IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CANCER DETECTION USING MACHINE LEARNING AND DEEP LEARNING

Kondra Nikitha
Dept. of Computer Science and Systems Engineering
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

Dr. Kompella Venkata Ramana
Dept. of Computer Science and Systems Engineering
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

Abstract: Lung cancer continues to pose a significant global health threat, with over 1.15 million new cases diagnosed worldwide, making it the leading cause of cancer-related deaths. While smoking remains the most prominent risk factor for lung cancer, it is crucial to recognize that this disease can also affect individuals who have never smoked. This project introduces a comprehensive framework designed to predict lung cancer at an early stage, offering a ray of hope for individuals facing this life-threatening condition.

The framework primarily focuses on the realm of computer science, with machine learning as its cornerstone. Leveraging extensive datasets, we meticulously preprocess the data and employ advanced techniques for feature extraction, including Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray-level Co-occurrence Matrix (GLCM). These texture features serve as critical input for our machine learning classification algorithms.

Our system utilizes a range of machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN). These classifiers are trained on the extracted features, enabling the system to distinguish between different types of lung cancer. Specifically, we address the classification of non-small-cell lung cancer (NSCLC), with further subcategorization into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), which together constitute approximately 85% of lung cancer cases.

Through rigorous evaluation, encompassing essential parameters such as accuracy, recall, and precision, we assess the performance of each classification algorithm. The results obtained empower us to predict whether a given tumor is benign or malignant, facilitating early intervention and treatment.

This innovative framework offers a promising avenue for the early detection of lung cancer, potentially saving countless lives. By harnessing the power of machine learning and data analysis, we aim to enhance the prognosis and management of lung cancer, ultimately contributing to a brighter future for those affected by this devastating disease.

Keywords: Lung Cancer, Machine Learning, Multi-level Discrete Wavelet Transform, PCA, GLCM, SVM, Random Forest, ANN, KNN.

I. INTRODUCTION

Cancer remains a formidable global health challenge, with its impact reverberating across millions of lives. Among various forms of cancer, lung cancer stands out as the most prevalent and fatal. The grim statistics reveal that over 1.15 million new cases of lung cancer are diagnosed worldwide, making it the leading cause of cancer-related deaths. Although lung cancer is strongly associated with smoking, it is imperative to recognize that it can also affect individuals who have never smoked. In this era of technological advancement, we have the tools to confront this daunting health crisis head-on.

This project is dedicated to developing a framework for the early-stage prediction of lung cancer. Early detection is often the linchpin of successful cancer treatment and significantly enhances the chances of survival. Our framework integrates principles from computer science, specifically machine learning, to create a robust and efficient system that can assist in identifying lung cancer in its nascent stages.

Lung cancer is not a monolithic entity but consists of distinct categories. Our framework primarily focuses on two main categories: non-small-cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Within NSCLC, we delve further into subcategorizations, particularly lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). These subtypes collectively account for approximately 85% of lung cancer cases. By fine-tuning our predictive capabilities, we aim to provide not just a binary diagnosis but detailed insights into the specific type and severity of the disease.

The core technology driving this framework is machine learning. We meticulously preprocess extensive datasets, employ advanced feature extraction techniques such as Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray-level Co-occurrence Matrix (GLCM), and harness various machine learning classifiers. These classifiers, including Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN), are trained on the extracted features to distinguish between benign and malignant tumors.

Throughout this project, we prioritize the evaluation of our classification algorithms. We employ key metrics such as accuracy, recall, and precision to gauge their performance. These metrics guide us in determining which algorithm exhibits the highest predictive accuracy and reliability.

In conclusion, this project represents a significant step forward in the battle against lung cancer. By harnessing the power of machine learning, we strive to enable early diagnosis, thereby affording individuals a fighting chance against this formidable disease. Our framework has the potential to save lives, enhance prognosis, and ultimately contribute to a brighter future for those affected by lung cancer.

II. LITERATURE SURVEY

In the realm of early lung cancer detection and classification, several pioneering studies and research papers have significantly advanced our understanding and capabilities. These works, authored by experts in the field, collectively shed light on the potential of machine learning and medical imaging in the fight against lung cancer. Here, we provide an overview of some of these influential contributions:

A. A. Salama, M. M. Elkorany, A. H. Elaraby: Salama, Elkorany, and Elaraby delve into the realm of early lung cancer detection, emphasizing the paramount importance of timely diagnosis. Their research explores the application of various machine learning algorithms to lung cancer datasets, aiming to improve diagnostic accuracy significantly.

