A Survey on Lung Cancer Prediction

B. Jyothi[1] and L. Mary Gladence[2]

Abstract

Cancerous tissue that develops in lung is called a lung cancer. It's the primary reason people die from cancer, everywhere, with smoking being the primary cause of the disease. Lung cancer prediction involves using various methods and technologies is to identify the individuals who are at higher risk of developing the disease. The goal of lung cancer prediction is to detect the disease early, when it is most treatable, or to prevent it from developing altogether. Machine learning may be used to forecast the risk that someone will acquire lung cancer by analysing vast volumes of data to find patterns. In this study, a comparison of various ML-based algorithms for detecting lung cancer has been published. The techniques were used to look for cancer. There has been an explosion in the number of techniques available in recent years for diagnosing lung cancer. The majority of these techniques make use of CT scan pictures, while some make use of x-ray images. In addition, many different classifier strategies are combined with a wide variety of segmentation techniques in order to employ image recognition for the purpose of locating lung cancer nodules. According to the findings of this research, multi-gene genetic programming is superior to other approaches in terms of producing reliable outcomes.

Keywords: *Machine Learning, Early Detection, Lung Cancer Prediction, Classifiers, Multi Gene Genetic Programming (MGGP).*

Introduction

Lung cancer is a type of cancer that develops in the lungs and is one of the most common and deadly forms of cancer worldwide. The use of prediction algorithms to identify people who have a high chance of acquiring lung cancer has gained popularity in recent years. Predictive models may help identify people who need screening or more inquiry, and early diagnosis and intervention are essential for improving patient outcomes [1].

Lung cancer is a significant public health issue globally and it marks among the top five leading causes of deaths due to cancer. The World Health Organization reports that lung cancer accounts for 11.6% of all cancer cases and 18.4% of cancer-related deaths. The year 2020 was projected to see 2.2 million new cases of lung cancer worldwide, resulting in 1.8 million deaths. The incidence of lung cancer varies widely across different regions, with Europe and North America having the highest rates while Africa and Asia have the lowest. This discrepancy is largely attributed to differences in smoking pervasiveness, which is a crucial imminence for lung cancer.

Due to the adoption of smoking cessation programmes and greater knowledge of the dangers of smoking, the incidence of lung cancer has gradually decreased in many industrialised nations. However, lung cancer remains a significant health challenge, particularly in developing countries where smoking rates continue to increase [3]. The global burden of lung cancer highlights the need for continued efforts to develop effective

---

[1] Research Scholar Department of CSE, Sathyabama Institute of Science and Technology, Chennai.

[2] Professor, School of Computing, Sathyabama Institute of Science and Technology, Chennai.

prevention and early detection strategies, as well as innovative treatment approaches to improve patient outcomes.

Predictive models for lung cancer utilize various types of data, including clinical, genetic, and imaging data, in order to form models which can predict accurately the likelihood of developing lung cancer. Before symptoms appear, these models may be able to identify individuals at a high risk for lung cancer, resulting in early intervention and higher survival rates [4]. In this context, the development of accurate and reliable predictive models for lung cancer holds significant promise for improving patient care and reducing the burden of this disease. With the increasing availability of medical data and advances in machine learning, the potential for developing accurate predictive models for lung cancer is rapidly increasing.

Lung cancer prediction using machine learning models has gained significant popularity in recent years. Machine learning algorithms are used to identify patterns and relationships within large datasets, allowing for the development of accurate predictive models. Several machine learning algorithms, including logistic regression, decision trees, and support vector machines, have been used for lung cancer prediction [5]. These algorithms utilize various clinical and imaging data, such as age, smoking history, tumor size, and tumor location, to develop predictive models that can accurately identify individuals at high risk of developing lung cancer.

Deep learning models, a type of machine learning, are also being applied to the prediction of lung cancer. These models utilize artificial neural networks to learn and identify patterns in large datasets. Convolutional neural networks (CNNs), is a type of DL model frequently used for image classification tasks, have been used to analyze medical images, such as CT scans, to detect lung cancer [6].

