IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.

IJSRA

International Journal of Science and Research Archive

Research Journal Archive, INDIA

(Review Article)

Check for updates

From data to diagnosis: A review of the current state of the art in lung cancer prediction using machine learning

Samuel Fanijo [1, *], Olusola Olabisi Ogunseye [2], Olumayowa Adeleke Idowu [3], Olufemi Olulaja [4] and Oghenetanure Ryan Enaworu [5]

[1] *Department of Computer Science, Iowa State University, USA.*
[2] *Department of Community, Environment and Policy, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, USA.*
[3] *Department of Economics, University of Pittsburgh, Pittsburgh, USA.*
[4] *African Center of Excellence for Genomics of Infectious Disease, Redeemers University, Nigeria.*
[5] *Department of Microbiology and Pharmacology, St. George's University, St. George, Grenada.*

International Journal of Science and Research Archive, 2022, 06(02), 292–314

Publication history: Received on 12 April 2022; revised on 21 June 2022; accepted on 24 June 2022

Article DOI: https://doi.org/10.30574/ijsra.2022.6.2.0118

Abstract

Lung cancer remains a leading cause of cancer-related mortality worldwide, with 1.8 million cases of lung cancer-related death recorded in 2022 alone. This is often due to late-stage diagnosis and the complexity of its molecular subtypes, necessitating the need for early detection and personalized treatment, to improve patient outcomes. Predictive biomarkers—biological indicators that help detect, monitor, and guide treatment—help in addressing this challenge. However, traditional methods of biomarker discovery often struggle to cope with the heterogeneity of lung cancer and the vast datasets generated from genomic, proteomic, and imaging technologies. In response, machine learning (ML) has emerged as a tool, capable of analyzing data and identifying novel biomarkers that may be overlooked through conventional techniques. This paper reviews the current state of predictive biomarker discovery in lung cancer, focusing on the application of machine learning approaches. It examines the types of biomarkers used in lung cancer diagnosis and treatment, recent advancements in ML-driven biomarker discovery, and the challenges that persist—such as data quality, model overfitting, and interpretability. This paper concludes with recommendations for future research directions, emphasizing the need for improved data integration, better model interpretability, and clinical validation of biomarkers to ensure that machine learning can fully realize its potential in revolutionizing lung cancer care.

Keywords: Lung cancer; Predictive biomarkers; Artificial Intelligence; Machine learning; Genomic data

1. Introduction

According to the World Health Organization, as of 2022, Lung cancer is the most common form of cancer worldwide, accounting for 12.4% of new cancer cases in 2022, and responsible for 18.7% of cancer-related deaths [1, 2]. In fact, according to the World Cancer Research Fund International, there were 2,480,675 new cases of lung cancer in 2022 [3]. The top 10 countries with the highest rates of lung cancer and lung cancer-related deaths in 2022 are shown in the Figure 1. Its aggressive progression, often undetected until advanced stages, makes treatment challenging and outcomes poor for many patients [4]. The high level of intra-tumor heterogeneity (ITH) and the complexity of cancer cells, which contribute to drug resistance, complicate cancer treatment. Also, the high mortality rate is due to late-stage diagnoses, when treatment options are limited and survival rates plummet [5]. Early detection of lung malignancies is therefore imperative for any anti-cancer treatment to reduce mortality and morbidity, especially in high-risk [6].

[*] Corresponding author: Samuel Fanijo

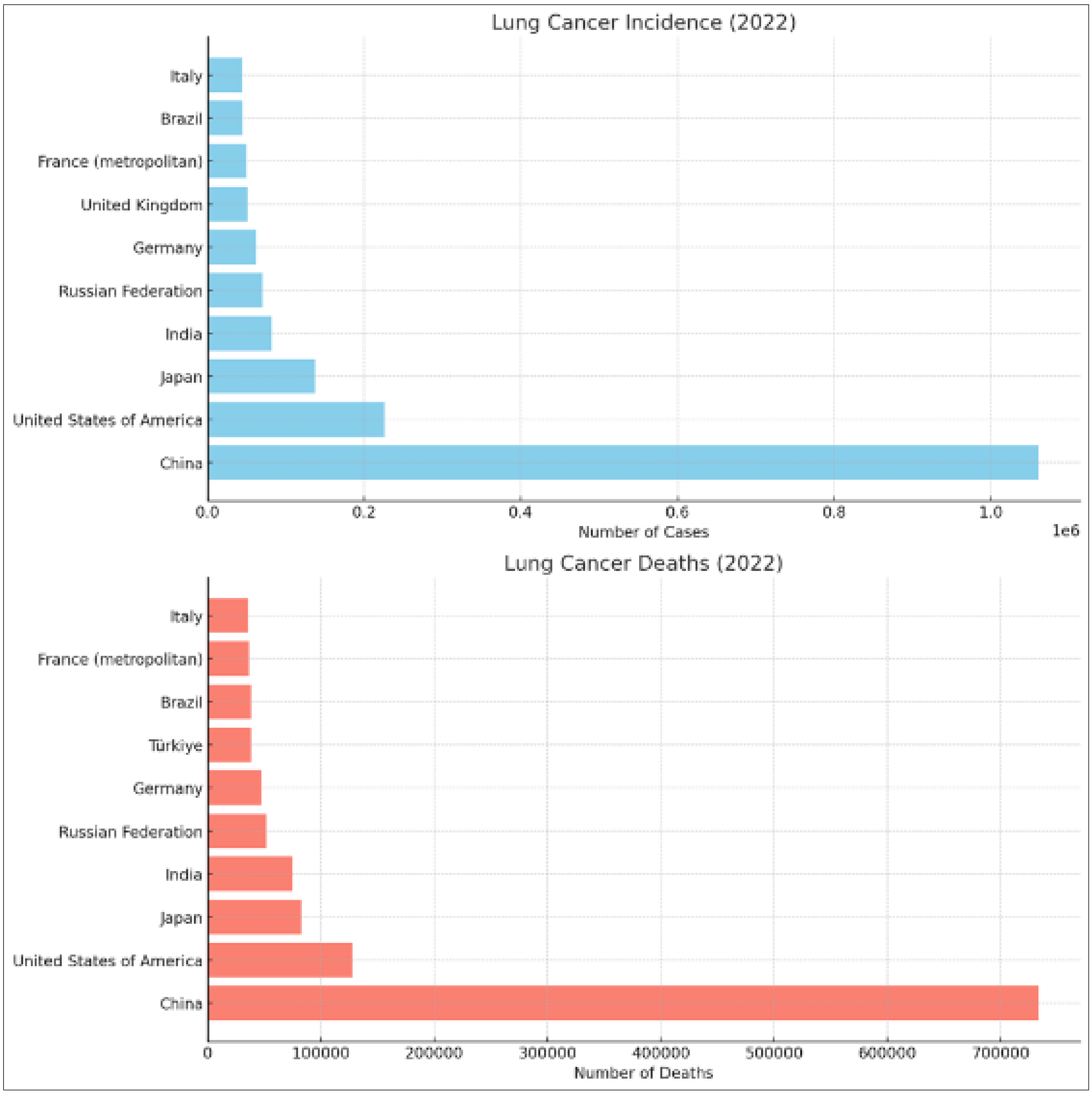

Figure 1 Top-10 countries with highest lung cancer incidence and deaths in 2022

Equally important is the development of personalized treatment strategies that are tailored to the unique characteristics of an individual's cancer, thus, enhancing treatment efficacy while reducing the chances of resistance to therapies [7].

Over the past decades, advancements in cancer research technologies have driven the development of numerous large-scale collaborative cancer projects, resulting in the creation of extensive clinical, medical imaging, and sequencing databases. These databases enable researchers to explore lung cancer comprehensively, from diagnosis and treatment to clinical outcomes [8, 9].

Notably, one of the tools advancing both early detection and personalized treatment in lung cancer is the use of predictive biomarkers, which are biological molecules that indicate the presence of cancer, predict disease progression, or inform the likely response to treatment [10]. Biomarkers can be found in various biological samples, such as blood, tissue, and other bodily fluids, and encompass a range of molecular entities such as genes, proteins, and metabolites [11]. Studies focusing on omics analysis, including genomics, transcriptomics, proteomics, and metabolomics, have further enhanced research tools and capabilities, facilitating the detection of these biomarkers. By identifying these biomarkers, clinicians can assess an individual's risk of developing lung cancer, enable earlier diagnoses, and determine the most appropriate treatments.

Among the various biological samples, blood is often considered the primary source for biomarker candidates due to its ability to reflect the overall condition of the patient, encompassing information from the primary tumor, metastases, immune response, and surrounding tissue. However, other samples more specific to lung cancer, such as sputum, bronchial lavage, or exhaled breath, can provide insights closer to the tumor and its microenvironment, potentially offering more precise information for clinical decisions, particularly in early-stage disease. Additionally, samples such as urine and saliva are emerging as valuable resources for biomarker detection, particularly in metabolomics-based studies. Fig. 2 provides a graphical overview of the current promising molecular biomarkers for lung cancer screening.