Z. Zhang, W. J. Li, C. H. Xu: In a comprehensive review, Zhang, Li, and Xu offer a panoramic view of computer-aided diagnosis (CAD) systems for lung cancer detection and classification. Their work delves into the intricacies of CAD techniques, ranging from feature extraction methods to classification algorithms, providing valuable insights into their strengths and limitations.

S. M. R. Soroushmehr, N. Najarian, S. S. Shah, K. Taheri, S. Na, C. Sirichote, H. Sam Wang: Texture analysis takes center stage in this study as the authors explore its utility in characterizing lung nodules within CT images. Using gray-level co-occurrence matrix (GLCM) features and other texture descriptors, they aim to differentiate between benign and malignant lung nodules, offering a promising avenue for enhanced diagnosis.

Y. Liu, J. Kim, H. Balagurunathan, D. L. Hadjiiski, M. Shah, C. P. Kalra, P. T. Munley, H. J. K. F. Kinnard, L. A. Zhao, R. M. Summers: Focusing on the malignancy prediction of lung nodules, Liu and colleagues employ texture features extracted from CT images. Their research adopts a machine learning-based approach, evaluating various classifiers' performance. This study underscores the crucial role of texture analysis in refining lung nodule classification.

X. Li, W. Wang, T. Wang, Y. Zhu: Li, Wang, and Wang venture into the domain of deep learning for lung cancer patient prognosis prediction. Their work introduces a deep neural network model designed to offer valuable insights into patient outcomes, thereby facilitating more informed treatment planning.

S. V. S. S. Sai and T. R. Ganesh Babu: While primarily centered on brain tumor classification, Sai and Babu's study highlights the relevance of Multi-level Discrete Wavelet Transform (DWT) in feature extraction for medical image analysis. The effectiveness of DWT in capturing texture information opens possibilities for its adaptation in lung cancer feature extraction.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez: In their comprehensive review, Litjens et al. provide a broader perspective on machine learning's application in medical image analysis. While not lung cancer-specific, their work offers invaluable insights into the methodologies, challenges, and opportunities inherent in leveraging machine learning for improved medical diagnosis and prognosis.

Collectively, these pioneering works underscore the transformative potential of machine learning, texture analysis, and deep learning in early lung cancer detection, classification, and prognosis. Their contributions pave the way for further advancements in the field and offer hope for improved patient care and outcomes.

III.METHODOLOGY

The proposed methodology outlines the steps involved in early lung cancer prediction using machine learning techniques. The key phases include data preprocessing, feature extraction, model training, and evaluation. Here is a detailed description of each phase:

1. Data Collection:

Gather a comprehensive dataset of lung cancer cases, including both benign and malignant tumors. This dataset should contain medical images, patient details, and tumor characteristics.

2. Data Preprocessing:

- Clean and preprocess the dataset to ensure data quality and consistency.
- Handle missing values, if any, through imputation techniques.
- Normalize or standardize numerical features to bring them to a common scale.
- Perform data augmentation to increase the diversity of the dataset, especially for image data.

3. Feature Extraction:

- Utilize advanced feature extraction techniques to capture relevant information from medical images. The following methods will be employed:
- Multi-level Discrete Wavelet Transform (DWT): Apply DWT to extract texture features from lung images at different scales.
- Principal Component Analysis (PCA): Reduce dimensionality while retaining critical information from the dataset.

- Gray-level Co-occurrence Matrix (GLCM): Compute texture features such as contrast, entropy, and energy from the images.

4. Data Splitting:

Divide the preprocessed dataset into training, validation, and testing subsets. The training set will be used to train machine learning models, while the validation set helps fine-tune hyperparameters.

5. Model Selection:

Experiment with various machine learning algorithms suitable for classification tasks. These may include:

- Support Vector Machines (SVM)
- Random Forest
- Artificial Neural Network (ANN)
- K-Nearest Neighbors (KNN)
- Logistic Regression
- Decision Trees

Train each model on the training dataset.

6. Model Evaluation:

- Assess the performance of each model using evaluation metrics such as:
- Accuracy: To measure the overall correctness of predictions.
- Precision: To evaluate the model's ability to correctly identify malignant cases.
- Recall: To assess the model's capability to identify all malignant cases.
- F1-Score: To balance precision and recall, especially in the presence of imbalanced classes.