The use of ML and DL models for prediction of lung cancer has shown promising results, with some studies reporting high accuracy rates. These models may increase the detection precision of lung cancer and enable focused screening and treatments. However, there are also challenges associated with the development and implementation of these models, such as the need for large, diverse datasets and the potential for bias in the data and algorithms. More research is needed to figure out how to deal with these problems and prove that machine learning and DL models can be utilised in clinical practise to predict lung cancer. In this paper, the state of art lung cancel prediction models are reviewed and analysed. In detail explanation of various prediction models are discussed in the literature section. The rest of the paper organized as follows: state of art prediction models are discussed in section-II. The MGGP based lung cancer prediction models are explained in section-III. In section-IV, conclusion of the paper is explained.

Related Works

There is a significant body of literature on lung cancer prediction, including research on the identification of risk factors, the development of predictive models, and the use of biomarkers for early detection. Some key studies and papers on lung cancer prediction include:

Lung Cancer Prediction Using Machine Learning

In order to figure out benign tumor while still maintaining a high lung cancer sensitivity, The CNN for Lung Cancer Prediction was developed by Marjolein A. Heuvelmans et al. [7]. In a European multicenter experiment, an unbiased dataset of ambiguous nodules was used as the input for the network, which had previously been trained using US screening data. The LCP-CNN was validated using CT images comprising 2106 nodules (205 lung malignancies) as part of the Early Lung Cancer Diagnosis.

Researchers Dakhaz Mustafa Abdullah et al. [8] examined the accuracy ratios of three different early-stage lung cancer classifiers are needed in order to save as many people's

lives as possible. These classifiers are the K Nearest Neighbor (KNN), CNN and SVM. The majority of the informative indexes that were employed as a part of this inquiry were gathered from databases that were kept by UCI and were related to people who were diagnosed with lung cancer. The execution of these investigations is the major emphasis of this work, which has the primary objective of conducting an investigation into accuracy of classification algorithms utilising the WEKA Tool.

In an urge to improve the efficiency of the process of identifying lung cancer from the dataset, C. Anil Kumar et.al [9] built a ML model that was based on SVMs. This has been done in order to make the process of detection as efficient as possible. A SVM classifier is utilised to categorise lung cancer patients according to the symptoms that they display, and concurrently, the Python is used in order to further the model's development. This is done in order to classify patients who have lung cancer. Based on a variety of different metrics, the effectiveness of this SVM model was assessed, and overall, the findings were encouraging. For the purpose of evaluating the model, many cancer datasets sourced from the library at the University of California, Irvine, were used. If the results of this study are confirmed to be positive, then smart cities will be able to provide its residents with superior medical care as a consequence of the findings of this study. Individuals suffering from lung cancer are now able to get real-time therapy at a reduced cost and with much reduced effort and latency, and they may do this from any place and at any time. The SVM and SMOTE methods that were already in use were contrasted with this model. As compared to the other approaches already in use, this method achieves an accuracy rate of 98.8 percent.

Mingzhou Liu et.al [10], proposed a Context Attention Network (CA-Net) that captures both nodule's properties and contextual features and then efficiently merges them throughout the process of classifying malignancy or benignancy. Authors use the characteristics of the nodule as a reference in order to precisely identify the surrounding features that include structures that have been warped or connected by the nodule. This is done using an attention process. In addition to this, Authors offer a feature fusion module that is able to dynamically modify the weights of a nodule's features as well as contextual features across nodules.

The objective of the study article that was recently published by B R Manju et.al [11] is to address the need for an interactive learning framework that makes it possible to effectively screen for abnormality in a patient. An outstanding method that significantly classifies the target groups is the principal component analysis, which stands out as a leader in its field. PCA is able to effectively combine connected attributes and provide a dispersed presentation of its components. This is made possible by PCA's ability to effectively combine properties. Clarity on the amount of important components that need to be maintained is provided by an inspection of the scree plot. Support Vector Machines, is a classification method that is both rapid and reliable, and it outperforms other approaches even when dealing with a little number of data. Support Vector Machines may also be referred to as SVM for short. The newly identified components are going to be entered into a SVM in order to further categorise them once they have been analysed. At the stage before it becomes hazardous, the clinical professionals will be urged to pay extra attention to the patients who are suffering this condition. While attempting to arrive at an accurate evaluation of the predicted ability, the confusion matrix is used. When it comes to the early identification of different stages of cancer, the model that was developed offers an extremely high degree of accuracy and leaves no room for mistake.