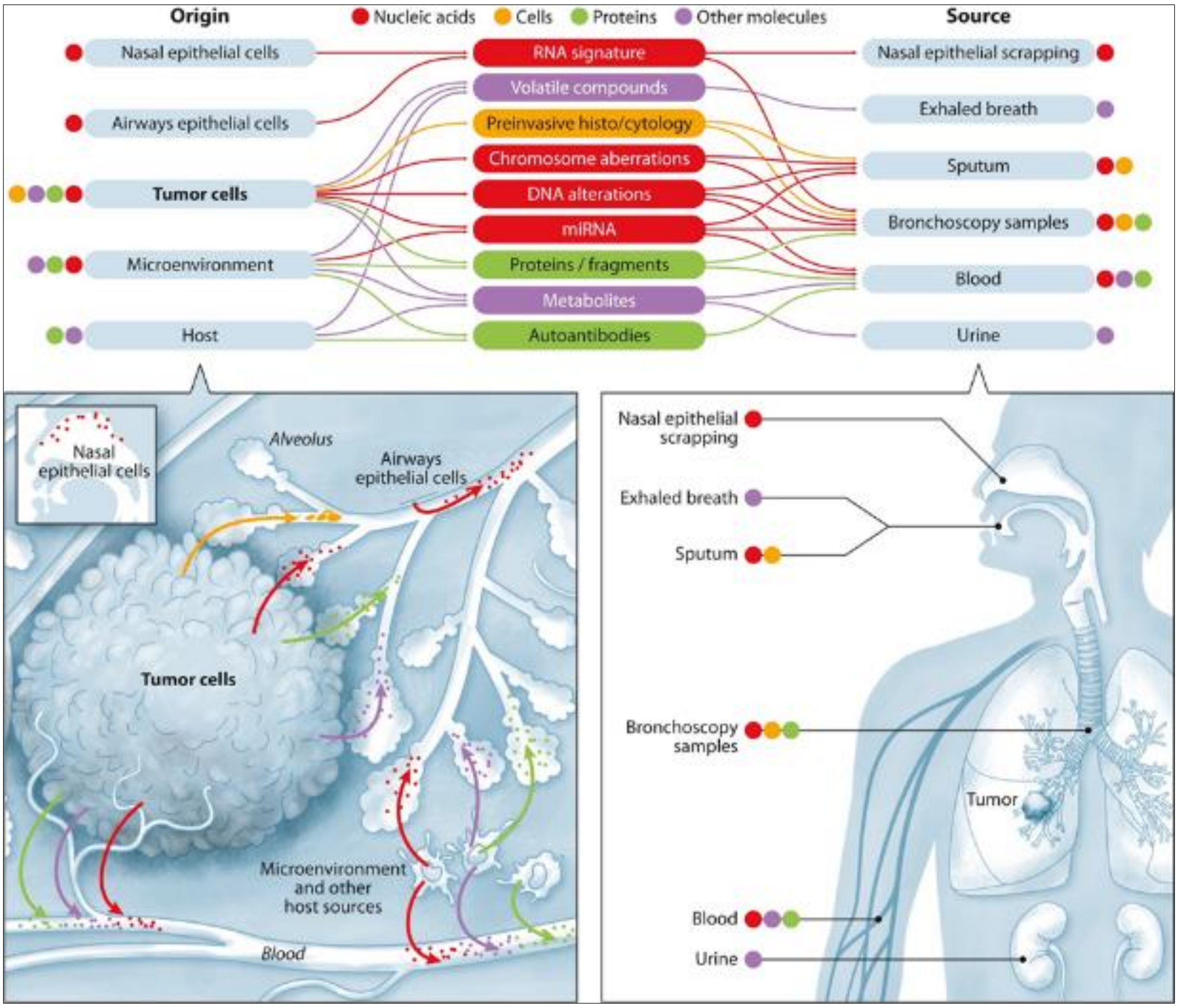

Figure 2 Current promising molecular biomarkers for lung cancer screening [12]

However, traditional methods of biomarker discovery are often limited by the complexity and heterogeneity of lung cancer. Lung cancer is not a single disease but rather a collection of different subtypes, each with distinct molecular characteristics. These variations make it challenging to identify biomarkers that are universally applicable across all patients or that can accurately predict disease outcomes in diverse populations [13, 14]. Intratumor heterogeneity can be caused by both intrinsic and extrinsic cell factors, ranging from genotypic alterations, epigenetic modification, signal transduction and plastic gene expression, as depicted in fig. 1 [13]. Moreover, the volume of data generated from high-throughput technologies such as genomics, proteomics, and imaging is enormous, and traditional statistical methods are often insufficient to handle such large datasets [15]. As a result, cancer research is increasingly moving towards the integration of multiple, large-scale data types.

Consequently, the use of machine learning (ML) models to automatically identify the internal patterns in different data types has emerged as a tool in biomarker discovery. ML is a subset of artificial intelligence (AI), dedicated to making predictions by recognising patterns in data through mathematical algorithms [16]. It refers to a set of computational techniques that allow computers to learn patterns from data and make predictions without being explicitly programmed [17]. These techniques can process vast amounts of data, identifying complex relationships that may not be apparent through traditional analytical methods. For many years, ML has been used as a supportive tool in cancer phenotyping and treatment. It has been extensively applied in advanced methods for early detection, cancer classification, signature extraction, tumour microenvironment (TME) deconvolution, prognosis forecasting, and drug response assessment [18-20].

The integration of ML into biomarker discovery represents a step forward in the fight against lung cancer. By automating the analysis of complex datasets and uncovering hidden patterns, ML has the potential to accelerate the identification of predictive biomarkers, which in turn can lead to earlier diagnosis, more precise treatment, and improved patient outcomes. This paper aims to review the current state of predictive biomarkers for lung cancer, with a particular focus on machine learning approaches. Through this review, this paper explores how machine learning has been applied to biomarker discovery, discuss recent advancements in the field, and consider the challenges and future directions for using machine learning in lung cancer research.

1.1. Predictive Biomarkers in Lung Cancer

Biomarkers are defined as measurable biological indicators that provide information about normal or pathological processes in the body [21]. According to the National Institutes of Health, a biomarker is 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes pathogenic processes, or pharmacologic responses to a therapeutic intervention [21]. In lung cancer, biomarkers signal the presence of cancer, predict its progression, and guide therapeutic interventions. These biomarkers are often derived from a range of biological sources such as tissue biopsies, blood samples, or imaging data, and they help clinicians make more informed decisions regarding diagnosis, prognosis, and treatment [22].

Predictive biomarkers for lung cancer can be broadly classified into several types:

- **Genetic Biomarkers**: These include specific gene mutations, amplifications, or deletions that can be directly linked to lung cancer development and progression [23]. For instance, as shown in Fig. 3, mutations in the *epidermal growth factor receptor* (EGFR) gene are present in a subset of non-small cell lung cancer (NSCLC) patients and are predictive of responsiveness to tyrosine kinase inhibitors (TKIs), a class of targeted therapies. Other well-known genetic biomarkers include *KRAS* mutations, which are more common in smokers and have implications for treatment resistance, and *ALK* gene rearrangements, which can be targeted by ALK inhibitors [24]. Advances in next-generation sequencing (NGS) technology have enabled the identification of multiple genetic alterations at once, providing a more comprehensive molecular profile of individual tumors.

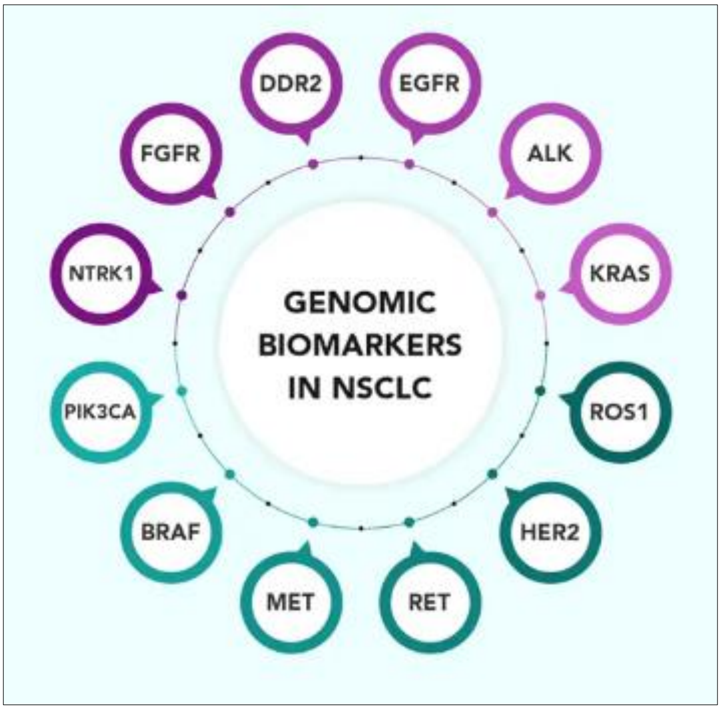

Figure 3 Genetic biomarkers in NSCLC [25]

