IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Strategies And Algorithms In Data Mining For Lung Cancer Deduction

[1]**N.Malathi,**[2]**Dr.A.Prakash,**
[1]Research Scholar, [2] Professor,
[1, 2]Department of Computer Science,
[1, 2]Hindusthan College of Arts & Science,
[1, 2] Coimbatore, Tamil Nadu, India.

Abstract

This paper utilizes data mining techniques for the deduction of lung cancer, a critical step in early diagnosis and treatment planning. Leveraging advanced algorithms, including decision trees and clustering methods, significant factors influencing lung cancer occurrence are identified from a comprehensive dataset. By analyzing patient demographics, lifestyle factors, and medical history, predictive models are developed to accurately classify individuals at risk. The study aims to enhance early detection efforts, potentially reducing mortality rates and healthcare burdens associated with lung cancer. Overall, the research contributes to the advancement of preventive healthcare strategies through effective data analysis and mining techniques.

Keywords: Lung Cancer, Data Mining, Early Detection, Predictive Modeling, Healthcare.

1. Introduction

Lung cancer is a devastating disease with significant implications for public health worldwide. It is one of the leading causes of cancer-related deaths globally, posing substantial challenges to healthcare systems and individuals alike. Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. However, the complexity of lung cancer, combined with the multitude of factors influencing its development and progression, makes diagnosis and prognosis challenging. In recent years, the emergence of data mining techniques has provided promising avenues for addressing these challenges by leveraging the vast amounts of available healthcare data to extract valuable insights. This introduction aims to explore the application of data mining techniques in the deduction of lung cancer, shedding light on their potential to enhance early detection, prognosis, and treatment planning.

Data mining techniques encompass a diverse set of computational methods and algorithms designed to extract patterns, trends, and knowledge from large datasets. These techniques have been increasingly utilized in healthcare settings to analyze clinical data, identify risk factors, predict disease outcomes, and inform decision-making processes. In the context of lung cancer, data mining holds immense potential for improving our understanding of the disease and its associated factors. The deduction of lung cancer using data mining techniques involves the analysis of various types of data, including clinical records, medical imaging, genomic data, and environmental factors. By integrating these diverse data sources, data mining algorithms can identify complex relationships and patterns that may not be apparent through traditional analysis methods. For example, machine learning algorithms such as decision trees, support vector machines, and neural networks can be trained on large datasets to classify patients into different risk groups based on their likelihood of developing lung cancer.

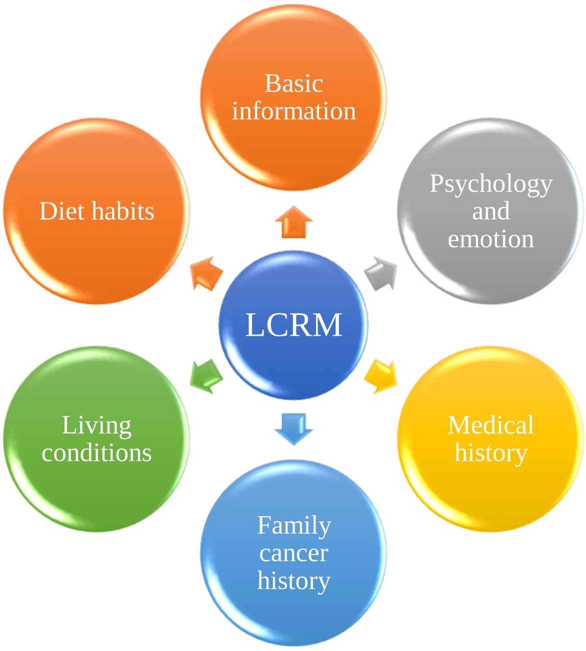

Figure 1. Lung Cancer Risk Model (LCRM)

One of the primary objectives of using data mining techniques in lung cancer deduction is to enhance early detection and diagnosis. By analyzing patient data and identifying specific risk factors or biomarkers associated with lung cancer, data mining algorithms can help healthcare providers identify individuals at higher risk and recommend targeted screening or diagnostic tests. Early detection is critical for improving patient outcomes, as it enables timely intervention and treatment initiation, potentially leading to better survival rates and quality of life for patients.