S. Shanthi et.al [12] proposed a novel technique for feature selection that is wrapper-based, and she does this by using a transformed version of the stochastic diffusion search (SDS) algorithm. For choosing the best feature subsets for the SDS, direct connection between the agents is going to be quite helpful. Among the several approaches to classification that have been tried and tested are the CNN, Naive Bayes (NB) algorithm, and the decision tree (DT). The experiment's findings demonstrate that the suggested strategy can perform at higher

levels than other methods already in use. These other methods include correlation-based feature selection and minimum redundancy maximum relevance.

In a small cohort of 41 healthy controls and 40 patients with lung cancer, the authors of the study that was published as research [13] investigated the possibility of using CNV from cfDNA as a biomarker for the detection of lung cancer. The cohort was tested on a modest scale. For the objective of making a diagnosis of cancer, arm-level CNV distributions were investigated with the use of XGBoost. Both the specificity and sensitivity levels in the test group were able to be raised up to 100% thanks to the adoption of a trained XGBoost classifier. In addition, a 5-fold cross-validation of the XGBoost model was carried out so that its consistency could be demonstrated.

In the paper [14] the authors categorise the data on lung cancer that is accessible in the UCI machine learning repository as benign or malignant based on their findings from an analysis of the different machine learning classifiers methodologies. After the pre-processing and conversion of the input data into binary form, the data set is then classified as either malignant or not cancerous using a well-known classifier approach that is implemented inside the Weka tool.

The image quality of lung cancer is improved throughout the research paper referred to as [15] by utilising and employing a variety of different methods of image enhancement. These methods include histogram equalisation methods, gamma correction, thresholding, picture adjustment and contrast stretching. The next paragraphs provide a more in-depth explanation of each of these approaches. The authors utilized and optimised robust Machine Learning classification techniques, such as the Decision Tree, Naïve Bayesian, SVM with Gaussian, RBF, and polynomial, and retrieved GLCM features from advanced photos. SVM, polynomial, and RBF methods all exhibits better results with 99.89% accuracy when image enhancement was not applied to the process. This analysis done on image dataset consists of 945 images that were provided by the LCA-MRI dataset. These techniques improved the predictive accuracy of our analysis. When this was taken into account, the results obtained with SVM, RBF, and polynomial kernels were 100% accurate and yielded an AUC of 1.00. The prediction performance was significantly improved through the use of image correction and contrast extending at cutoff point (0.02, 0.98). According to the results, the method that was offered has the potential to be of great aid in boosting the lung cancer prediction for the sake of next level diagnosis by experienced radiologists for mortality rate reduction. This would be done in order to lessen the number of people who pass away from the disease.

Sharmila Nageswaran's [16] study work demonstrates an accurate prediction of lung cancer by using technology that is made possible by ML and image processing. For the purpose of experimental inquiry, the dataset consisted of 83 CT scan images taken from around 70 various patients. During the phase of image pre-processing known as the geometric mean filter is used. The end result is an enhancement in the image's overall quality. Then after, the means approach is utilized to the photos in order to segment them. This segmentation may be used to locate the portion of the picture in question. After this step, machine learning-based categorization approaches are used. Various ML strategies, including ANN, KNN, and RF, were used throughout the categorization process. It has been discovered that the models of ANN yields higher accurate findings in prediction of lung cancer.

These studies and others demonstrate the potential for machine learning techniques to improve lung cancer prediction and risk assessment, and highlight the importance of leveraging big data and advanced analytics to improve cancer outcomes.

Lung Cancer Predication Using Multi Gene Genetic Programming

Multi Gene Genetic Programming (MGGP) is a type of genetic programming algorithm that uses multiple genes instead of a single gene to encode a solution. In traditional genetic programming, a single string of genetic code represents a solution, which can be optimized

through a process of mutation and recombination. MGGP extends this concept by allowing multiple genetic codes to work together to form a solution. Each gene in MGGP represents a distinct sub-solution, and the combination of genes produces a more complex solution. The algorithm optimizes the genes and their interaction through a process of selection, crossover, and mutation [17].

