Comprehensive Review on Lung Cancer Detection with Metabolite Profile, CT Scans using Machine Learning

Yash P Khairnar
Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune, Maharashtra, 411057, India

Amisha P Patil
Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune, Maharashtra, 411057, India

Atharva Pingale
Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune, Maharashtra, 411057, India

Parth Deshpande
Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune, Maharashtra, 411057, India

Prof. Kimi Ramteke
Department of Computer Engineering
International Institute of Information Technology
Hinjawadi, Pune, Maharashtra, 411057, India

***Abstract*—Globally, lung cancer is the second most frequent type of cancer. Lung cancer is typically identified in advanced stages, which limits the spectrum of potential treatment choices. Implementing screening for high-risk persons has the prospect of early detection, which could lead to considerable increase in survival rate. This review article's goal is to examine different machine learning and deep learning approaches that use metabolomics (plasma, serum) and CT scan images to identify lung cancer subtypes (adenocarcinoma, squamous cell carcinoma, large cell carcinoma), as well as cancer stage type (I, II, III, IV).**

***Index Terms*—Lung Cancer, Metabolomics, Metabolites, CT Scan, Machine Learning, Deep Learning**

I. Introduction

A. Lung Cancer

Lung Cancer, like any other form of cancer, is the result of genetic mutations which lead to abnormal and uncontrolled division of cells in the lungs, leading to the formation of masses called as tumor. Lung cancer tumors can be particularly harmful because they have the potential to metastasize, which means cancer cells can break away from the primary tumor and spread to other parts of the body through the bloodstream or lymphatic system. This metastatic spread can lead to the formation of secondary tumors (metastases) in vital organs like the brain, liver, bones, and other parts of the body, making the cancer even more challenging to treat.

B. Types of Lung Cancer

The types lung Cancer as represented in Fig. 1 are described below

Non-Small Cell Lung Cancer (NSCLC):

1) Adenocarcinoma: This is the most common type of NSCLC, and it often occurs in the outer regions of the lung. It is more common in non-smokers and tends to grow more slowly.

2) Squamous Cell Carcinoma: This type usually develops in the lining of the bronchial tubes and is often associated with a history of smoking. It tends to grow more centrally in the lung.

3) Large Cell (undifferentiated) Carcinoma: This is a less common type of NSCLC and can appear in any part of the lung. It tends to grow and spread quickly.

Small Cell Lung Cancer (SCLC):

4) Small Cell Carcinoma: This is a highly aggressive and fast-growing type of lung cancer. It is typically found in the central part of the lung and is strongly associated with smoking.

C. Metabolomics and metabolites

The scientific discipline of metabolomics is centered on the investigation of metabolites, which are tiny molecules or substances that play a variety of roles in an organism's metabolic processes. Sugars, amino acids, lipids, and other

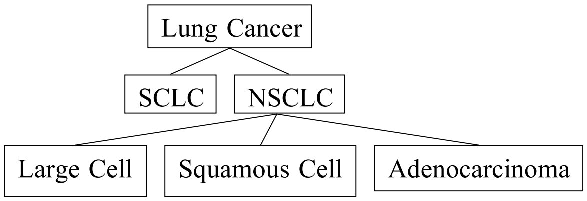

Figure 1. Tree diagram illustrating different types of lung cancer.

organic substances are examples of these tiny molecules, but they are not the only ones. The byproducts of cellular functions, metabolites are essential to the control and upkeep of biological systems.

The metabolome is "downstream" from the proteome and genome networks, which means that it is impacted by the proteins that are made by the proteome and genome as well as the genetic information that is encoded in the organism's genome.

A person's metabolomic profile is a direct indicator of their physiological and psychological health.
Metabolomics enable the collection, detection and analysis of various types of cancer related metabolites.