2. Literature Survey

1. Thamilselvan P (2022) et.al proposed Lung cancer prediction and classification using Adaboost data mining algorithm[1]. Adaptive Boosting, or AdaBoost, is a potent AI ensemble method that serves as a meta-algorithm classifier. It trains several weak classifiers iteratively, giving each sample an equal initial weight. In order to improve the identification of the incorrectly categorized samples in later iterations, sample weights are modified after every training round based on the sample error rate. This iterative procedure yields a group of k weak learners, which are then combined in a weighted way to create a strong learner. The high incidence of cancer, especially lung cancer, emphasizes how important early identification and intervention strategies are. The likelihood of a successful course of therapy and survival are greatly increased with early detection, hence developing prediction tools is critical. The suggested Adaboost algorithm shows impressive predictive power from CT scans for lung cancer, particularly in differentiating between small- and non-small-scale cancer cells. More than 100 CT images, both cancer and non-cancer are used to assess the algorithm's performance, demonstrating its accuracy and effectiveness. With a minimum misclassification rate of 1.54%, the Adaboost method's classification accuracy of 98.46% is achieved by a rigorous analysis of true positives, true negatives, false positives, and false negatives. The algorithm's potential goes beyond lung cancer detection to other medical applications including brain tumor detection and breast cancer identification, even though processing time is not discussed in this paper.

Merits

1. Effectively distinguishes between benign and malignant cells, aiding in accurate diagnosis and early intervention.

Demerits

1. Lack of consideration for processing time may hinder real-time application in clinical settings.

2. W. Abdul (2020) et.al proposed An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network[2]. Although it is still difficult, early identification of lung cancer is essential for saving lives. The precision of current systems, which rely on hand-engineered processes, is limited. An Automatic Lung Cancer Detection and Classification (ALCDC) system based on Convolutional Neural Networks (CNN) is introduced, leveraging the success of deep learning in numerous identification tasks. Using CT scan images, this approach attempts to detect and categorize lung cancers as benign or malignant. When tested using data from the Image Database Resource Initiative (IDRI) and the Lung Image Database Consortium (LIDC), the suggested ALCDC system outperforms state-of-the-art methods with an astounding accuracy of 97.2%. In order to prevent overfitting, the CNN design consists of convolutional and pooling layers followed by dropout layers.The network efficiently extracts features utilizing ReLU activation functions, 32 and 16 filters in the first convolutional layers, and a pooling layer with a kernel size of 2. The classification is further improved by a fully connected layer that has 16 neurons and ReLU activation. Stochastic gradient descent and cross-entropy loss are used in the system's training to reduce the discrepancy between expected and desired outputs. The CNN architecture determines if nodules are benign or cancerous by analyzing parameters including learning rate, epoch counts, and filter sizes. With 97.2% accuracy, 95.6% sensitivity, and 96.1% specificity, this deep learning solution performs exceptionally well, highlighting its potential to improve medical diagnosis research and healthcare systems. Such technologies' ability to enable early diagnosis shows promise for enhancing patient outcomes and lung cancer prognoses outcomes.

Merits

1. Utilizes deep learning, specifically CNN architecture, offering improved performance compared to traditional hand-engineered techniques.

Demerits

1. Overfitting risk addressed with dropout layers, but further optimization may be needed to enhance robustness.

3.Maleki N (2021) et.al proposeda k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection[3]. This research proposes a k-Nearest-Neighbors (kNN) method for patient stage diagnosis of lung cancer, enhanced by a genetic algorithm (GA) for effective feature selection. The kNN method's accuracy is increased when the GA optimizes feature selection to decrease dataset dimensions and speed up the classifier. The ideal value for k is ascertained by means of experimental approaches, yielding 100% accuracy when applied to a database of lung cancer cases. This indicates the possibility of effectively diagnosing lung cancer staging through the correlation of clinical data with data mining techniques. The GA iteratively improves solutions represented as chromosomes through heuristic search techniques inspired by natural selection. The process consists of five stages: the creation of the original population, fitness assessment, selection, crossover, and mutation. The genetic algorithm selects chromosomes based on fitness and performs genetic operations on them to create diversity and prevent premature convergence. This process is repeated generations later. The convergence of the cost function and the identification of the ideal features and k value demonstrate the significant improvement in classification accuracy that results from hybridizing the kNN approach with GA-based feature selection. Standard hardware was used for computational research, and real clinical data was used to validate the conclusions by a lung cancer specialist. To further increase performance and resilience, future studies could investigate different feature selection strategies and classification algorithms.