MGGP has been applied to various problems, including feature selection, classification, regression, and optimization. For example, MGGP is widely used to develop predictive models for medical diagnosis, energy consumption prediction, and financial forecasting. MGGP has also been used for feature selection in image classification, natural language processing, and bioinformatics. Compared to traditional genetic programming, MGGP offers several benefits, including improved search efficiency, better scalability, and increased diversity in the solution space. However, MGGP also requires careful design and tuning of parameters, as well as specialized techniques for interpreting the evolved solutions.

There are few studies specifically using MGGP for lung cancer prediction by selecting automatic features. However, there have been several studies that have used genetic programming algorithms for feature selection in cancer prediction and classification tasks. One example is a study published in the journal BMC Bioinformatics in 2019 [18], which used a genetic programming algorithm to select features from genomic data for breast cancer classification. The authors used a variant of the genetic programming algorithm called Gene Expression Programming (GEP) to select a set of features from genomic data that were most predictive of breast cancer subtype. The resulting model achieved high accuracy in predicting breast cancer subtypes.

Another study published in the journal Artificial Intelligence in Medicine in 2017 [19] used a genetic programming algorithm to select features from gene expression data for bladder cancer classification. The authors used a variant of the genetic programming algorithm called Multi Expression Programming (MEP) to select a set of features that were most predictive of bladder cancer. The resulting model achieved high accuracy in distinguishing between cancerous and non-cancerous bladder tissue. These studies demonstrate the potential of genetic programming algorithms for feature selection in cancer prediction tasks. While these studies did not focus specifically on amino acid sequences, the use of genetic programming algorithms for feature selection in other types of genomic data suggests that the approach could be applicable to amino acid sequences as well. Future research may explore the use of genetic programming algorithms, including MGGP, for feature selection in amino acid sequences for lung cancer prediction.

The authors conducted an analysis of altered genes associated to lung cancer because these genes provide important information in protein amino acid sequences [20] Their goal was to identify beneficial new patterns that might easily detect early stages of cancer. For this purpose, the author created a novel evolutionary learning approach that makes use of the discriminant information of protein amino acids. This technique utilises a biologically inspired multi-gene genetic programming algorithm. This model efficiently identifies 23 discriminant characteristics out of 1500 features. After that, it optimises the combination of the chosen characteristics and the corresponding basic functions in order to make a prediction about lung cancer. As a result, a powerful prediction model has been developed, which contributes to an improved understanding of the complicated and diverse nature of lung cancer.

Barath Narayanan et al. [21] presented a unique cluster-based classifier model for use in lung nodule detection systems. The authors presented a unique and enhanced model of feature selection that was adaptable which could be utilised for either the cluster or the classifier aspects of the study. The authors make use of a separate database in order to conduct training, and this database includes a total of 173 nodules spread across 160 different cases. CRs are processed using this database. The evaluations are carried out with

the assistance of a database that can be accessed by anybody and was conceived of by the JRST. The JRST database is comprised of 154 CRs, each of which contains a single nodule that has been confirmed by a radiologist. Because some of the cases in this research had lung swelling in the retro-cardiac and sub-diaphragmatic areas of the lungs, the authors of this study decided not to include those instances in the JRST database. The research was based on the execution of triple cross validation on 107 cases drawn from open source dataset which consisted of 280 nodules and developed by the Lung Imaging Database Collaboration.

In order to determine which genes are necessary, Lei Chen [22] looked into the important and non-important genes mentioned in prior research which retrieved gene-ontology (GO) concepts. Authors determined each important/non-important gene into a vector using the enhancement theory of GO and KEGG pathways. The MR-MR method was used to assess these associations. Subsequently, significant KEGG pathways and GO keywords were extracted using the Incremental-Feature and SVM methods. With a Matthews correlation value of 0.951, the prediction model developed utilising the retrieved GO keywords and KEGG pathways performed almost flawlessly in identifying important and non-important genes. The twentyone most significant GO keywords and 3 KEGG pathways were thoroughly examined to completely study the major variables impacting the basic functions of essential genes. Also, a number of genes that our prediction model identified to be important genes were supplied in this research. We propose that our work offers additional functional and pathway details on the crucial genes and offers a fresh approach to researching related issues [23-24].