D. Computer Tomography

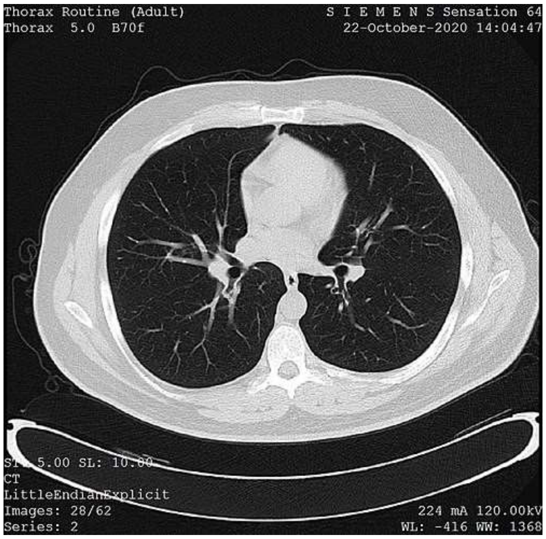

Figure 2. CT scan
(https://en.wikipedia.org/wiki/Medical_imaging)

A CT (computed tomography) scan is a popular and valuable diagnostic tool for lung cancer detection and evaluation. It is frequently used to diagnose lung cancer, assess the severity of the disease, determine cancer stage, and guide treatment decisions.

The imaging appearance of the original tumor is heterogeneous. NSCLCs can present as peripheral lesions that penetrate the chest wall or as centrally positioned masses that infect the mediastinal tissues. The margins of tumors can be uneven and spiculated, smooth, or lobulated. They may have cavitation andcentral necrosis, or they may be uniformly solid. Tumors that are cavitating and positioned centrally are more likely to have a squamous histology. The tumor may appear as a ground-glass opacity, a region of consolidation, or a combination of both and may resemble an infectious disease. It is more typical for adenocarcinoma and its subgroups to have this appearance. In bronchoalveolar carcinomas, now known as adenocarcinomas in situ, consolidation with an air bronchogram and mixed density or pure ground-glass nodules are observed. Currently TNM (Tumor, Nodal, Metastasis) staging is followed in clinicalpractice for lung cancer staging.

In this paper, we propose a new methodology to detect lung cancer in both early and advanced stages using serum, plasma metabolic profiling and computer tomography images.

II. Literature Review

Metabolomics presents issues in dealing with massive amounts of data, but modern deep learning (DL) approaches provide solutions. Recently, DL has played an important role in metabolic phenotyping, biomarker discovery, multiomics integration, pathway prediction, and metabolic modeling. [2] focuses on the practical applications of deep learning (DL) in metabolomics for data integration, prediction, statistical inference, and overcoming data gathering and processing challenges.

Suren Makajua's suggested approach [5] detects cancerous nodules in lung CT scan images utilizing watershed segmentation for detection and Support Vector Machine (SVM) for categorization of nodule as malignant or benign. Their proposed model diagnoses cancer with 92% accuracy. However, it does not differentiate between cancer stages I, II, III, and IV.Based on the metabolite profile of the patient, we can determine the type of lung cancer. In the process proposed by Ashvin [6], metabolites are classified based on their moleculardescriptors, followed by feature extraction and dimensionality reduction. They separated multiclass lung tumors into binary classifiers, each of which is a tiny neural network. Their model achieves 92.0% overall test accuracy. Each classifier has an accuracy of higher than 90%. By more than 14%, our two-stage classifier surpasses the classic naive multiclass classifier. In [21], they proposed a strategy for detecting metabolomic biomarkers with excellent accuracy. It achieved 100% accuracywith the Ridge Classifier for the Plasma sample. Using the XGBoost Classifier, they achieved 90.91% accuracy on the
Serum sample. In the case of Plasma and Serum samples, the number of most dominating metabolites was only 19 and 7, respectively.

Many computer-aided detection methods systems, in addition to deep learning and radiomic techniques exist, however the gold standard for these techniques hasn't yet to be determined. Of all the techniques covered in this paper [8], Deep learning and radiomics-based ones appear to be the most promising by giving an accuracy of 95.41%.

By combining a number of high-throughput Both Liquid Chromatography Quadrupole Time-of-Flight Mass Spectrometry (LC-QTOF/MS) and Gas Chromatography-Mass Spectrometry (GC/MS)-based untargeted metabolomics technology, as well as Liquid Chromatography Tandem Mass Spectrometry(LC-MS/MS).