Merits

1. Hybridization increases classification accuracy, aiding in patient risk assessment for effective treatment.

Demerits

1. Genetic Algorithm's stochastic nature may not always find optimal solutions.

4. Vinmalar FL (2019) et.al proposedGenetic Algorithm (GA) for prediction of lung cancer using data mining techniques[4].Numerous lives are lost due to cancer, a dangerous and widespread disease that highlights the importance of early detection for a better prognosis. This research explores the field of data mining tools for early lung cancer prediction. The lungs are important but delicate organs that are frequently destroyed by cancer's unrelenting attack. Lung cancer is mostly caused by smoking, occupational smoke exposure, air pollution, industrial toxins, and genetic predispositions. In order to forecast numerous models, our work suggests using a dataset classification strategy based on genetic algorithms (GAs), which authors show to be superior to both Differential Evolution and Particle Swarm Optimization. By introducing a new weighted average ensemble strategy driven by GA, we improve forecast accuracy over traditional approaches. Using distinct data partitioning for improved precision, authors carefully test eight machine learning models to determine which ones perform best and then fine-tune them. A thorough data analysis and graphical representations show that the integration of GA-based weighted averaging produces promising results, outperforming conventional techniques in accuracy and performance criteria.

Merits

1. GA-based ensemble method improves performance over classical weighted average approaches.

Demerits

1. Interpretability of GA-generated models might be challenging for medical practitioners.

5. Pontes B (2021) et.al proposed a data mining based clinical decision support system for survival in lung cancer[5]. To improve clinical recommendations for predicting lung cancer patients' overall survival, a clinical decision support system (also known as a CDSS) has been developed. Using sophisticated algorithms like XGBoost and Generalized Linear Models, the CDSS examined 1167 factors by using data from 543 patients between 2013 and 2017. The AUCs of the CDSS exceeded 0.90, suggesting that it performed better than guideline-based predictions overall and especially well in small cell lung cancer cases. Conversely, AUCs for predictions based on guidelines were frequently less than 0.70. The CDSS showed statistically significant improvements; in some cases, the AUC increased from 0.60 to 0.84 ($p = 0.0009$).Specifically, the CDSS showed notable improvements in small cell lung cancer patient survival prediction, as well as for patients with prolonged follow-up (≥ 18 months). This highlights the ability of the CDSS to offer individualized management guidance and to support evidence-based prognostic talks among patients with lung cancer.

Merits

1. Assists in formulating evidence-based management advice, facilitating individualized discussions according to prognosis.

Demerits

1. Potential biases or errors in data mining algorithms could affect accuracy and reliability of predictions.

6. J. Taveira De Souza (2019) et.al proposed Dimensionality Reduction in Gene Expression Data Sets[6].In microarray data analysis for lung cancer pathologic staging, dimensionality reduction techniques are crucial for improving prediction accuracy, reducing computational burden, and constructing robust models. This paper conducts a comprehensive comparison between two such techniques: attribute selection and principal component analysis (PCA), focusing on gene expression datasets. Both methods are implemented during pre-processing and rigorously evaluated experimentally. Additionally, a novel approach combining consistency-based subset evaluation (CSE) with minimum redundancy maximum relevance (mRMR), termed CSE-mRMR, is introduced to enhance classification efficiency. Results demonstrate a significant improvement in classifier hit rates when employing either reduction method compared to using all attributes. Notably, attribute selection consistently outperforms PCA across various classifiers and datasets, as validated by cross-validation techniques. Furthermore, CSE-mRMR exhibits strong classification performance across the datasets. Collectively, these findings suggest that attribute selection holds promise for analyzing and predicting gene expression datasets effectively. This research contributes to the growing body of literature advocating for the relevance of attribute selection in the analysis and future prediction of gene expression data sets, underscoring its potential significance in advancing lung cancer pathologic staging diagnosis.

