is discovered that multiple regression algorithm has a high accuracy of 96% in predicting the lung cancer [4]. Lal Hussain et al. used the dataset provided by the Lung Cancer Alliance. They have used different Image Enhancement methods to improve the Image Quality. With that they have computed the texture features for predicting the lung cancer detection. Among the various machine learning algorithms, SVM Gaussian , RBF, Decision Tree, SVM Polynomial, Naïve Bayes – the SVM Gaussian , RBF and SVM Polynomial provided the highest performance with an accuracy of 99.89% in predicting the lung cancer [5].

Dr S VenkataLakshmi et al. used Kaggle Dataset for predicting the Lung cancer. The research team implemented various classification algorithms like Logistic Regression, Linear Discriminant Analysis, K- Nearest Neighbours and Naïve Bayes. In predicting the lung cancer detection, Logistic Regression and KNN algorithms provided 100% accuracy with the model they have proposed [6]. In [7], Hemant Jaiman et al. analyzed various machine learning classification algorithms such as Naïve Bayes, Decision Tree, SVM and Logistic regression in predicting the lung cancer. An accuracy of 99.2% is obtained for SVM model for predicting the lung cancer. Rahat Idrees et al. utilized various Supervised Machine Learning algorithms to detect the early lung cancer stage. They have undergone ANN, SVM, Random Forest and Multiple Linear Regression algorithms. Out of which Random Forest model gave the best result with an accuracy of 99.99% in predicting the lung cancer [8]. C.S.Anita et al. used UCI Machine Learning Repository for predicting the Lung Cancer. They have implemented various machine learning techniques like SVM, Random Forest, Naive Bayes, Artificial Neural Networks and Gaussian Naïve Bayes and an accuracy of 98% is predicted from GNB model [9]. Makarov et al. analyzed the dataset which is taken from EROB database for predicting the accuracy of the machine learning model. They have implemented the following machine learning algorithms like Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, KNN Classifier and predicted an accuracy of 70% by Gradient Boosting Classifier and Random Forest [10]. R Patra analyzed various machine learning techniques to predict the lung cancer from the dataset taken from UCI Machine Learning Repository. The comparison technique reveals RBF classifier has resulted with an accuracy rate of 81.25% among all other techniques like KNN, Naïve Bayes and J48 classifier [11]. Y.Gultepe used the dataset taken from the UCI machine learning repository. The researcher applied the dataset to the following classification algorithms like Random Forest, K-Nearest Neighbor, Naïve Bayes, Logistic regression, Decision Trees and SVM and predicted am accuracy of 83% modeled with KNN [12]. Yunpeng Cuit et al. analyzed the performance of machine learning algorithms like Logistic Regression, XGBoosting Machine, Random Forest, Gradient Boosting Machine, Neural Network and decision tree in predicting the early stage lung cancer. The team came with an accuracy of 77.2% when the model is implemented with Logistic Regression, XGBoosting Machine and Gradient Boosting Machine [13].

Dakhaz et al. used the lung cancer dataset that was made available in the UCI machine learning repository. The Dataset is implemented in the following machine learning algorithms SVM, KNN and CNN and predicted an accuracy of 95.56% with SVM Model [14]. Kasthuri et al. used the dataset taken from UCI machine learning repository and implemented in the following machine learning algorithms- SVM, Naïve Bayes, KNN and Logistic Regression. The team predicted an accuracy of 82.25% with SVM model [15].

III Methodology

The Primary Objective of this survey paper is to review and analyze various researches that have implemented machine learning algorithms for early stage detection of lung cancer. The Primary Sources were research papers published between 2019 and 2022, with a focus on the methodologies used, datasets employed and the achieved results.

Studies that specifically utilized machine learning techniques for lung cancer prediction and provided clearresults in terms of accuracy percentages were selected for review and it was analyzed.

IV Results

A broad spectrum of machine learning techniques for the detection of early-stage lung cancer were used.These techniques ranged from ensemble learning techniques to classification and regression algorithms.

Muntasir Mamun et al. reported a 94.42% accuracy using the XGBoost ensemble learning technique.

Radhika P R et al. noted a 96.9% accuracy with logistic regression using the UCI Machine Learning Repository dataset and a 99.2% accuracy with SVM using the Data World dataset.

Dr.K.Sivanagireddy et al. reported a remarkable 96% accuracy with Multiple Regression with the Kaggledataset.

Lal Hussain et al. recorded an accuracy rate of 99.89% with SVM Gaussian, SVM Polynomial and RBF withLung Cancer Alliance dataset.

Dr.S.Venkatalakshmi et al. reported a 100% accuracy rate for both logistic regression and KNN algorithmusing the Kaggle dataset.

Hemant Jaiman et al. achieved an accuracy of 99.2% using the SVM model.

Rahat Idrees et al. recorded a near perfect 99.99% accuracy with Random Forest model.

C.S.Anita et al. reported an accuracy rate of 98% with Gaussian Naïve Bayes Model using the UCI Machine Learning Repository Dataset.

Makarov et al. recorded an accuracy rate of 70% predicted from Gradient Boosting Classifier and RandomForest algorithm.