Analytical technique	Metabolites found	Publication
K-nearest neighbor (KNN), Naive Bayes, AdaBoost,Support Vector Machine (SVM), Random Forest, and Neural Network with 10-cross fold technique	L-Kynurenine, Proline, Spermidine, Amino-hippuric acid, Palmitoyl-l-carnitine, Taurine, Phenylalanine, Valine, o-Tyr, Carnitine	Xie Y. et al. [3]
Logistic regression model	Palmitic acid, Heptadecanoic acid, 4-Oxoproline, Tridecanoic acid, Ornithine	Qi S. et al. [4]
Random Forest model	tryptophan, kynurenine, xanthurenic acid	Sun R. et al. [9]
Partial least-squares discriminant analysis (PLS-DA) model	Histidine, glutamine, glycine, threonine, alanine, valine, citrate, tyrosine, proline, leucine, isoleucine, pyruvate, 3-hydroxybutyrate, betaine, succinate, creatinine, lactate, creatine, acetoacetate,myo-inositol, phenylalanine, TMAO, acetate, glutamate, glucose, formate	Singh A. et al. [12]

Table I. Metabolites associated with lung cancer

Machine learning algorithms with metabolomics technologies Algorithms, this research paper [9] developed eight computational frameworks to assess pemetrexed effectiveness in the treatment of NSCLC and compare their performance further. At last, they selected a superior than other techniques and a collection of three metabolites that include kynurenine, tryptophan, and xanthurenic acid which overperforms the clinical biochemical criteria for lymphocytes and neutrophils to assess pemetrexed efficiency.

Zhu W. [10] proposed SGHF-Net, a unique deep learning network for reliably detecting lung cancer subtypes on CT scans, in this research. A pathological feature synthesis module (PFSM) and a radiological feature extraction module (RFEM) were used to build the suggested model, which built on a feature fusion framework. These two modules aided the network in gathering high-level pathological characteristics as well as high-level radiological features from CT scans alone, with the pathological features containing information on the related pathological pictures. By combining the two modules, the hybrid features covering both pathological priors and radiographic information may be produced using the feature fusion framework. They established the superiority of the proposed framework via a number of strategy comparisons and confirmed the complementarity of synthetic pathological priors (i.e., pathological features) could significantly improve the accuracy of CT-based lung cancer and subtypes classification approaches.

Wang S. et al. [11] achieved an impressive maximal accuracy of 95.75% in their thorough analysis of lung cancer image classification, outperforming both state-of-the-art models and classical methods. The potential of their method for a precise and trustworthy diagnosis of lung cancer was highlighted by this exceptional accuracy. The robustness of their results was enhanced by the balanced dataset they used for training and testing, which included 168 images for testing and 2054 images for training. Their transfer learning approach improved the model's capacity to distinguish between four different pathological types of lung cancer, with a particular focus on correctly classifying IA. The findings underscored the practical significance of their methodology and its capacity to facilitate prompt and precise identification of lung cancer, offering significant perspectives for healthcare professionals.

The study on lung cancer (LC) patients in India supports the ability of serum metabolomics analysis based on nuclear magnetic resonance (NMR) to differentiate LC from non-LC subjects and to identify clinical subtypes of LC [12]. The metabolic disturbances that have been observed, such as elevated oxidative stress, activated glutaminolysis, altered energy metabolism, cancer-related inflammation, increased amino acid utilization, and support for anabolic metabolism, indicate the diagnostic utility of this approach.

Dimililer K. et al [13]. presents an image enhancement tumor detection system. To provide meaningful representations of lung patterns, image processing techniques are used to reduce the amount of data with computational and time costs. A relationship between the patterns of input and output has been established in order to identify the tumors present in the images. Raw 256x256 input images are obtained, and black and white pixels are removed through thresholding. Noise is then eliminated using erosion and median filtering, the image's tiny objects are eliminated. The location of the tumor is determined by subtracting the small objects removed image from the median filtered image.