Merits

1. Reduces computational complexity, improves visualization.

Demerits

1. May lose important information, requires careful selection of reduction techniques.

7.S. Senthil (2018) et.al proposed Predicting Lung Cancer Using Data mining Techniques with the AID of SVM Classifier[7]. The recommended techniques provide a reliable means to predict lung cancer classification, playing a crucial role in medical data diagnosis and classification. Various lung cancer diagnostic systems, supported by SVM, accurately predict both typical and abnormal lung cancers. Our research focuses on predicting lung cancer classification, distinguishing between typical and abnormal cases. Lung cancer remains a prevalent disease, responsible for a significant portion of cancer diagnoses and deaths. According to a 2003 report from the American Cancer Society, lung cancer accounted for 13% of all cancer diagnoses, with 28% of cancer-related deaths attributed to it. The survival rate for lung cancer diagnosed within five years is only 15%, increasing to 49% when the disease is localized and detected early. Diagnosis primarily relies on computerized tomography (CT) data. In this study, authors employ several data mining techniques, such as SVMs, for lung cancer identification and characterization, along with pre-processing and feature selection methods. Authors utilize characterization techniques to classify cases based on their characteristics, and introduce an optimal feature selection approach using firefly optimization. This comprehensive approach aims to enhance the accuracy and effectiveness of lung cancer classification and diagnosis.

Merits

1. As this structure will be available on the web, patients from remote spots can moreover benefit its benefits.

Demerits

1. The use of fine needle or other biopsy kinds of stuff to the influenced region is more challenging to the patient.

8.M. M. Abdelwahab (2017) et.al proposed four layers image representation for prediction of lung cancer genetic mutations based on 2DPCA[8]. This paper introduces a novel approach for early prediction of lung cancer somatic mutations and identification of substitution types. Lung cancer, a prevalent genetic disease, claims numerous lives annually due to gene damage triggered by various factors, including cigarette smoking. Oncogenes are activated and tumor suppressor genes are deactivated, leading to mutations and tumor formation in lung cells. To address this, a method integrating four-layer image representation and 2DPCA technique is proposed. Optimal image layer sizes representing gene symbolic sequences are investigated, with analysis revealing that high prediction accuracy can be achieved with fewer eigenvectors, reducing feature dimensionality and computational complexity. Experimental results demonstrate the algorithm's effectiveness, achieving a peak accuracy of 98.55% in predicting lung cancer somatic mutations and 88.18% in identifying gene substitution types, outperforming other methods. This approach not only aids in early diagnosis of lung cancer but also offers insights into the underlying genetic mechanisms, potentially informing targeted treatment strategies.

Merits

1. Consequently, the dimensionality of features and computational complexity were reduced and the prupossed algorithm achieved the highest accuracy.

Demerits

1. The computational time to predict and identify the mutations for one gene is approximately.

9. P. Nanda (2020) et.al proposed Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest[9]. Endurance rate prediction plays a crucial role in providing patients with valuable insights into the success of their treatment. However, determining the factors to consider for this prediction, especially in cases of lung cancer, presents challenges. This study employs various algorithms, including the proposed Improved Random Forest method, to classify the survival rate of Non-Small Cell Lung Cancer (NSCLC) patients. NSCLC, a prevalent form of lung cancer, is a leading cause of cancer-related deaths. Common treatments for NSCLC include Radiation Therapy and Chemotherapy. Staging, the process of determining the extent of cancer spread, is essential for treatment planning. The T-N-M staging system describes the cancer's size and location within the body. This paper introduces an innovative framework for predicting the survival rate of NSCLC patients using the Improved Random Forest method. Results demonstrate a significant improvement in prediction performance to 98% with the proposed approach. Comparative analysis against other methods such as SVC, Naive Bayes, Decision Tree, and Random Forest reveals the superiority of the proposed method. The findings underscore the effectiveness of the Improved Random Forest algorithm in predicting NSCLC patient survival rates, offering valuable insights for treatment planning and patient care.