R Patra noted an accuracy rate of 81.25% with Radial Basis Function Network(RBF) Classifier using theUCI Machine Learning Repository Dataset.

Y.Gultepe reported an accuracy rate of 83% with KNN model using UCI Machine Learning Repository.

Yunpeng Cuit et al. noted an accuracy rate of 77.2% with Logistic Regression, XGBoosting Machine and Gradient Boosting Machine.

Dakhaz et al. reported a 95.56% accuracy using the Support Vector Machine model.

Kasthuri et al. recorded an accuracy of 82.25% with SVM Model which is implemented in UCI Machine Learning Repository Dataset.

The following table shows the Summary of Literature on Machine Learning Approaches for Lung Cancer Prediction.

Reference	Dataset Details	ML Classifier	Accuracy Rate
Muntasir Mamun et al.[2]	**Kaggle** 309 Instances 16 Attributes	**XGBoost**	**94.42%**
		LightGBM	92.558%
		AdaBoost	90.70%
		Bagging	89.76%
Radhika P R et al.[3]	**UCI MachineRepository Learning** 32 Instances 57 Attributes	Logistic Regression	96.9%
		Decision Tree	85.71%
	Data World 1000 Instances 25 Attributes	Logistic Regression	66.7%
		Decision Tree	90%
		Naïve Bayes	87.87%
		SVM	**99.2%**
Dr.K.Sivanagireddy et al.[4]	**Kaggle** 15 Attributes	Linear Regression	85%
		Logistic Regression	95%
		Logarithmic Regression	86%
		Multiple Regression	**96%**
		Exponential Regression	85%
Lal Hussain et al.[5]	**Lung Cancer Alliance** 945 Instances	**SVM Gaussian**	**99.89%**
		RBF	**99.89%**
		SVM Polynomial	**99.89%**
		Decision Tree	99.35%
		Naïve Bayes	88.57%
Dr.S.Venkatalakshmi et al.[6]	**Kaggle** 20 Attributes	**Logistic Regression**	**100%**
		Linear Discriminant Analysis	95.89%
		K-Nearest Neighbours	**100%**
		Naïve Bayes	90.41%
Hemant Jaiman et al.[7]	**Data World** 1000 Instances 25 Attributes	**Support Vector Machine**	**99.2%**
		Decision Tree	90%
		Naïve Bayes	87.87%
		Logistic Regression	66.7%
Rahat Idrees et al.[8]	**Data World** 1000 Instances 25 Attributes	Artificial Neural Network	65.75%
		Multiple Linear Regression	77.54%
		Random Forest	**99.99%**
		Support Vector Machine	98.91%

C.S.Anita et al.[9]	**UCI Machine Learning Repository** 14 attributes	Support Vector Machine	78%
		Random Forest	87%
		Naïve Bayes	88%
		Artificial Neural Networks	89%
		Gaussian Naïve Bayes	**98%**
Makarov et al.[10]	**EROB Database** 19379 Instances 15 Attributes	Decision Tree	63%
		Random Forest	**70%**
		Gradient Boosting	**70%**
		Logistic Regression	69%
		KNN	68%
R Patra[11]	**UCI Machine Learning Repository** 32 Instances 57 Attributes	KNN	75%
		Naïve Bayes	78.125%
		Radial Basis Function Network	**81.25%**
		J48	78.12%
Y.Gultepe[12]	**UCI Machine Learning Repository** 32 Instances 57 Attributes	Random Forest	33%
		KNN	**83%**
		Naïve Bayes	67%
		Logistic Regression	50%
		Decision Tree	33%
		Support Vector Machine	50%
Yunpeng Cuit et al.[13]	**Oncologic Database of US** https://seer.cancer.gov/19887 Instances 12 Attributes	**Logistic Regression**	**77.2%**
		XGBoostingMachine	**77.2%**
		Random Forest	77.1%
		Neural Network	77.1%
		Gradient Boosting Machine	**77.2%**
		Decision Tree	76.8%
Dakhaz et al.[14]	**UCI Machine Learning Repository** 32 Instances 56 Attributes	**SVM**	**95.56%**
		KNN	89.65%
		Convolutional Neural Network	92.11%
Kasthuri et al.[15]	**UCI Machine Learning Repository** 32 Instances 57 Attributes	KNN	76%
		Naïve Bayes	79.125%
		SVM	**82.25%**
		Logistic Regression	79.12%

XGBoost, Logistic Regression, SVM and Random Forest were some of the top performing algorithms across different studies. The accuracy rate of these algorithms ranged from 70% to 100% with some studies reporting perfect accuracy in their prediction.

Conclusion

In Summary, the reviewed literature underscores the potential of machine learning as an effective tool for early-stage lung cancer detection, with several algorithms demonstrating high accuracy rates across various datasets. While the current literature showcases the promise of ML in early stage lung cancer detection, there are ample avenues for exploration. Future work cannot only focus on enhancing accuracy but also on ensuring that tools are ethically sound, transparent , integrated and accessible to all.

Future Work

Based on the reviewed literature and the observations made, the following future work directions can be identified.