The comprehensive review [14] highlights how Deep Learning (DL) is revolutionizing many aspects of cancer research and treatment. The potential of DL methodologies has been shown in the areas of precision medicine, drug design, metastasis prediction, cancer diagnosis, and drug response prediction. They present promising paths toward the discovery of new chemical structures, the customization of therapeutic approaches, and the advancement of our knowledge of the biology of cancer. The utilization of multi-omics and medical imaging data, such as MRI scans, histology slides, DNA methylation, and gene expression, by DL has created novel and promising opportunities in the fight against cancer. DL's ability to process mass spectrometry data and find metabolic biomarkers helps us better understand metabolic heterogeneity within cancer as we dive into metabolomics analysis.

Pomyen Y. [15] highlights how Deep Learning (DL) is revolutionizing many aspects of cancer research and treatment. The potential of DL methodologies has been shown in the areas of precision medicine, drug design, metastasis prediction, cancer diagnosis, and drug response prediction. They present promising paths toward the discovery of new chemical structures, the customization of therapeutic approaches, and the advancement of our knowledge of the biology of cancer. The utilization of multi-omics and medical imaging data, such as MRI scans, histology slides, DNA methylation, and gene expression, by DL has created novel and promising opportunities in the fight against cancer. DL's ability to process mass spectrometry data and find metabolic biomarkers helps us better understand metabolic heterogeneity within cancer as we dive into metabolomics analysis.

The article presented by Raza R. et al. [16] Lung-EffNet, a novel method for classifying lung cancer that makes use of EfficientNet's powers on CT scan Images. The study overcomes inherent challenges in medical image analysis by fine-tuning five EfficientNet variants through transfer learning, outperforming conventional convolutional neural network architectures. Surprisingly, the suggested approach achievesa remarkable 99.10% accuracy rate along with strong ROC scores between 0.97 and 0.99, confirming its excellent per- formance in the classification of lung cancer. This ground- breaking discovery represents a major improvement in the ac- curacy and efficacy of lung cancer diagnosis, with potentially profound effects on the field of medical image analysis as a whole. The article also outlines potential directions for future research, laying the groundwork for ongoing advancements and innovation in this crucial field.

The thorough analysis of Li Y. [17] shows how the application of artificial intelligence (AI) to lung cancer researchis a game-changer that is bringing in a new era of pre- cision oncology. The potential for enhanced drug response assessment, immunotherapy procedures, prognosis prediction, early detection, diagnosis, and diagnosis is highlighted by the investigation of AI-driven decision support tools in various aspects of lung cancer therapy. In addition, the review offers insightful information about baseline methods, datasets, and method characteristics that will serve as a basis for further machine learning research in this important area. However, there are still issues with data quality, interpretability, model robustness, and performance metrics. A multi-modal system that seamlessly integrates imaging and omics data is thoughtto represent the future of lung cancer therapies.

The article [18] explores the field of supervised learning approaches for cancer detection with a particular emphasis on the LUNA16 dataset. The suggested hybrid model, SVM and CNN, surpasses previous approaches and shows a notable improvement in performance. It emphasizes how important feature selection from intricate datasets is to improving the accuracy of machine learning models. Deep learning techniques have a higher computational cost but have the benefit of automatic feature extraction, even though they have the potential for over-fitting. However, the hybrid approach, whichuses SVM for classification and CNN for feature selection, combines the advantages of both deep learning and conven- tional machine learning. This synergy produces a definitive result that is faster than current methods and highly accurate, eliminating the need for a separate feature selection process. The present study offered by Galal A. [19] thorough analysis of how machine learning (ML) approaches are developing in the field of metabolomics. Although support vector machines(SVM) and random forests (RF) have been used extensively, metabolomic analyses are undergoing a paradigm shift as deep learning (DL) becomes more and more popular. The main focus has been on cancer stratification, whereby machine learning algorithms have been able to classify different types of cancer by comparing them to control sample sets. Because dataquality and characteristics vary, selecting an ML algorithm is still a crucial choice, and no single method performs better than others in all situations. Additional research into method selection is necessary because hyperparameter tuning and feature selection have a major impact on the outcomes. The identification of biomarkers, predictive validation, enrichment analyses, and possible disease treatment avenues are all made
possible by these ML-driven investigations.