Merits

1. This information assists the specialist with picking the best treatment choices.

Demerits

1. The prediction framework has lackluster showing when it thinks about just the staging information.

10. P. -W. Soh (2018) et.al proposed Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations[10]. Air pollution has emerged as a complex and pressing issue, with particulate matter, particularly PM2.5, posing significant health risks. The small size of PM2.5 allows it to penetrate deeply into the lungs, affecting gas exchange and potentially leading to respiratory and cardiovascular diseases, as well as an increased risk of lung cancer. Consequently, accurate monitoring of air quality has become essential for public health management. This study aims to forecast air quality up to 48 hours in advance by employing a combination of artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks to capture spatial-temporal relationships. The proposed predictive model integrates meteorological data from recent hours and elevation-related information to account for landscape effects on air quality. It considers trends across different regions, including neighboring and similar areas, both spatially and temporally. Experiments conducted using datasets from Taiwan and Beijing demonstrates that the proposed model exhibits outstanding performance, surpassing existing state-of-the-art systems. By leveraging advanced machine learning techniques and comprehensive data analysis, this research contributes to the advancement of air quality forecasting, thereby facilitating more effective management strategies to mitigate the impact of air pollution on public health.

Merits

1. CNN module consideration being more helpful for longer time span expectations, since CNN can separate the fleeting postpone factor from encompassing objective highlights by learning spatial data.

Demerits

1. Consider explicit Air box sensor source models as highlights to tune and mitigate noise because of machine contrasts.

11. Yu T, He Z (2015) et.al proposed C5.0 algorithm for Analysis of the factors influencing lung cancer hospitalization expenses using data mining[11]. Hospitalization expenses incurred during lung cancer therapy pose a significant economic burden on patients and are a focal point for medical insurance departments. Thus, there is a shared interest among hospitals and insurance institutions in accurately classifying and analyzing these expenses to predict reasonable medical costs. To address this, a C5.0 algorithm is utilized to analyze factors influencing hospitalization expenses among 731 lung cancer patients. The C5.0 algorithm, a data mining method for classification, reveals that variables such as length of stay (LOS), major therapy, and medication costs hold greater importance. These factors significantly impact the hospitalization costs of lung cancer patients. Remarkably, classification accuracy rates of training and testing partition sets exceed 84%. Additionally, upon incorporating cost variables, the accuracy rate surges beyond 95%. Notably, the derived classification rules align with actual clinical practice. Furthermore, the research's established model exhibits potential applicability beyond lung cancer, offering insights into disease hospitalization costs based on selected feature variables. Overall, the utilization of the C5.0 algorithm demonstrates its efficacy in accurately predicting and analyzing lung cancer hospitalization expenses, contributing to informed decision-making in healthcare resource allocation and cost management.

Merits

1. C5.0 algorithm offers high accuracy in analyzing factors impacting lung cancer hospitalization expenses, providing valuable insights for healthcare cost management.

Demerits

1. However, it may struggle with handling large datasets and requires careful parameter tuning for optimal performance.

12.M. V. Dass (2014) et.al proposed J48 algorithm with improved decision tree for Classification of lung cancer subtypes by data mining technique[12]. This study delves into the foundational role of gene mutations and their expressions in cancer progression, particularly focusing on lung cancer. The analysis involves examining genomic and proteomic datasets, encompassing biomarkers like microRNAs, genes, and proteins, for Non-Small Cell Lung Cancer (NSCLC) and its subtypes: Squamous Cell Cancer (SCC) and adenocarcinoma (ADC). A coordinated classification decision tree induction algorithm is applied to these biomarkers to predict cancer types accurately. The algorithm achieves high classification accuracy, bolstered by cross-validation techniques enhancing the J48 algorithm's performance. Additionally, top ten classification rules are derived using the apriori algorithm, refining the decision tree model. While the typical accuracy rate nears 99.7%, user interest-driven pruning is employed to streamline the rules. The resultant decision tree, customizable to user preferences, exhibits significant improvement over the baseline J48 algorithm. These findings serve as valuable guidelines for diagnosis and therapeutic interventions in SCC and ADC cancers. Leveraging biomarker data for precise lung cancer diagnosis could alleviate the burdens associated with histopathological assessments, offering more efficient patient management.