- **Proteomic Biomarkers**: Early detection of lung cancer can be facilitated by identifying protein markers and autoantibodies specific to different cancer types. For instance, the EarlyCDT®-Lung test, applied to 1,613 patients in the United States, successfully identified stage I lung cancer with greater specificity than imaging methods, especially in asymptomatic patients [26]. Similar outcomes were reported in studies by Caroline J. Chapman, where autoantibodies were found to have comparable sensitivity and specificity for lung cancer detection [27]. These findings underscore the importance of autoantibodies as an enhancement to traditional diagnostic approaches. Currently, several biomarkers are used clinically for lung cancer detection, though many potential tumor-associated markers are still under investigation. For example, CEACAM, CYFRA 21-1, and ProGRP are used in serum-based testing, but their individual concentrations are often insufficient for early diagnosis. However, combining these markers enhances detection capabilities, particularly for adenocarcinoma [28, 29, 30]. Other studies have demonstrated significant diagnostic differences between lung cancer patients and healthy individuals using a combination of markers, including CA125, NY-ESO, and NSE (neuron-specific enolase) [31, 32].
- **Imaging Biomarkers**: Imaging biomarkers refer to structural or functional changes in the body, detected through technologies such as computed tomography (CT) scans, positron emission tomography (PET), or magnetic resonance imaging (MRI) [33]. In lung cancer, changes in tumor size, density, or metabolic activity serve as indicators of disease progression or response to treatment. For example, reductions in tumor size on CT scans after chemotherapy or targeted therapy serve as imaging biomarkers of treatment success. Similarly, PET scans that show decreased metabolic activity in tumors indicate a favorable response [33, 34]. Despite their utility, imaging biomarkers often require correlation with molecular or histological findings to provide a complete picture of disease status. Fig 4. Illustrates the common image modalities in lung cancer screening/detection.

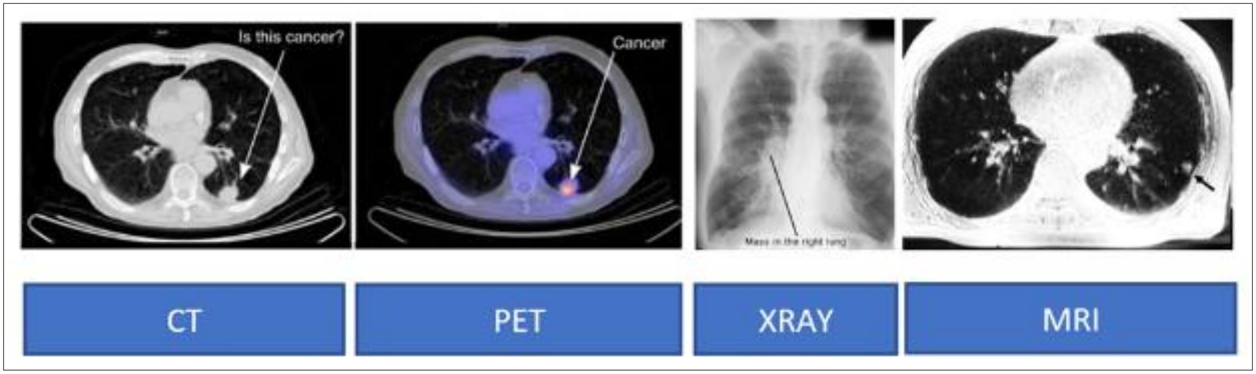

Figure 4 Common image modalities in lung cancer screening/detection

1.2. Limitations and Challenges

As noted above, a range of biomarkers has already been incorporated into the diagnosis and treatment of lung cancer. For instance, the aforementioned EGFR mutations and ALK rearrangements guide the use of targeted therapies, while PD-L1 expression helps in selecting patients for immunotherapy [24, 35]. Additionally, liquid biopsies—tests that detect tumor DNA or RNA in blood—are emerging as non-invasive alternatives to traditional tissue biopsies, providing a means of tracking tumor evolution and therapeutic resistance over time [36]. Despite these advances, limitations and challenges remain in biomarker development for lung cancer:

- **Variability across populations**: Genetic and environmental differences result in variability in biomarker expression across different populations. For instance, EGFR mutations are more common in East Asian populations and in non-smokers, while KRAS mutations are more prevalent in Western populations and smokers [37]. This variability poses challenges in ensuring that biomarkers are universally applicable across diverse demographic and ethnic groups.
- **Complexity of lung cancer subtypes**: Lung cancer is not a single disease but a collection of histologically and molecularly distinct subtypes. For instance, NSCLC accounts for about 85% of all lung cancers, but it includes several subtypes, such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Each of these subtypes has unique molecular characteristics, making it difficult to identify biomarkers that can reliably predict outcomes across all forms of lung cancer [38].
- **Dynamic nature of cancer evolution**: Cancer is a dynamic disease, constantly evolving in response to environmental pressures such as treatment. This makes it challenging to identify static biomarkers that can reliably predict outcomes over time. For example, a patient with an EGFR-mutated tumor may initially respond

to EGFR inhibitors, but over time, the tumor may develop secondary mutations (e.g., T790M mutation) that confer resistance to the treatment [39]. This dynamic nature warrants the need for continuous monitoring and updating of biomarker profiles during the course of the disease.

Given these challenges, traditional biomarker discovery methods are often insufficient to capture the complexity and heterogeneity of lung cancer.

2. Machine Learning in Biomarker Discovery

Machine learning (ML), a subset of AI, is revolutionizing biomarker discovery by enabling the analysis of vast, complex datasets that are characteristic of modern biological research. With respect to lung cancer, where data from genomics, proteomics, imaging, and clinical records can be overwhelming, traditional statistical approaches often fall short in capturing the patterns and relationships that exist in this high-dimensional data. ML, by contrast, is designed to handle such complexity and can uncover hidden insights that may lead to the identification of new biomarkers [40]. ML has been instrumental in cancer prediction and diagnosis by analyzing pathology profiles, imaging studies, and converting images into mathematical sequences (Fig 5). For instance, a model called ViT-Patch has proven effective in detecting malignancies and localizing tumors [41].

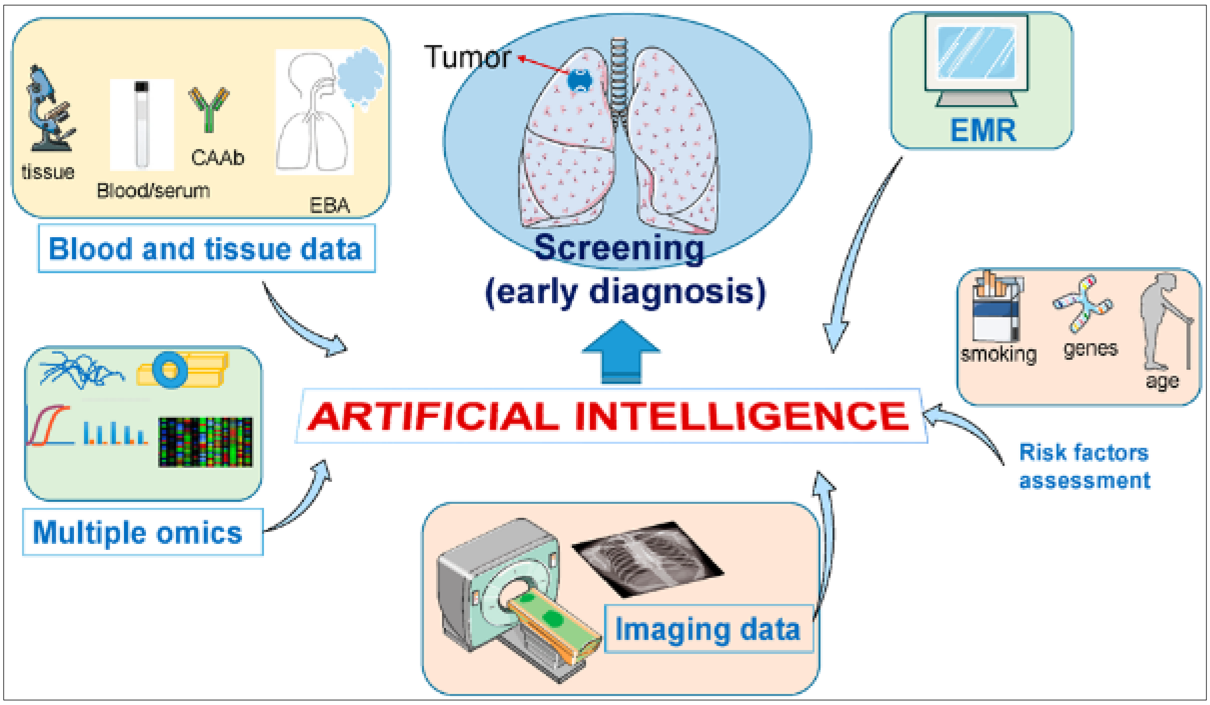

Figure 5 Overview of Artificial Intelligence for lung cancer diagnosis [42]

Several studies have applied various machine learning techniques to classify cancer data, with support vector machine classifiers, probabilistic neural networks, and K-nearest neighbours showing promising results. One study demonstrated that the random forest model achieved a 96% accuracy in detecting cancers [43]. Comparative research on different ML algorithms, including support vector machines, artificial neural networks (ANN), and Naive Bayes, revealed ANN as a reliable method for real-time cancer predictions. Additionally, principal component analysis was used to enhance prediction accuracy by reducing dimensionality [44]. Moreover, Pulse-Coupled Neural Networks have been used in image processing, with various neural network designs evaluated for cancer prediction accuracy [45].