The results of the research [20] indicate that diagnostic features have a higher discriminating power than texture features primarily because the tiny size of the lymph nodes makesit difficult to calculate texture features. CNN is a popular technique that classifies images based on their appearancepatterns. In the research, the CNN's performance is compara- ble to the best classical methods, despite not utilizing critical diagnostic features such as SUV and tumor size. Furthermore, CNN eliminates the need for tumor segmentation and feature selection, making the process much more convenient and less susceptible to user bias. CNN also does not accept handcrafted features as input. CNN also stays away from the contentious texture features that are influenced by the size of the tumor.

III. Proposed Methodology

We suggest a fresh methodology (figure no.3) that is distinct in its own right. The system's general design is a hybrid of three independent machine learning models, each with its own set of algorithms. We intend to extract the most essential metabolites from the metabolite's dataset of serum and plasma samples as done in [21] that have a substantial influence or show variation in themselves for different types and stages of lung cancer. We employ feature extraction methods to extract these metabolites and then train a classification model based

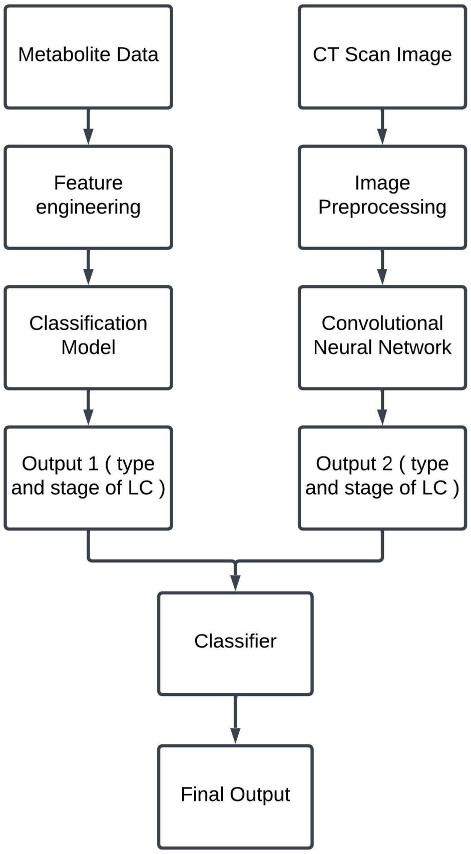

Figure 3. Proposed methodology

on these retrieved features. To ensure that the total system's correctness is not relied on a single source, we incorporate a convolution neural network (CNN) that can take CT Scan images as input and produce an output that indicates the kind and stage of cancer.

The final step is to integrate or combine these results. To do this, we utilize the two results from the two separate models that we received as features for the third and final classifier. This gives us the final output, which is the kind and stage of lung cancer for the patient in question.

IV. Conclusion

A number of studies that are pertinent to the identification and staging of lung cancer are covered in this paper. We reviewed how deep learning and machine learning algorithms can help in accurate lung cancer subtyping and staging. It's evident from the reviews that there is an association between a few of the metabolites and the lung cancer. Additionally, this review also highlights the use of image classification and image preprocessing to identify lung cancer from the CT Scans images.

We hope that our proposed methodolgy, which integrates multiple approaches, improves the lung cancer detecion accuracy.

Thus, we would be able to analyze the type and stage of lung cancer without the need for any form of invasive medical testing by utilizing machine learning, deep learning techniques, and statistical analysis. Lung cancer early diagnosis is greatly encouraged by this strategy. We plan to apply this method to additional types of cancer in the future.

V. References

[1] Guan X, Du Y, Ma R, Teng N, Ou S, Zhao H, Li X, "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices," BMC Medical Informatics and Decision Making, June 2023.

[2] Sen P, Lamichhane S, Mathema VB, McGlinchey A, Dickens AM, Khoomrung S, Oresic M, "Deep learning meets metabolomics: a methodological perspective," Oxford: Briefings in Bioinformatics, September 2020.

[3] Xie Y, Meng WY, Li RZ, Wang YW, Qian X, Chan C, Yu ZF, Fan XX, Pan HD, Xie C, Wu QB, Yan PY, Liu L, Tang YJ, Yao XJ, Wang MF, Leung ELH, "Early lung cancer diagnostic biomarker discovery by machine learning methods," Elsevier: Translational Oncology, September 2020.