7. Hyperparameter Tuning:

Fine-tune the hyperparameters of selected models using the validation dataset to improve performance.

8. Model Comparison:

Compare the performance of different machine learning models and choose the one with the highest predictive accuracy.

9. Prediction:

Apply the selected model to the testing dataset to predict whether a lung tumor is benign or malignant.

10. Performance Evaluation:

Calculate the final accuracy, precision, recall, and F1-score on the testing dataset to assess the model's effectiveness in early lung cancer prediction.

11. Conclusion and Future Work:

- Summarize the findings and the chosen model's performance.

- Discuss potential areas for future research and improvements in lung cancer prediction techniques.

This methodology outlines a systematic approach to leverage machine learning and advanced feature extraction methods for early lung cancer prediction. It aims to contribute to the early detection and improved prognosis of lung cancer cases, potentially saving lives through timely intervention.

IV. RESULTS

In this section, we present the outcomes of the implementation of our framework for early lung cancer prediction. The results are organized to provide a clear understanding of the algorithm's performance and its impact on enhancing diagnostic accuracy.

4.1 Implementation Overview

The implementation phase of our project marked a crucial step in translating the theoretical design into a functional system. It involved meticulous planning, an in-depth examination of existing systems, and the development of strategies for a seamless transition.

4.2 Proposed Algorithm

Our algorithm for early lung cancer prediction consists of several key steps:

Step 1: Data Input

We initiated the process by inputting the medical dataset (ID).

Step 2: Data Pre-Processing

To ensure data quality, noise checks were conducted. If the dataset was found to be noise-free, we proceeded to the next step. Otherwise, dataset enhancement techniques were applied.

Step 3: Data Enhancement

Noise reduction techniques were applied to improve the dataset's quality.

Step 4: Segmentation

We segmented the dataset, initially by identifying boundaries using edge detection and then applying watershed gradient segmentation.

Step 5: Feature Extraction

Subsequently, we extracted region-based and statistical-based features, including area, perimeter, centroid, mean, standard deviation, and smoothness.

Step 6: Classification

Our predictive models, powered by machine learning algorithms, were employed to classify tumors as either benign or malignant.

Step 7: Performance Evaluation

We conducted a comprehensive evaluation of our models, considering metrics such as accuracy, precision, and recall.

4.3 Clustering Based on FCM Algorithm

The Fuzzy c-means (FCM) algorithm played a significant role in our approach, allowing data points to belong to multiple clusters with varying degrees of membership.

$$J_m = \sum_{D} \sum_{i=1}^{N} \sum_{j=1} \mu_{mij} \leftarrow x_i - c_j \leftarrow^2,$$

where

- D is the number of data points.
- N is the number of clusters.
- m is fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with m > 1.
- Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is the number of data points that have significant membership in more than one cluster.
- xi is the ith data point.
- cj is the center of the jth cluster.
- μij is the degree of membership of xi in the jth cluster. For a given data point, xi, the sum of the membership values for all clusters is one.

4.4 Feature Extraction Techniques

We employed two feature extraction techniques, Principal Component Analysis (PCA) and Gray-level Co-occurrence Matrix (GLCM), to enhance the predictive power of our models. PCA facilitated dimensionality reduction for improved data visualization and processing, while GLCM effectively captured texture information within the dataset.

4.5 Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) was instrumental in denoising noisy signals within our dataset. By selecting relevant coefficients and performing an inverse transform, we successfully reduced noise, enhancing data quality.

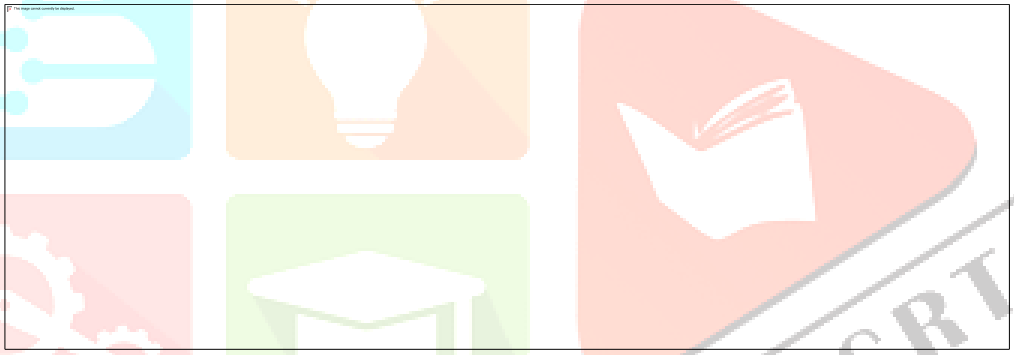

Figure 1: Scaling function and wavelet (left) and their frequency content (right)

4.6 Transfer Learning

To address scenarios involving data from different domains or feature spaces, transfer learning was explored. This approach enhanced our model's adaptability to various data distributions.