In summary, MGGP is a machine learning technique can be used for predicting lung cancer. It involves identifying a set of genetic markers that are associated with the disease, and using these markers to build a model which can predict the possibilities of a person developing lung cancer. The process of multi-gene genetic programming starts with collecting a large amount of genetic data from individuals with and without lung cancer. The data is then analyzed to identify genetic markers that are differentially expressed between the two groups. These markers are then used to build a model that can predict the likelihood of an individual developing lung cancer. The model is typically trained using a ML algorithm such as a s SVM or a decision tree [25]. Once the model is trained, it can be used for prediction of the likelihood of an individual developing lung cancer based on their genetic data. MGGP has shown promising results in predicting lung cancer, with studies reporting high levels of accuracy and sensitivity. However, further research is needed to validate the technique and optimize its performance [26].

Conclusion

When lung cancer is detected at a very early stage, it is to the patient's benefit since treatment may then be initiated to prevent the sickness from having a detrimental influence on the patient's life and prolong the patient's life expectancy. As a result of this, the purpose of this study is to present a summary of a extensive evaluation of a number of machine learning algorithms that may recognise lung tumours by using either CT scan photographs or X-ray images as input. Researchers have used a broad range of classifiers in their published studies, including MLP, SVM, Naive Bayes, Neural Networks, Gradient Boosted Trees, Decision Trees, K-Nearest Neighbors, Multinomial Random Forest Classifiers, Naive Bayes, Stochastic Gradient Descent, and Ensemble Classifier. It is feasible to draw the conclusion, as a result of this, and on the basis of the extensive survey that has been carried out as a part of this study, that the approaches which used multi-gene genetic programming techniques got greater results in terms of accuracy than other typical machine learning techniques.

References

Oudkerk, Matthijs, ShiYuan Liu, Marjolein A. Heuvelmans, Joan E. Walter, and John K. Field. "Lung cancer LDCT screening and mortality reduction—evidence, pitfalls and future perspectives." *Nature reviews Clinical oncology* 18, no. 3 (2021): 135-151.

Ferlay, Jacques, Murielle Colombet, Isabelle Soerjomataram, Donald M. Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. "Cancer statistics for the year 2020: An overview." International journal of cancer 149, no. 4 (2021): 778-789.

Bade, Brett C., and Charles S. Dela Cruz. "Lung cancer 2020: epidemiology, etiology, and prevention." Clinics in chest medicine 41, no. 1 (2020): 1-24.

Tu, Yuqin, Yan Wu, Yunfeng Lu, Xiaoyun Bi, and Te Chen. "Development of risk prediction models for lung cancer based on tumor markers and radiological signs." Journal of clinical laboratory analysis 35, no. 3 (2021): e23682.

Tuncal, Kubra, Boran Sekeroglu, and Cagri Ozkan. "Lung cancer incidence prediction using machine learning algorithms." Journal of Advances in Information Technology Vol 11, no. 2 (2020).

Subramanian, R. Raja, R. Nikhil Mourya, V. Prudhvi Teja Reddy, B. Narendra Reddy, and Srikar Amara. "Lung Cancer Prediction Using Deep Learning Framework." International Journal of Control and Automation 13, no. 3 (2020): 154-160.

Heuvelmans, Marjolein A., Peter MA van Ooijen, Sarim Ather, Carlos Francisco Silva, Daiwei Han, Claus Peter Heussel, William Hickes et al. "Lung cancer prediction by Deep Learning to identify benign lung nodules." Lung Cancer 154 (2021): 1-4.

Abdullah, Dakhaz Mustafa, Adnan Mohsin Abdulazeez, and Amira Bibo Sallow. "Lung cancer prediction and classification based on correlation selection method using machine learning techniques." Qubahan Academic Journal 1, no. 2 (2021): 141-149.