2.1. Types of Machine Learning Techniques

ML techniques can be broadly divided into three categories, each with its own strengths and applications in biomarker discovery:

- **Supervised Learning**: This approach involves training models on labeled datasets, where the outcome (e.g., presence or absence of disease) is known. Supervised learning algorithms, such as support vector machines (SVMs), decision trees, and random forests, are particularly useful in classification tasks [46]. For example, in lung cancer biomarker discovery, supervised learning can be used to train a model on genomic data to predict whether a patient has cancer based on the presence of specific genetic mutations. The model learns the relationship between the input features (e.g., gene expression profiles) and the known outcomes, which allows it to make predictions on new, unseen data. In addition to classification, supervised learning is also used for regression tasks, where the goal is to predict a continuous outcome, such as the likelihood of treatment success or disease progression based on biomarker levels [40].

- **Unsupervised Learning**: Unlike supervised learning, unsupervised learning works with unlabeled data, meaning the model does not know the outcomes beforehand. Instead, it aims to identify patterns or groupings in the data [40]. Clustering algorithms like k-means, hierarchical clustering, or principal component analysis (PCA) are common unsupervised learning methods used to find subgroups of patients with similar molecular profiles [47]. In lung cancer, unsupervised learning can help to discover new cancer subtypes based on biomarker data, which may lead to more personalized treatment strategies [40]. This technique is especially valuable when working with large, complex datasets where it's not clear how the variables relate to one another.
- **Deep Learning**: A subset of machine learning, deep learning models are inspired by the structure of the human brain and consist of neural networks with multiple layers. These models excel at learning from raw, unstructured data, such as images or sequences of genomic data, and can automatically extract relevant features without the need for manual intervention. Convolutional neural networks (CNNs), for example, are widely used in medical imaging to detect tumor features in lung CT scans, while recurrent neural networks (RNNs) are useful for time-series data, such as monitoring changes in biomarker levels over time [48]. Deep learning models have shown promise in biomarker discovery because of their ability to process large-scale, high-dimensional datasets, but they often require substantial computational power and large datasets for training.

2.2. Applications of Machine Learning in Biomarker Discovery

Machine learning has a variety of applications in the discovery and validation of biomarkers for lung cancer, spanning across different tasks such as feature selection and classification. A demonstration of Machine Learning application, adapted from [49] is shown in Fig 6.

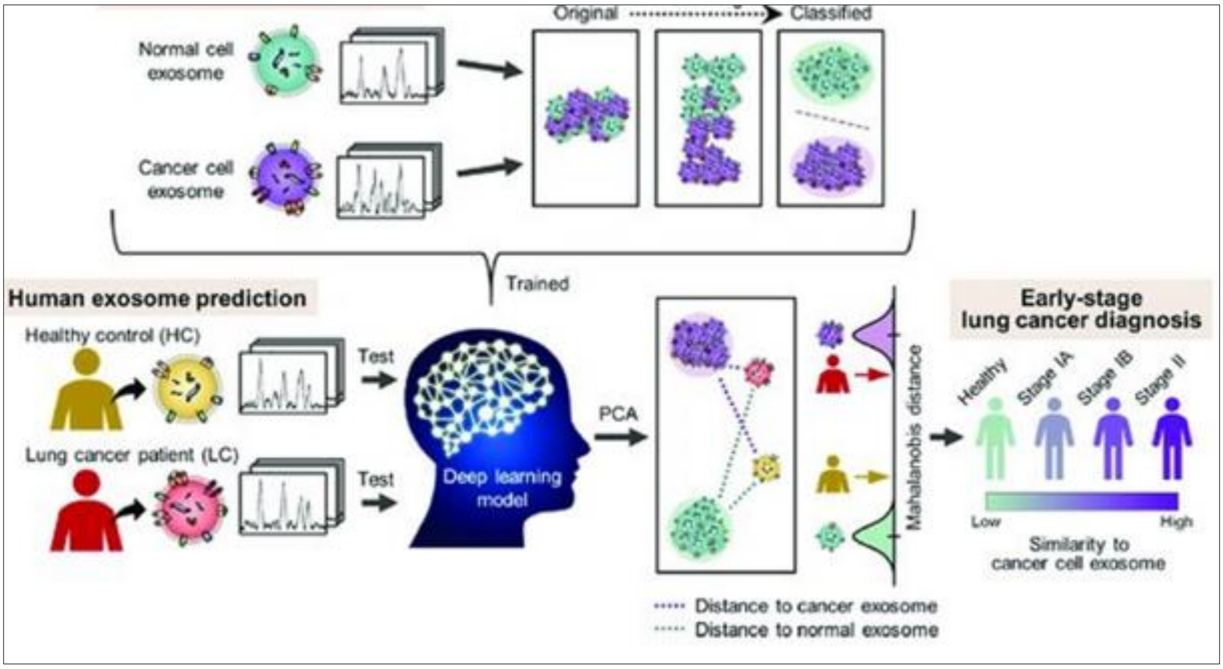

Figure 6 Applications of Machine Learning for lung cancer diagnosis [42]

- **Feature Selection**: One of the most crucial steps in biomarker discovery is identifying the most relevant features (e.g., genes, proteins, or imaging characteristics) from a large pool of potential candidates [50]. Machine learning algorithms, particularly those used in supervised learning, can rank and select the features that are most predictive of disease presence or treatment outcomes. Techniques such as random forests and support vector machines are commonly used for feature selection because they can handle large numbers of features and assess their importance in predicting an outcome [51]. Furthermore, deep learning algorithms such as CNN are common applications due to their advanced learning techniques. Feature selection helps in reducing the dimensionality of the data, making the model more interpretable and less prone to overfitting [50, 51].

- **Classification**: Classification tasks in biomarker discovery involve predicting a categorical outcome, such as whether a patient has lung cancer or not, based on biomarker data [52]. Machine learning models, such as decision trees SVMs, and deep learning, CNNs are commonly used for this purpose. In lung cancer research, classification models trained on genomic or proteomic data can help distinguish between cancerous and non-cancerous tissues or predict which patients are more likely to respond to a particular treatment based on their biomarker profiles [53].

2.3. Advantages of Machine Learning in Biomarker Discovery

The use of ML in biomarker discovery offers several advantages:

- **Handling High-Dimensional Data**: Lung cancer datasets are often high-dimensional, with thousands of genes, proteins, or imaging features being measured for each patient. ML algorithms, especially those in deep learning, are designed to handle these complex datasets and can identify patterns that may not be evident using traditional statistical methods [52].
- **Improved Accuracy and Precision**: ML models, once trained, can make highly accurate predictions, often outperforming traditional approaches. For example, in lung cancer, ML models have been shown to improve the prediction of treatment responses or the likelihood of disease recurrence based on a patient's molecular profile [54].
- **Identification of Complex Relationships**: Unlike traditional statistical methods, which may assume linear relationships between variables, ML models can capture complex, non-linear relationships between biomarkers and clinical outcomes [55]. This capability allows researchers to discover novel interactions and associations that may have been previously overlooked.

3. State of the Art in Predictive Biomarkers for Lung Cancer using Machine Learning

The application of ML in predictive biomarker discovery for lung cancer has evolved in recent years, driven by advances in computational power, access to high-dimensional data, and novel ML techniques. Lung cancer, being highly heterogeneous at the molecular and cellular levels, poses challenges for traditional biomarker identification methods. However, ML approaches have demonstrated the ability to analyse diverse datasets, extract complex patterns, and identify predictive biomarkers with greater accuracy [56]. This section explores the state-of-the-art machine learning methodologies used in lung cancer biomarker discovery, highlighting key studies and comparing the performance of various techniques.

3.1. Common Machine Learning Algorithms used for Lung Cancer Prediction

Random Forests: Random Forests (RF) are a widely-used ensemble learning method, with extensive applications in data mining and machine learning that have been employed in lung cancer biomarker research [57]. As a nonparametric, tree-based ensemble approach, RF combines the principles of adaptive nearest neighbours with bagging, allowing for effective data-adaptive inference. The bagging process, or bootstrap aggregating, involves training each decision tree in the forest on a random subset of the data (with replacement), which reduces variance and prevents overfitting.