Merits

1. This classification methodology reveals an approach for differential diagnosis of cancers based on knowledge of deviated/disregulated biomolecules.

Demerits

1. The current trend of common therapy for both the cancer can’t be improved to cancer phenotype specific therapy.

13. T. Turki (2017) et.al proposed Boosted Transfer Learning Approach for Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients[13].Traditional machine learning methods for drug sensitivity prediction typically assume that training and test data must share the same feature space and distribution. However, this assumption often does not hold in real-world scenarios. For instance, in predicting drug sensitivity for multiple myeloma patients, there may be limited training data available. However, there might be ample auxiliary data for predicting drug sensitivity in patients with a different cancer type, where the auxiliary data exist in a distinct feature space or distribution. In such cases, leveraging transfer learning techniques could enhance prediction algorithm performance on the test data for the target task by leveraging auxiliary data from the related task. This study introduces two transfer learning approaches that integrate auxiliary data from the related task with the training data of the target task to enhance prediction performance on the test data for the target task. The performance of these transfer learning approaches is evaluated using three auxiliary datasets and compared against baseline methods using the area under the ROC curve (AUC) on the test data for the target task. Experimental findings demonstrate the effectiveness of the proposed approaches, highlighting their superiority over baseline methods when auxiliary data are incorporated. By leveraging transfer learning, these approaches offer promising avenues for improving drug sensitivity prediction in scenarios with limited training data availability.

Merits

1. Boosted transfer learning enhances drug sensitivity prediction in multiple myeloma patients improving model accuracy and generalization.

Demerits

1. However, it may require significant computational resources and extensive fine-tuning to optimize performance across datasets.

14.Krishnaiah V (2013) et.al proposed One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) for Diagnosis of lung cancer prediction system using data mining classification techniques[14]. Cancer, a leading cause of death globally, underscores the critical need for early detection to ensure successful treatment. Lung cancer, often misdiagnosed, requires prompt identification to prevent severe consequences. Timely diagnosis significantly impacts cure rates and patient prognosis. However, diagnostic errors remain prevalent in medical practice, highlighting the importance of leveraging knowledge discovery and data mining in healthcare. Classification-based techniques such as Rule-based systems, Decision Trees, Naïve Bayes, and Artificial Neural Networks hold promise for analyzing vast healthcare datasets. Yet, the healthcare sector often overlooks valuable insights hidden within these data. To address this, One Dependency Augmented Naïve Bayes (ODANB) and Naïve Creedal Classifier 2 (NCC2) enhance data pre-processing and decision-making processes. These classifiers extend traditional methods to handle imprecise probabilities and incomplete datasets, ensuring robust classifications. Uncovering hidden patterns and relationships facilitate accurate diagnosis, even for complex scenarios beyond the capabilities of conventional decision support systems. By incorporating common lung cancer symptoms such as age, sex, wheezing, and pain, predictive models aid in assessing disease likelihood. The paper's objective is to propose a model for early detection and accurate diagnosis of lung cancer, empowering healthcare professionals to save lives through timely intervention.

Merits

1. Boosted transfer learning enhances drug sensitivity prediction in multiple myeloma patients improving model accuracy and generalization.

Demerits

1. However, it may require significant computational resources and extensive fine-tuning to optimize performance across datasets.