There's a potential to analyze more extensive and diverse datasets, which can capture a broader spectrum of patient data from various demographics, geographies and clinical settings.

Hybrid models that combine the strengths of multiple algorithms can be explored to possibly improve accuracy and robustness of predictions.

With the advances in Neural Networks and Deep Learning, Convolutional Neural Networks and Recurrent Neural Networks could be explored, especially for Image Data from CT Scan and Radiographs.

Integrating patient history, genetics and lifestyle factors into models for a more holistic approach to early lung cancer detection.

References

1. https://www.who.int/news-room/fact-sheets/detail/cancer
2. Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, "Lung Cancer Prediction model using Ensemble Learning Techniques and a systematic review analysis", IEEE World AIIoT Congress , Dec' 2022, DOI: 10.1109/AIIoT54504.2022.9817326
3. Radhika P R, Rakhi A S Nair, Veena G, " A Comparative study of Lung Cancer Detection using Machine Learning Algorithm" , 2019 IEEE International Conference on Electrical, Computer and Communication Technologies, 20-22 February 2019, DOI : 10.1109/ICECCT.2019.8869001
4. Dr.K.Sivanagireddy, Dr.Srinivas Yerram, S.Sri Nandhini Kowsalya, S.S. Sivasankari, J.Surendiran, R.G.Vidhya, "Early lung cancer Prediction using Correlation and Regression" , 2022 International Conference on Computer, Power and Communications (ICCPC), 14-16 Dec' 2022, DOI : 10.1109/ICCPC55978.2022.10072059
5. Lal Hussain,Alsolai H, Hassina S B H, Nour M K, Duhayyium M A, Hilal A M, Salama A S, Motwakel A, Yaseen I, Rizwanullah M, " Lung Cancer Prediction using Robust Machine Learning and Image Enhancement Methods on extracted Gray-level Co-occurrence Matrix Features", Applied Science June 2022,12,6517, DOI: 10.33.90/app121366517
6. Dr.S.VenkataLakshmi, Bhasetty Greeshma, M J Thanooj, K Revanth Reddy, K Rohith Rakesh, "Lung Cancer Detection and Stage Classification using Supervised Algorithms", Turkish Journal of Physiotherapy and Rehabilitation 2021.
7. Hemant Jaiman, Dr.Kuldeep Sharma, Sujatha K, "Survey on Lung Cancer Detection using Machine Learning", International Journal for Research in Applied Science & Engineering Technology , June 2020. DOI : 10.22214/ijraset.2020.6323
8. Rahat Idrees, Muhammed Kamran Abid, Saleem Raza, Muhammed Kashif, Muhammed Waqas, Mubashir Ali, Laiba rehman, "Lung Cancer Detection using Supervised Machine Learning Techniques", LGU Research Journal of Computer Science & IT Jan'-Mar' 2022. DOI : 10.54692/lgurjcsit.2022.0601276
9. C.S.Anita, Vasukidevi G., D.Rajalakshmi,K.Selvi, Ramesh T., "Lung Cancer Prediction using Machine Learning Techniques", International Journal of Health Sciences 2022, DOI : 10.53730/ ijhs.v6ns2.8306

10. V.A. Makarov, D.R. Kaidarova, S.E.Yessentayeva, J.Kalmatayeva, M.E.Mansurova, N.Kadyrbek, R.E.Kadyrbayeva, S.T. Olzhayer, I.I.Novikov, "Using Machine Learning Algorithms to develop a model for predicting the survival of Lung Cancer Patients in the Republic of Kazakhstan", Oncojournal 2022,65-3,Pages 4-11, DOI:10.52532/2663-4864-2022-3-65-4-11
11. R Patra, "Prediction of Lung Cancer using Machine Learning Classifier", https://link.springer.com/chapter/10.1007/978-981-15-6648-6_11, International Conference on Computer Science, Communication and Security,July 2020,PP-132-142
12. Y.Gultepe, "Performance of Lung Cancer Prediction Methods using different Classification Algorithms", Computer,Materials and Continua, Feb' 2021, Vol.67, No.2,PP: 2015-2025
13. Yunpeng Cui,Xuedong Shi, Shengjie Wang, Yong Qin, Bailin Wang, Xiaotong Che, Mingxing Lei, "Machine Learning approaches for prediction of early death among Lung Cancer patients with bone metastases using routine clinical characteristics : An analysis of 19,887 patients", Front Public Health Oct' 2022, DOI : 10.3389/fpubh.2022.1019168
14. Dakhaz Mustafa Abdullah , Adnan Mohsin Abdulazeez, Amira Bibo Sallow, "Lung Cancer Prediction and Classification based on Correlation Selection method using machine learning technique", Qubahan Academic Journal, May 2021,1(2),PP 141-149,DOI:10.48161/qaj.v1n2a58
15. Dr.M.Kasthuri,M.Riyana Jency "Improving the performance of Lung Cancer Prediction using Machine Learning Techniques on Big Data", International Journal of Computer Science and Mobile Computing, October 2020,9(10),Pg:64-72,DOI : 10.47760/IJCSMC.2020.v09i10.008