Each tree in the random forest makes predictions, and the final prediction is the aggregation of these predictions. For classification tasks, RF predicts the class by taking a majority vote among the individual trees' predictions. Mathematically, for a given input *x*, the classification prediction from a random forest is:

$$\hat{y} = \text{mode}\ \{h_1(x), h_2(x), \ldots, h_T(x)\}$$

where $h_1(x)$ is the prediction of the *t*-th tree and *T* is the total number of trees. In regression tasks, RF outputs the average prediction across all trees:

$$\hat{y} = \left(\frac{1}{T}\right) \sum_{t=1}^{B} \text{h1}(x)$$

The greedy process of node splitting in each decision tree is performed step-by-step, where the algorithm selects the best split based on impurity measures like Gini impurity for classification or variance reduction for regression. The Gini impurity at node t is calculated as:

$$\mathrm{G}(t) = 1 - \sum_{k=1}^{K} P_k^2$$

where P_k represents the proportion of samples of class k in node t.

Additionally, the 'grouping property' of trees enables RF to manage correlations and interactions between variables effectively [58]. RF also offers a measure of variable importance, ranking and selecting variables by calculating how much a feature reduces impurity across all trees. This feature importance measure is particularly useful for analyzing high-dimensional genomic or proteomic data.

These features make RF particularly suitable for analysing genomic or proteomic data and conducting bioinformatics research [57]. Applications of RF in this field include prediction, variable selection, pathway analysis, genetic association studies, and epistasis detection. For instance, in a study aimed at identifying genetic mutations predictive of lung cancer, random forests were used to process genomic data from thousands of patients, identifying mutations in genes such as EGFR and TP53 as predictive biomarkers [59].

The Out-of-Bag (OOB) error provides an internal measure of model performance by using the samples not included in the bootstrap for evaluation. The OOB error is computed by averaging the errors over all out-of-bag samples, giving a robust estimate of the model’s generalization ability.

The versatility and robustness of random forests, particularly their ability to handle 'large p, small n' problems, make them a valuable tool in tackling complex datasets typical of genomic research.

- **Support Vector Machines (SVMs)**: Support Vector Machines (SVMs) are a type of supervised machine learning algorithm commonly used for classification and regression tasks [60]. The idea behind SVMs is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional space. By maximising the margin between the data points of different classes, SVMs aim to achieve classification, even in cases where the data is not linearly separable [61]. Mathematically, the decision boundary (hyperplane) is represented by the equation:

$$w.\, x+b = 0$$

Where w is the weight vector (normal to the hyperplane), x is the input feature vector, and b is the bias term. The goal of SVM is to maximize the margin, which is the distance between the hyperplane and the nearest data points from both classes, called support vectors. This is done by minimizing $\frac{1}{2}||w||^2$ subject to the constraint:

$$y_i\,(w.\, x+b) \geq 1,\ \forall_i$$

where y_i is the class label for data point x_i, ensuring that data points are correctly classified.

SVMs aim to achieve classification even in cases where the data is not linearly separable [61]. To handle non-linear data, SVMs use kernel functions to map the input data into a higher-dimensional space, where a linear hyperplane can effectively separate the classes [62]. In the dual form of the optimization problem, instead of directly working with input vectors, the dot products are replaced by kernel functions. Common kernel functions include the Radial Basis Function (RBF) and polynomial kernels, which allow SVMs to address non-linear relationships between features. The dual optimization problem becomes:

$$\min_{\alpha} \sum_{i=1}^{n} \alpha i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha i \alpha j y i y j K(xi, xj)$$

where $K(xi,xj)$ is the kernel function used to compute the dot product in the transformed space. These kernels make SVMs suitable for complex datasets like those in genomic research.

SVMs have been applied extensively in lung cancer research to classify tumor types or predict treatment responses. For example, SVMs can classify tumors as malignant or benign based on molecular and imaging biomarkers by identifying the optimal hyperplane that separates the data into classes [63]. In one study, SVMs were used to classify lung cancer subtypes based on microRNA expression profiles, achieving high accuracy in distinguishing between adenocarcinoma and squamous cell carcinoma [64]. The ability of SVMs to handle high-dimensional feature spaces is particularly useful in genomic data analysis, where the number of features often exceeds the number of samples [61].

- **Neural Networks and Deep Learning**: Neural Networks and Deep Learning: Neural networks, particularly deep learning models, have made strides in processing and analysing data types such as medical images, genomic sequences, and proteomic profiles. CNNs, a type of deep learning model, have been effective in analysing imaging data, such as CT and PET scans, to detect lung cancer nodules, predict tumor malignancy, and even estimate tumor progression [65]. CNNs excel in image processing tasks because they can automatically learn features from raw data, unlike traditional methods that require manual feature extraction.

The mathematical foundation of neural networks, including CNNs, lies in how each neuron in the network computes a weighted sum of the inputs, applies a bias, and then passes this sum through an activation function to produce an output. This process can be mathematically represented as:

$$y = f\left(\sum_{i=1}^{n} wixi + b\right)$$

where x_i represents the input features, w_i the associated weights, *b* the bias term, and *f* the activation function that introduces non-linearity into the model.

In the case of CNNs, they leverage specialized layers like convolutional layers, which apply filters to the input image, and pooling layers, which reduce the dimensionality of the data. The convolution operation in CNNs is mathematically defined as:

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(i+m, j+n)K\ (m,n)$$

where *S*(*i,j*) represents the output of the convolution at position (*i,j*), *I* is the input image, and *K* is the filter applied to the image [65]. This allows CNNs to capture spatial hierarchies and important features in the image.

For example, in a recent study, a deep learning model was trained on thousands of CT images to predict the likelihood of lung cancer progression, outperforming traditional radiological assessments [65]. These models have also been used to integrate imaging data with genomic information, improving the prediction of treatment responses based on a combination of visual and molecular biomarkers.

3.2. Recent Studies and Advancements in Lung Cancer Diagnosis Using Machine Learning

Machine Learning has significantly impacted the early detection and diagnosis of lung cancer, offering improvements in speed, accuracy, and precision. Recent advancements in deep learning algorithms, radiomics, and machine learning have demonstrated the potential to transform lung cancer diagnosis, aiding in earlier detection and improved patient outcomes. This section discusses key studies that illustrate the application of Machine Learning in lung cancer diagnosis, focusing on the models, datasets, and outcomes achieved in each case over the last 5-8 years.

One of the most notable studies is by Ardila et al. (2019), which developed a deep learning algorithm for the analysis of low-dose computed tomography (CT) scans to diagnose lung cancer. Their model demonstrated an AUC of 0.94, showing high diagnostic accuracy. This was a pivotal study as it showcased the capacity of AI models to handle large-scale clinical datasets and derive patterns that human eyes might overlook, which can be critical for early detection of lung cancer. The success of this study emphasized the importance of incorporating AI into routine lung cancer screenings, particularly in high-risk populations.

Radiomics has been explored in various studies for differentiating benign from malignant lung nodules. Delzell et al. (2019) utilized CT scan data of 200 lung nodules and applied radiomics to verify whether nodules were benign or malignant. Their model achieved an AUC of 0.72, demonstrating moderate success in classifying lung nodules. Although the model showed promise, its results highlight the continued need for refinement in AI algorithms for radiomics to further increase accuracy and reliability.

Schwyzer et al. (2018) applied a deep machine learning algorithm to fluorodeoxyglucose positron emission tomography (FDG-PET) imaging data to diagnose lung cancer, achieving a sensitivity of 95.9% and a specificity of 98.1%. The study underscores how AI can be effectively used with advanced imaging techniques, such as PET scans, to enhance diagnostic accuracy in lung cancer.

Sun et al. (2020) also used radiomics on CT scans of pure ground glass nodules from 385 patients to predict whether the nodules were invasive, achieving an AUC of 0.77. Radiomics models like these are particularly valuable for determining the aggressiveness of tumors, helping guide treatment decisions and improving patient outcomes.

Feng et al. (2019) employed a radiomics model on CT images of sub-solid nodules in 100 patients to differentiate between minimally invasive and invasive adenocarcinoma. Their model achieved an AUC of 0.912, highlighting the power of radiomics to make fine distinctions between cancer subtypes. Such tools are critical in tailoring treatments and improving survival rates for lung cancer patients.

Avanzo et al. (2020) used support vector machines (SVMs) on nodule data from low-dose CT scans to differentiate between adenocarcinoma and focal pneumonia, achieving an accuracy of 87.6%. This study demonstrated the potential for AI models to avoid misdiagnosis by distinguishing between cancerous and non-cancerous conditions with high precision.