15. Sowmiya T (2014) et.al proposed Reduced-Order Constrained Optimization (ROCO) for Optimization of lung cancer using modern data mining techniques[15]. In the modern era, cancer remains among the most prevalent and perilous diseases globally, with lung cancer emerging as one of its most formidable types. This malignancy spreads rapidly through uncontrolled cell growth in lung tissues, underscoring the critical importance of early detection for patient survival. This paper comprehensively surveys various data mining methodologies employed in lung cancer prediction, emphasizing their pivotal role in classification tasks. Specifically, authors delve into the utilization of ant colony optimization (ACO) techniques within data mining frameworks. ACO aids in refining disease prediction values, thereby facilitating more accurate prognoses. Through a case study, authors amalgamate data mining and ACO methodologies to generate precise rules and classifications for lung cancer diagnosis. Furthermore, this study lays the groundwork for enhancing medical diagnosis in lung cancer by proposing the use of the Reduced-Order Constrained Optimization (ROCO) method. ROCO streamlines the creation of clinically acceptable Intensity-Modulated Radiation Therapy (IMRT) plans for advanced lung cancer patients, offering a swift and automated solution. Our novel ROCO implementation interfaces seamlessly with treatment planning systems and employs dose constraints based on previous work, thereby advancing the field of lung cancer diagnosis and treatment planning.

Merits

1. ROCO optimizes lung cancer treatment efficiently with reduced computational complexity.

Demerits

1. It may overlook certain nuances in patient data, potentially impacting treatment efficacy.

Table 1.Proposed Methods, Merits and Demerits

Paper Title	Algorithm/Method	Description	Merits	Demerits
Lung cancer prediction and classification using Adaboost data mining algorithm	AdaBoost	Lung cancer prediction using ensemble method.	Effective in distinguishing between benign and malignant cells.	Lack of consideration for processing time.
An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network	ALCDC System (CNN)	Automatic lung cancer detection and classification using CNN.	Utilizes deep learning for improved performance.	Risk of overfitting, requires further optimization.
A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection	k-NN with GA	K-Nearest-Neighbors method enhanced by genetic algorithm for feature selection.	Hybridization increases classification accuracy.	Genetic Algorithm's stochastic nature may not always find optimal solutions.
Genetic Algorithm (GA) for prediction of lung cancer using data mining techniques	GA for Prediction	Genetic Algorithm-based ensemble method for lung cancer prediction.	Improves performance over classical approaches.	Interpretability of GA-generated models might be challenging.
A data mining based clinical decision support system for survival in lung	CDSS with XGBoost	Clinical Decision Support System for survival prediction in lung	Assists in evidence-based management	Potential biases or errors in data mining algorithms could

cancer		cancer.	advice.	affect predictions.
Dimensionality Reduction in Gene Expression Data Sets	Dimensionality Reduction	Utilizes attribute selection and PCA for gene expression data sets.	Reduces computational complexity and improves visualization.	May lose important information and requires careful selection of techniques.
Predicting Lung Cancer Using Data mining Techniques with the Aid of SVM Classifier	SVM Classifier	Predicting lung cancer classification using SVM.	Provides reliable prediction for lung cancer classification.	The use of biopsy tools may be challenging for patients.
Four layers image representation for prediction of lung cancer genetic mutations based on 2DPCA	2DPCA	Image representation for prediction of lung cancer genetic mutations.	Reduces feature dimensionality and computational complexity.	Computational time for prediction and identification may be high.
Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest	Improved Random Forest	Predicts survival rate of NSCLC patients.	Assists in selecting optimal treatment options.	May not perform well with staging information alone.
Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations	Air Quality Prediction	Predicts air quality using ANN, CNN, and LSTM networks.	Offers improved performance in air quality forecasting.	Requires consideration of explicit Air box sensor source models.
C5.0 algorithm for Analysis of the factors influencing lung cancer hospitalization expenses using data mining	C5.0 Algorithm	Analyzes factors influencing lung cancer hospitalization expenses.	Offers high accuracy in analyzing factors impacting hospitalization expenses.	May struggle with large datasets and parameter tuning.
J48 algorithm with improved decision tree for Classification of lung cancer subtypes by data mining technique	J48 with Improved DT	Classifies lung cancer subtypes using an enhanced decision tree.	Offers an approach for differential diagnosis of cancers.	The trend of common therapy for both cancer types may not be improved.
Boosted Transfer Learning Approach for Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients	Boosted Transfer Learning	Improves drug sensitivity prediction in multiple myeloma patients.	Enhances model accuracy and generalization.	Requires significant computational resources and fine-tuning.
One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) for Diagnosis of lung cancer prediction	ODANB and NCC2	Enhances diagnosis of lung cancer using rule-based classifiers.	Handles imprecise probabilities and incomplete datasets.	May require computational resources and careful selection of features.

system using data mining classification techniques				
Reduced-Order Constrained Optimization (ROCO) for Optimization of lung cancer using modern data mining techniques	ROCO	Optimizes lung cancer treatment using ACO techniques.	Efficiently optimizes treatment with reduced computational complexity.	May overlook nuances in patient data, impacting treatment efficacy.