Anil Kumar, C., S. Harish, Prabha Ravi, Murthy Svn, B. P. Kumar, V. Mohanavel, Nouf M. Alyami, S. Shanmuga Priya, and Amare Kebede Asfaw. "Lung cancer prediction from text datasets using machine learning." BioMed Research International 2022 (2022).

Liu, Mingzhou, Fandong Zhang, Xinwei Sun, Yizhou Yu, and Yizhou Wang. "Ca-net: Leveraging contextual features for lung cancer prediction." In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, pp. 23-32. Springer International Publishing, 2021.

Manju, B. R., V. Athira, and Athul Rajendran. "Efficient multi-level lung cancer prediction model using support vector machine classifier." In IOP Conference Series: Materials Science and Engineering, vol. 1012, no. 1, p. 012034. IOP Publishing, 2021.

Shanthi, S., and N. Rajkumar. "Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods." Neural Processing Letters 53 (2021): 2617-2630.

Yu, Daping, Zhidong Liu, Chongyu Su, Yi Han, XinChun Duan, Rui Zhang, Xiaoshuang Liu, Yang Yang, and Shaofa Xu. "Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier." Thoracic cancer 11, no. 1 (2020): 95-102.

Patra, Radhanath. "Prediction of lung cancer using machine learning classifier." In Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1, pp. 132-142. Springer Singapore, 2020.

Hussain, Lal, Hadeel Alsolai, Siwar Ben Haj Hassine, Mohamed K. Nour, Mesfer Al Duhayyim, Anwer Mustafa Hilal, Ahmed S. Salama, Abdelwahed Motwakel, Ishfaq Yaseen, and Mohammed Rizwanullah. "Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features." Applied Sciences 12, no. 13 (2022): 6517.

Nageswaran, Sharmila, G. Arunkumar, Anil Kumar Bisht, Shivlal Mewada, J. N. V. R. Kumar, Malik Jawarneh, and Evans Asenso. "Lung cancer classification and prediction using machine learning and image processing." BioMed Research International 2022 (2022).

Niazkar, Majid, and Hamid Reza Niazkar. "Covid-19 outbreak: application of multi-gene genetic programming to country-based prediction models." Electronic Journal of General Medicine 17, no. 5 (2020).

Granata, Ilaria, Enrico Troiano, Mara Sangiovanni, and Mario Rosario Guarracino. "Integration of transcriptomic data in a genome-scale metabolic model to investigate the link between obesity and breast cancer." BMC bioinformatics 20, no. 4 (2019): 1-11.

Hamet, Pavel, and Johanne Tremblay. "Artificial intelligence in medicine." Metabolism 69 (2017): S36-S40.

Sattar, Mohsin, Abdul Majid, Nabeela Kausar, Muhammad Bilal, and Muhammad Kashif. "Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences." Computational Biology And Chemistry 98 (2022): 107638.

Narayanan, Barath Narayanan, Russell C. Hardie, Temesguen M. Kebede, and Matthew J. Sprague. "Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities." Pattern Analysis and Applications 22 (2019): 559-571.

Chen, Lei, Yu-Hang Zhang, ShaoPeng Wang, YunHua Zhang, Tao Huang, and Yu-Dong Cai. "Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways." PloS one 12, no. 9 (2017): e0184129.

Venkateswara Rao, K., Mary Gladence, L., Raja Lakshmi, V., "Research of feature selection methods to predict breast cancer", in International Journal of Recent Technology and Engineering(IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.

L.Mary Gladence,T.Ravi,M.Karthi, "Heart Disease Prediction using Naïve Bayes Classifier-Sequential Pattern Mining" in the International Journal of Applied Engineering Research ISSN 0973-4562 Volume 9, Number 21 (2014) pp. 8593-8602

Rao, K. Venkateswara, and L. Mary Gladence. "Support vector machine based disease classification model employing hasten eagle Cuculidae search optimization." Concurrency and Computation: Practice and Experience (2022): e7259.

Rajan, Juliet Rani, and V. Maria Anu. "Application of data mining on cloud for lung cancer prognosis." International Journal of Cloud Computing 11, no. 5-6 (2022): 449-455.