Aydin et al. (2021) developed a convolutional neural network (CNN) model for classifying different lung cancer subtypes. Their model, which analyzed 301 lung cancer CT scans, demonstrated a sensitivity of 90% for distinguishing between squamous cell carcinoma, adenocarcinoma, and small cell carcinoma, though specificity was lower at 44%. This finding indicates that while AI can effectively classify cancer subtypes, there is room for improvement in distinguishing non-cancerous anomalies.

Teramoto et al. (2017) utilized conventional deep neural networks on histopathological images to classify lung cancer types, achieving classification accuracies ranging from 60% to 89% for adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. This study highlights the capability of AI to perform complex tissue classification tasks that would typically require expert pathologists.

Coudray et al. (2018) also employed conventional deep neural networks, specifically applying them to adenocarcinoma histopathology images to predict gene mutations. Their model achieved accuracy rates of 73.3%–85.6%, showcasing the role of AI in not only identifying cancer subtypes but also predicting underlying genetic mutations that can inform targeted therapies.

Flores-Fernandez et al. (2012) applied an artificial neural network model to a panel of serum biomarkers from 63 lung cancer patients, achieving a correct classification rate of 93.3%. This study underscores the promise of using AI to integrate biomarker data with imaging and histopathology to enhance diagnostic accuracy and predict patient outcomes.

Chen et al. (2020) used a radiomics approach to differentiate between small cell lung cancer (SCLC) and NSCLC in a dataset of 69 patients. The model achieved an AUC of 0.93, reflecting its strong ability to distinguish between these two critical subtypes of lung cancer. Such distinctions are crucial for determining the appropriate treatment, as SCLC and NSCLC have different therapeutic approaches and prognoses.

Yu et al. (2016) applied a combination of SVM and random forest models to histopathological images of lung adenocarcinoma, achieving an AUC of 0.81 for distinguishing malignant tumors from healthy tissue. By applying AI to histopathological data, researchers can enhance the precision of tissue classification, which is key for accurate diagnosis and treatment planning.

3.3. Recent Studies and Advancements in Lung Cancer Treatment using Machine Learning

AI advancements in lung cancer treatment have also progressed significantly, with various studies focusing on predicting treatment outcomes and supporting therapeutic decisions. Several AI-based models have shown promise in predicting responses to different types of lung cancer therapies.

Coroller et al. (2016) developed a radiomics-based model for predicting pathologic complete response to chemoradiation in non-small cell lung cancer (NSCLC) patients. Their study achieved an AUC of 0.61, indicating moderate performance in predicting therapeutic outcomes. Radiomics models were key in extracting quantitative imaging features that could assess patient response to treatments like chemoradiation.

Kureshi et al. (2016) focused on predicting responses to epidermal growth factor receptor-tyrosine kinase inhibitor (EGFR-TKI) therapy in NSCLC. They developed a radiomics-based model that showed an AUC of 0.76, highlighting its potential for improving precision medicine in lung cancer by targeting treatments more effectively.

Tian et al. (2021) combined radiomics and deep learning approaches to predict responses to PD-1 and PD-L1 immunotherapy in NSCLC patients. The model achieved an AUC of 0.71, suggesting that AI can play an essential role in personalizing immunotherapy treatments by identifying patients more likely to benefit from these therapies.

Liu et al. (2018) evaluated the feasibility of Watson for Oncology (WFO) in recommending treatments for NSCLC and small cell lung cancer (SCLC). Their study found that WFO showed consistency in 65.8% of cases and concordance of 92.4% with multidisciplinary teams (MDT). This highlights the potential of AI-driven tools like WFO to complement human decision-making in oncology settings.

Kim et al. (2020) further examined WFO and found strong concordance (92.4%) with MDT treatment recommendations for NSCLC and SCLC. Their findings support the role of AI systems in enhancing clinical decision-making and offering consistent treatment pathways.

Dercle et al. (2020) utilized a radiomics-based model to predict the sensitivity of NSCLC patients to treatments with nivolumab and docetaxel. Their models achieved AUCs of 0.77 and 0.67, respectively. This underscores the role of AI in evaluating treatment sensitivity, thereby enabling more tailored therapeutic interventions.

Zhang et al. (2021) focused on adenocarcinoma, applying radiomics models to predict EGFR mutations for targeted therapy. Their study demonstrated high predictive performance, with an AUC of 0.84. This finding indicates that AI can significantly enhance mutation-based treatment approaches in lung cancer.

Mu et al. (2020) used deep learning models to predict EGFR mutations in NSCLC patients for targeted therapy. Their model achieved an AUC of 0.83, further emphasizing the role of AI in supporting precision oncology by identifying genetic mutations critical for treatment planning.

3.4. Recent Studies and Advancements in Lung Cancer Detection Sequencing Data and Machine Learning

The role of AI in early detection of lung cancer has become increasingly vital, with a specific focus on machine learning (ML) models applied to sequencing data. These advancements use data derived from circulating tumor DNA (ctDNA), RNA sequencing (RNA-seq), and copy number variations (CNVs) to improve diagnostic accuracy. Below are several studies that demonstrate how AI and ML models are applied to sequencing data to detect and classify lung cancers.

Mathios et al. (2021) demonstrated a logistic regression (LR) model with a LASSO penalty for identifying lung cancer using cfDNA fragmentomes. Their study involved 799 samples and utilized 10-fold cross-validation, showing high performance with an AUC of 0.98. They successfully combined cfDNA fragmentation profiles with clinical risk factors and imaging features to detect lung cancer. However, DNA variations in late-stage diseases may affect cfDNA detection.

Chabon et al. (2020) explored the Lung-CLiP framework using 5-nearest neighbor (KNN), 3-nearest neighbor, naïve Bayes (NB), logistic regression (LR), and decision trees (DT) to detect lung cancer from cfDNA. Their model, based on 160 samples, demonstrated an AUC between 0.69 and 0.98. The study provided an early detection framework with feature selection incorporating single nucleotide variants (SNVs) and copy number variations (CNVs), but it faced a sampling bias due to high smoking rates among participants.

Liang et al. (2019) used logistic regression on 296 samples to identify lung cancer via ctDNA methylation markers. Their model, validated with 10-fold cross-validation, had an AUC of 0.816. The study featured nine DNA methylation markers, which imposed a limitation on the overall assay performance despite the model's high accuracy.

Whitney et al. (2015) developed a logistic regression model for RNA-seq data obtained from bronchial epithelial cells (BECs). Their model achieved an AUC of 0.81 with 10-fold cross-validation on 299 samples. Despite its high sensitivity for detecting small and peripheral lesions, the wide variance in selected genes poses a challenge for feature selection under different cohorts.

Podolsky et al. (2016) compared various models, including KNN, NB normal distribution of attributes, and DTs, using RNA-seq data for lung cancer detection. Based on 529 samples, the model achieved an AUC of 0.91 but faced overfitting risks, with performance variability observed across different datasets.

3.5. Recent Studies and Advancements in Lung Cancer Diagnosis Using Imaging Data and Machine Learning

McWilliams et al. (2013) developed a logistic regression (LR) model using nodule characteristics and clinical risk factors from CT images to estimate the malignancy risk of pulmonary nodules. The model achieved high accuracy, with an area under the curve (AUC) between 0.907 and 0.960. This model is particularly effective for small nodules (<10mm), although its predictive performance could be affected by variations in nodule characteristics and the subjective nature of nodule segmentation.

Van Riel et al. (2017) compared a computer-based malignancy risk model against human observers for estimating the risk of pulmonary nodules using CT scans. Their model achieved an AUC ranging from 0.706 to 0.932, performing comparably to radiologists. However, the model's reliance on binary classifications for smaller nodules presents challenges in refining diagnostic accuracy.

Kriegsmann et al. (2016) applied linear discriminant analysis (LDA) to matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry data to subtype non-small cell lung cancer (NSCLC). The model achieved high accuracy (0.991), maintaining performance across formalin-fixed paraffin-embedded (FFPE) biopsies. However, the model's effectiveness was hindered by the quality of stratification, as no benchmarking comparisons were conducted.

Buty et al. (2016) combined spherical harmonics and deep convolutional neural networks (DCNN) for characterizing lung nodules, integrating both shape and radiologist-provided nodule annotations. This model achieved AUCs between 0.793 and 0.824 across multiple test datasets. While the model showed significant accuracy improvements compared to previous benchmarks, arbitrary ground-truth scores may impact its generalizability.

Hussein et al. (2017) implemented a 3D CNN-based multi-task model to stratify lung nodules using CT imaging. Their model reached an accuracy of 0.9126, surpassing other models in several benchmarking tests. Despite its effectiveness, the model's performance is affected by the lack of standardized benchmarks, as some of the ground-truth data was arbitrarily defined by radiologists.