3. Conclusion

In this paper, employing data mining techniques for lung cancer deduction proves invaluable in healthcare. Through analysis of extensive datasets, patterns and insights crucial for early detection and personalized treatment are uncovered. Despite challenges such as data complexity and model optimization, the potential to significantly enhance diagnostic accuracy and patient outcomes is evident. Moving forward continued research and integration of advanced data mining methods hold promise for further improving lung cancer deduction and patient care.

4.References

1. Thamilselvan P. Lung cancer prediction and classification using Adaboost data mining algorithm. *International Journal of Computer Theory and Engineering*. 2022 Nov; 14(4):149-54.
2. AbdulW.An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network.*13th International Conference on Developments in eSystems engineering (DeSE), Liverpool, United Kingdom*.2020, pp. 443-446, doi: 10.1109/DeSE51703.2020.9450778.
3. Maleki N, Zeinali Y, Niaki ST. A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*. 2021 Feb 1; 164:113981.
4. Vinmalar FL, Kombaiya AK. Prediction of lung cancer using data mining techniques.*In2022 3rd International Conference on Smart Electronics and Communication* (ICOSEC). 2022 Oct 20 (pp. 975-977). IEEE.
5. Pontes B, Núñez F, Rubio C, Moreno A, Nepomuceno I, Moreno J, Cacicedo J, Praena-Fernandez JM, Rodriguez GA, Parra C, León BD. A data mining based clinical decision support system for survival in lung cancer. reports of practical Oncology and radiotherapy. 2021; 26(6):839-48.
6. Taveira De SouzaJ., A. Carlos De Francisco and D. Carla De Macedo. Dimensionality Reduction in Gene Expression Data Sets.*IEEE Access*, vol. 7, pp. 61136-61144, 2019, doi: 10.1109/ACCESS.2019.2915519.
7. S. Senthil and B. Ayshwarya. Predicting Lung Cancer Using Data mining Techniques With the AID of SVM Classifier.*Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, Bangalore, India, 2018, pp. 210-216, doi: 10.1109/ICGCIoT.2018.8753095.
8. AbdelwahabM. M., and S. A. Abdelrahman, Four layers image representation for prediction of lung cancer genetic mutations based on 2DPCA,*2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, USA, 2017, pp. 599-602, doi: 10.1109/MWSCAS.2017.8052994.
9. Nanda P. and N. Duraipandian, Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest,*2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2020, pp. 93-97, doi: 10.1109/ICICT48043.2020.9112558.
10. Soh PW, Chang JW, Huang JW. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *Ieee Access*. 2018 Jun 22; 6:38186-99.
11. Yu T, He Z, Zhou Q, Ma J, Wei L. Analysis of the factors influencing lung cancer hospitalization expenses using data mining. Thoracic Cancer. 2015 May;6(3):338-45.
12. DassM. V., M. A. Rasheed and M. M. Ali, Classification of lung cancer subtypes by data mining technique,*Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, Calcutta, India, 2014, pp. 558-562, doi: 10.1109/CIEC.2014.6959151.

13. TurkiT., Z. Wei and J. T. L. Wang, Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients, in *IEEE Access*, vol. 5, pp. 7381-7393, 2017, doi: 10.1109/ACCESS.2017.2696523.
14. Krishnaiah V, Narsimha G, Chandra NS. Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*. 2013 Dec;4(1):39-45.
15. Sowmiya T, Gopi M, Begin M, Robinson LT. Optimization of lung cancer using modern data mining techniques. *International Journal of Engineering Research*. 2014;3(5):309-14.