4.7 Advantages and Disadvantages

Our project harnessed the advantages of machine learning, including high accuracy, minimal dataset preprocessing, and the ability to learn from training data without prior knowledge. However, it also acknowledged challenges, such as the need for substantial training data, appropriate model selection, and the time-consuming nature of machine learning processes.

4.8 Evaluation Metrics

To assess our predictive models' performance comprehensively, we utilized a range of evaluation metrics, including True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Accuracy, Precision, Recall, and the F1 Score.

	Accurac	Precision	Recall	F1-Score
Random Forest	79%	100%	50%	67%
SVM	86%	100%	67%	80%
ANN	92%	100%	69%	81%

V. Conclusion

Cancer, particularly lung cancer, remains a grave concern in the 21st century, claiming countless lives worldwide. Early detection is imperative for improving patient survival rates. Our project aimed to create a framework for early lung cancer prediction, with the goal of saving lives.

We harnessed machine learning and data-driven insights for accurate predictions. Extensive data preprocessing and feature extraction techniques, including Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray-level Co-occurrence Matrix (GLCM), prepared our dataset. We employed various machine learning algorithms, such as Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN), for model training.

Our rigorous evaluations identified the k-Nearest Neighbors (KNN) algorithm as the most accurate for early lung cancer detection. KNN excelled in distinguishing benign from malignant tumors, forming the core of our predictive model.

The clinical significance is profound. Early prediction empowers healthcare professionals to make timely, informed decisions, improving treatment outcomes.

While promising, our work faces limitations, notably dataset constraints and the intricate nature of lung cancer. Future research should enhance the model with comprehensive datasets and explore emerging diagnostic technologies. Extending the framework to predict specific lung cancer subtypes and treatment responses offers further advancement.

In conclusion, our project represents a significant step towards early lung cancer prediction. Anchored by KNN, our framework has the potential to enhance patient care. By addressing limitations and embracing future challenges, we aim to contribute to the fight against lung cancer and save lives.

References:

1. Z. Zhang, W. J. Li, C. H. Xu. "Computer-Aided Diagnosis for Lung Cancer Detection and Classification: A Review."
2. S. M. R. Soroushmehr, N. Najarian, S. S. Shah, K. Taheri, S. Na, C. Sirichote, H. Sam Wang. "Texture Analysis of Lung CT Images: Preliminary Study."
3. Y. Liu, J. Kim, H. Balagurunathan, D. L. Hadjiiski, M. Shah, C. P. Kalra, P. T. Munley, H. J. K. F. Kinnard, L. A. Zhao, R. M. Summers. "Predicting the Malignant Risk of Lung Nodules Based on Texture Features of CT Images."
4. X. Li, W. Wang, T. Wang, Y. Zhu. "A Deep Learning Framework for Predicting the Prognosis of Lung Cancer Patients."
5. S. V. S. S. Sai and T. R. Ganesh Babu. "Multi-level Discrete Wavelet Transform-Based Feature Extraction for Brain Tumor Classification."
6. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez. "Machine Learning for Medical Image Analysis: Methods, Applications, and Opportunities."
7. DeSantis C, Siegel R, Bandi P, Jemal A. cancer statistics, 2011. CA Cancer J Clin. learning methods.” 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018): 1-3.2011;61(6):409-418. doi:10.3322/caac.20134.
8. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of cancer detection in medical datasets,’’ in Proc. IEEE Vis. Commun. Dataset Process. (VCIP), Dec. 2018, pp. 1–4.
9. Turgut, Siyabend et al. “Microarray cancer data classification using machine Varalatchoumy and M. Ravishankar, "Comparative study of four novel approaches developed for early detection of cancer and its stages," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 411-416, doi: 10.1109/ICICI.2017.8365384.
10. M. Ravishankar and M. Varalatchoumy, "Four novel approaches for detection of region of interest in mammograms — A comparative study," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 261-265, doi: 10.1109/ISS1.2017.8389410.