[4] Qi S, Wu Q, Chen Z, Zhang W, Zhou Y, Mao K, Li J, Li Y, Chen J, Huang Y, Huang Y, "High-resolution metabolomic biomarkers for lung cancer diagnosis and prognosis," Scientific Reports, 2021. prognosis," scientific reports, 2021.

[5] Makajua S, Prasad PWC, Alsadoona A, Singhb AK, Elchouemic A, "Lung Cancer Detection using CT Scan Images," pp. 107–114, Elsevier, ScienceDirect: Procedia Computer Science, 125 (2018) in press.

[6] Choudhary A, Yu J, Kouznetsova VL, Kesari S, Tsigelny IF, "Two-Stage Deep-Learning Classifier for Diagnostics of Lung Cancer Using Metabolites," MDPI: Metabolites, October 2023.

[7] Bhattacharjee A, Dey J, Kumari P, "A combined iterative sure independence screening and Cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer," Elsevier: Healthcare Analytics 2, 2022.

[8] Binczyk F, Prazuch W, Bozek P, Polanska J, "Radiomics and artificial intelligence in lung cancer screening", Review Article on Implementation of CT-based Screening of Lung Cancer, Jan 2021.

[9] Sun R, Fei F, Wang M, Jiang J, Yang G, Yang N, Jin D, Xu Z, Cao B, Li J, "Integration of metabolomics and machine learning revealed tryptophan metabolites are sensitive biomarkers of pemetrexed efficacy in non-small cell lung cancer," Wiley: Cancer Medicine, August 2023.

[10] Zhu W, Jin Y, Ma G, Zhejiang Lab, Chen G, Egger J, Zhang S, Metaxas DN, "Classification of Lung Cancer subtypes on CT Scan Images with Synthetic Pathological priors," August 2023.

[11] Wang S, Dong L, Wang X, Wang X, "Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy," Open Med., pp. 190-197, December 2019.

[12] Singh A, Prakash V, Gupta N, Kumar A, Kant R, "Serum Metabolic Disturbances in Lung Cancer Investigated through an Elaborative NMR-Based Serum Metabolomics Approach," ACS Omega, vol. 7, pp. 5510-5520, 2022.

[13] Dimililer K, Uger B, Ever YK, "Tumor Detection on CT Images using Image Enhancement," TOJSAT-The Online Journal of Science and Technology, Volume 7, Issue 1, January 2017.

[14] Mathema VB, Sen P, Lamichhane S, Oresic M, Khoomrung S, "Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine," Elsevier: Computational and Structural Biotechnology Journal 21, pp. 1372–1382, 2023

[15] Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S, "Deep metabolome: Applications of deep learning in metabolomics," Elsevier: Computational and Structural Biotechnology Journal 18, pp. 2818–2825, 2020.

[16] Raza R, Zulfiqar F, Khan MO, Arif M, Alvi A, Iftikhar MA, Alam T, "Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images," Elsevier: Engineering Applications of Artificial Intelligence 126, July 2023.

[17] Li Y, Wu X, Yang P, Jiang G, Luo Y, "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis," Elsevier: Genomics Proteomics Bioinformatics, vol. 20, pp. 850–866, November 2022.

[18] Shafi I, Din S, Khan A, D íez I. L. Torre Ḿıez, Casanova Rí J, Pifarre K T, Ashraf, "An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network". MDPI: Cancers, 2022.

[19] Galal A, Talal M, Moustafa A. "Applications of machine learning in metabolomics: Disease modeling and classification". Fronteirs in Genetics, 2022.

[20] Wang W, Rong Z, Wang G, Hou Y, Yang F, Qiu M., "Cancer metabolites: promising biomarkers for cancer liquid biopsy". Biomarker Research 2023.

[21] Ghosh U.K, Abir F, Rifaat N, Shovan S.M, Sayeed A, Hasan M.A. "Most dominant metabolomic biomarkers identification for lung cancer". Elsevier: Informatics in Medicine Unlocked 2022.