Khosravan et al. (2019) also employed a 3D CNN-based multi-task model but extended it by incorporating a sparse attentional model and eye-tracking for diagnosing lung nodules. This model demonstrated an accuracy of 0.97 for classification and a Dice similarity coefficient (DSC) of 0.91 for segmentation. However, segmentation accuracy decreased if the regions of interest extended outside of the lung regions, which might limit the model's robustness in detecting peripheral nodules.

Ciompi et al. (2015) proposed an ensemble of 2D views combined with CNNs for classifying peri-fissural nodules from CT images. Their model achieved an AUC of 0.868, performing well in distinguishing benign from malignant nodules. While the model's 3D CT volume features were a strength, its reliance on radiologist-identified positions and diameters limited the system's automatic generalization ability.

Venkadesh et al. (2021) leveraged a deep learning model combining 2D-ResNet50 and 3D-InceptionV1 to estimate malignancy risk in pulmonary nodules detected from low-dose CT screenings. The model achieved AUCs ranging from 0.86 to 0.96 and outperformed other models in malignancy risk estimation. One key limitation, however, was the need for precise nodule positioning in CT images, which poses challenges for cases with nodules that are difficult to locate.

Ardila et al. (2019) used Mask-RCNN and RetinaNet architectures for malignancy classification in lung cancer using CT images. Their deep learning model demonstrated a high AUC (0.944), outperforming human radiologists. Nevertheless, the study was based on data from a single cohort, limiting its generalizability across larger populations.

AbdulJabbar et al. (2020) proposed a multi-task CNN-based ensemble model for analyzing histological images of lung adenocarcinoma. Their model achieved a high classification accuracy of 0.9133, demonstrating robust segmentation of various cell types. However, the annotation quality in the reference dataset impacted the model's consistency across different institutions.

Coudray et al. (2018) used multi-task CNN models based on Inception-V3 for mutation prediction from histological slides in NSCLC patients. The model achieved AUC values between 0.733 and 0.856, but it faced limitations in classifying mutations related to certain genes, such as STK11 and EGFR, which were underrepresented in the training dataset.

Lin et al. (2021) employed a deep convolutional GAN (DCGAN) model combined with AlexNet to classify CT images of lung tumors. The model exhibited outstanding performance, achieving an accuracy of 0.9986. The use of GAN-generated synthetic data significantly improved the model's ability to generalize, though the lack of benchmarking comparisons with alternative models limited the assessment of its effectiveness.

Table 1 Summary of Recent Studies and Advancements Reviewed

Reference	Authors	Year	ML Methods	Dataset/Study Type	Predicted Outcome	Performance Metrics
[49]	Ardila et al.	2019	Deep Learning Algorithm	Low-dose CT Scan (NSCLC)	Diagnosis of Lung Cancer	AUC = 0.94
[66]	Delzell et al.	2019	Radiomics	CT scan of 200 lung nodules	Benign or Malignant Nodules Verification	AUC = 0.72
[67]	Schwyzer et al.	2018	Deep Machine Learning	FDG-PET Imaging	Diagnosis of Lung Cancer	Sensitivity = 95.9%, Specificity = 98.1%
[68]	Sun et al.	2020	Radiomics	CT scans of 385 patients with glass nodules	Invasiveness Prediction	AUC = 0.77
[69]	Feng et al.	2019	Radiomics	CT scans of 100 sub-solid nodules	Differentiate Minimally Invasive & Invasive Adenocarcinoma	AUC = 0.912
[70]	Avanzo et al.	2020	SVM	Nodules of Low-dose CT scans	Differentiate Adenocarcinoma from Pneumonia	Accuracy = 87.6%
[71]	Aydin et al.	2021	CNN	301 lung cancer CT scans	Squamous vs Adenocarcinoma, Small Cell	Sensitivity = 90%, Specificity = 44%
[75]	Chen et al.	2020	Radiomics	CT Radiomics of 69 lung cancer patients	Differentiate NSCLC from SCLC	AUC = 0.93
[76]	Yu et al.	2016	SVM, Random Forest	2480 histopathological images of lung adenocarcinoma	Distinguish malignant tumors from healthy tissue	AUC = 0.81
[72]	Teramoto et al.	2017	Conventional Deep Neural Networks	298 histopathological images	Classified Adenocarcinoma, Squamous, & Small Cell	Accuracy = 89%, 60%, 70%
[73]	Coudray et al.	2018	Deep Neural Networks	Pathological images of adenocarcinoma	Predicted 10 prevalent genes	Accuracy = 73.3%–85.6%
[74]	Flores-Fernandez et al.	2012	Artificial Neural Network Modeling	Serum biomarkers of 63 patients	Classifying Lung Cancer	Correct classification rate = 93.3%

[77]	Coroller et al.	2016	Radiomics-based Model	NSCLC	Predicting Pathologic Complete Response to Chemoradiation	AUC = 0.61
[78]	Kureshi et al.	2016	Radiomics-based Model	NSCLC	Predicting Response to EGFR-TKI Therapy	AUC = 0.76
[79]	Tian et al.	2021	Radiomics, Deep Learning	NSCLC	Predict Response to PD-1 and PD-L1 Immunotherapy	AUC = 0.71
[80]	Liu et al.	2018	Watson for Oncology (WFO)	NSCLC and SCLC	Feasibility in Treatment Recommendations	Consistency = 65.8%
[81]	Kim et al.	2020	Watson for Oncology (WFO)	NSCLC and SCLC	Treatment Concordance Between MDT and WFO	Concordance = 92.4%
[82]	Dercle et al.	2020	Radiomics-based Model	NSCLC	Treatment Sensitivity to Nivolumab & Docetaxel	AUC = 0.77 (Nivolumab), AUC = 0.67 (Docetaxel)
[83]	Zhang et al.	2021	Radiomics-based Model	Adenocarcinoma	Predicting EGFR Mutation	AUC = 0.84
[84]	Mu et al.	2020	Deep Learning Model	NSCLC	Predicting EGFR Mutation	AUC = 0.83
[90]	McWilliams et al.	2013	Logistic Regression (LR)	CT Imaging Data	Lung Nodule Malignancy	AUC = 0.907–0.960
[91]	Van Riel et al.	2017	Logistic Regression (LR)	CT Imaging Data	Malignancy Risk Estimation of Pulmonary Nodules	AUC = 0.706–0.932
[92]	Kriegsmann et al.	2016	Linear Discriminant Analysis (LDA)	Mass Spectra from MALDI Imaging	Subtyping Non-Small Cell Lung Cancer (NSCLC)	Accuracy = 0.991
[93]	Buty et al.	2016	Spherical Harmonics, DCNN	CT Imaging Data	Lung Nodule Malignancy Characterization	AUC = 0.793–0.824
[94]	Hussein et al.	2017	3D CNN-based Multi-Task Model	CT Imaging Data	Risk Stratification of Lung Nodules	Accuracy = 0.9126
[95]	Khosravan et al.	2019	3D CNN-based Multi-Task Model	CT Imaging Data	Lung Nodule Segmentation and Classification	Segmentation DSC = 0.91, Accuracy = 0.97
[96]	Ciompi et al.	2015	SVM, Random Forest (RF)	CT Imaging Data	Classification of Pulmonary Nodules	AUC = 0.868
[97]	Venkadesh et al.	2021	2D-ResNet50, 3D-Inception-V1	CT Imaging Data	Malignancy Risk Estimation of Pulmonary Nodules	AUC = 0.86–0.96

[98]	Ardila et al.	2019	Mask-RCNN, RetinaNet, 3D-Inception-V1	CT Imaging Data	Lung Cancer Screening	AUC = 0.944
[99]	AbdulJabbar et al.	2020	Micro-Net, 3D SC-CNN	Histological Images	Annotation of Cell Types in Single-Cell Level	Accuracy = 0.913
[100]	Coudray et al.	2018	Inception-V3, CNN-based Model	Histological Images	Lung Cancer Mutation Classification	AUC = 0.733–0.856
[101]	Lin et al.	2021	DCGAN, AlexNet	CT Imaging Data	Lung Cancer Classification	Accuracy = 0.9986

4. Challenges

While machine learning has shown great promise in the discovery of predictive biomarkers for lung cancer, several challenges still hinder its full potential in clinical applications. These challenges span across data quality, model performance, interpretability, and the integration of diverse data types.

4.1. Data Quality and Availability

One of the challenges in using ML for biomarker discovery is the quality and availability of data. ML models, especially deep learning algorithms, require large, high-quality, and well-annotated datasets to perform effectively. In lung cancer biomarker discovery, these datasets must include diverse sources such as genomic, proteomic, and imaging data, along with accurate clinical annotations. However, obtaining such data presents several difficulties:

- **Incomplete or Missing Data**: Clinical datasets frequently contain gaps or incomplete information due to factors such as patient dropout, inconsistent data collection methods, or errors during data entry. These missing data points can result in biased predictions and diminish the effectiveness of machine learning models [102]. Additionally, ongoing concerns about patient privacy and data security, particularly in light of breaches involving major organisations, add to the controversy surrounding data collection. For instance, patient confidentiality limits data availability, which restricts model training and prevents the full potential of the model from being realised [54].
- **Small Sample Sizes**: In many cases, particularly in rare lung cancer subtypes, the number of available samples is limited. Small datasets increase the risk of overfitting, where models perform well on training data but fail to generalise to new, unseen data. This issue is especially pronounced in deep learning models, which require large amounts of data for training [103].
- **Data Imbalance**: In lung cancer datasets, certain classes (e.g., specific genetic mutations or tumor types) may be overrepresented, while others are underrepresented. This imbalance can skew ML models toward the majority class, reducing the accuracy of predictions for minority classes. For example, a model trained to detect common mutations like *EGFR* may not perform as well on rare mutations due to the lack of sufficient data [59].

4.2. Overfitting and Model Generalizability

Overfitting is another challenge in ML, particularly when working with high-dimensional data such as genomics or proteomics, where the number of features often exceeds the number of samples. Models that overfit capture noise and irrelevant patterns in the training data, leading to poor generalization when applied to new datasets. Deep learning models, which are inherently complex, are especially prone to overfitting [104].

- **Cross-validation and Regularization**: While techniques like cross-validation and regularization are commonly used to mitigate overfitting, they may not always be sufficient when dealing with highly variable or limited datasets [105]. Improving generalization requires the development of more robust models that can learn underlying patterns in the data without being influenced by noise or irrelevant features.
- **Transfer Learning and Data Augmentation**: One potential solution to the problem of small datasets and overfitting is the use of transfer learning, where models pre-trained on large datasets are fine-tuned on smaller, domain-specific datasets. Additionally, data augmentation techniques, such as generating synthetic data to simulate rare cases, could help increase the diversity and volume of training data, improving model performance on real-world lung cancer datasets.

4.3. Model Interpretability

ML models, particularly deep learning algorithms, are often criticized for their lack of interpretability. These models function as 'black boxes,' where predictions are made based on complex, multi-layered computations that are difficult to trace or understand [106]. This opacity poses challenge in clinical settings, where explainability is important for decision-making and trust.

- **Explainable AI (XAI)**: Recent efforts in the field of explainable AI (XAI) aim to address this issue by developing techniques that make machine learning models more transparent and interpretable. For example, methods such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) help in interpreting how individual features contribute to a model's predictions [107]. However, further research is needed to integrate these tools into clinical workflows, ensuring that machine learning models provide actionable insights for clinicians.
- **Feature Importance and Visualisation**: Enhancing interpretability also involves identifying which features, such as specific genetic mutations or imaging patterns, are driving the model's predictions. Visualisation techniques, like heatmaps in medical imaging, can highlight the regions of interest that the model focuses on, aiding clinicians in understanding and validating the model's outputs.

4.4. Integration of Multi-Omics and Clinical Data

Lung cancer is a complex and heterogeneous disease, influenced by a multitude of genetic, proteomic, and environmental factors. Traditional approaches to biomarker discovery often focus on single data types, such as genomic mutations or imaging features. However, integrating multi-omics data (e.g., genomics, transcriptomics, proteomics) with clinical features (e.g., patient history, comorbidities) can provide a more comprehensive understanding of disease mechanisms and improve biomarker accuracy.

- **Multi-Modal Machine Learning**: Integrating multi-omics and clinical data requires the development of multi-modal ML approaches capable of processing and analysing diverse data types simultaneously. By combining data from multiple sources, such models can uncover complex interactions between biological processes and disease outcomes, leading to the discovery of more robust and reliable biomarkers.
- **Challenges in Data Integration**: While multi-omics integration holds promise, it also presents challenges, including data standardisation, dimensionality reduction, and managing missing or inconsistent data across different sources. Effective data harmonisation techniques and algorithms capable of handling heterogeneous data are needed to maximise the potential of multi-omics analysis in biomarker discovery.

5. Future Directions

To advance the use of ML in lung cancer biomarker discovery, several future directions should be considered:

- **Improved Data Collection and Sharing**: Initiatives to improve data quality, standardisation, and sharing across institutions and research groups are essential for building large, high-quality datasets. Collaborative efforts, such as cancer registries and multi-center clinical trials, can provide more diverse and comprehensive datasets for ML models.
- **Advances in Model Interpretability**: Continued research into explainable AI techniques is needed to bridge the gap between ML predictions and clinical applicability. Models that provide clear, interpretable insights into how biomarkers are identified and how predictions are made will be crucial for integrating ML into routine clinical practice.
- **Personalised Medicine and Precision Oncology**: ML has the potential to drive the development of personalised medicine, where treatments are tailored to the specific molecular profile of an individual's tumor. More efforts need to go into this direction to take application of ML in lung cancer to the next level.

6. Conclusion

This paper has examined the state of predictive biomarker discovery for lung cancer, with a particular focus on the role of ML. ML techniques have demonstrated potential in processing vast amounts of genomic, proteomic, and imaging data, revealing novel biomarkers that can enhance the early diagnosis of lung cancer and optimise treatment approaches. These technologies are not only capable of identifying patterns in complex datasets but also offer opportunities to refine personalised treatment plans, thereby improving patient outcomes. However, despite the promise shown by ML, several challenges have been highlighted that must be addressed to fully realise its potential in clinical practice. Data quality is a concern, as the accuracy of ML models is heavily dependent on the availability of well-curated, annotated datasets. Additionally, issues such as model overfitting, where models perform well on training data but fail to generalise to new, unseen cases, remain barriers. The interpretability of ML models also poses a challenge, as the complexity of many algorithms, particularly deep learning models, can make it difficult for clinicians to understand how predictions are made, reducing the trust and transparency necessary for clinical integration.

Looking ahead, future research should concentrate on refining ML models to enhance their robustness and reliability while also investing significantly into application of AI for personalized medicine and precision oncology. This includes developing methods to mitigate overfitting, improving data preprocessing techniques, and creating more interpretable models that provide clear, actionable insights for healthcare professionals. Additionally, efforts should be intensified to integrate diverse data sources—such as multi-omics datasets, imaging data, and clinical features—to create more comprehensive and accurate predictive biomarkers. Moreover, the clinical validation of biomarkers discovered through ML is essential to ensure their applicability in real-world settings. Rigorous validation across diverse patient populations and clinical environments will be necessary to confirm the efficacy and reliability of these biomarkers before they can be widely adopted in practice.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declared that there is no conflict of interest.

References

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020 (https://gco.iarc.fr/today, accessed February 2021).

[2] De Martel C, Georges D, Bray F, Ferlay J, Clifford GM. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health. 2020;8(2):e180-e190.

[3] World Cancer Research Fund. (202). Lung cancer statistics. https://www.wcrf.org/cancer-trends/lung-cancer-statistics/

[4] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: a cancer journal for clinicians. 2024 Jan 1;74(1).

[5] Ashique S, Bhowmick M, Pal R, Khatoon H, Kumar P, Sharma H, Garg A, Kumar S, Das U. Multi drug resistance in colorectal cancer-approaches to overcome, advancements and future success. Advances in Cancer Biology-Metastasis. 2024 Jan 12:100114.

[6] Brockley LJ, Souza VG, Forder A, Pewarchuk ME, Erkan M, Telkar N, Benard K, Trejo J, Stewart MD, Stewart GL, Reis PP. Sequence-based platforms for discovering biomarkers in liquid biopsy of non-small-cell lung cancer. Cancers. 2023 Apr 13;15(8):2275.

[7] Labrie M, Brugge JS, Mills GB, Zervantonakis IK. Therapy resistance: opportunities created by adaptive responses to targeted therapies in cancer. Nature reviews Cancer. 2022 Jun;22(6):323-39.

[8] Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, Jones L, Hamblin A, Arumugam P, Chan G, Chubb D, Noyvert B. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Nature Medicine. 2024 Jan;30(1):279-89.

[9] Lee JH, Hwang EJ, Kim H, Park CM. A narrative review of deep learning applications in lung cancer research: from screening to prognostication. Translational Lung Cancer Research. 2022 Jun;11(6):1217.

[10] Li C, Wang H, Jiang Y, Fu W, Liu X, Zhong R, Cheng B, Zhu F, Xiang Y, He J, Liang W. Advances in lung cancer screening and early detection. Cancer biology & medicine. 2022 May 5;19(5):591.

[11] Ahmad A, Imran M, Ahsan H. Biomarkers as biomedical bioindicators: approaches and techniques for the detection, analysis, and validation of novel Biomarkers of diseases. Pharmaceutics. 2023 May 31;15(6):1630.

[12] Seijo, L. M., Peled, N., Ajona, D., Boeri, M., Field, J. K., Sozzi, G., Pio, R., Zulueta, J. J., Spira, A., Massion, P. P., Mazzone, P. J., & Montuenga, L. M. (2019). Biomarkers in lung cancer screening: Achievements, promises, and challenges. Journal of Thoracic Oncology, 14(3), 343–357. https://doi.org/10.1016/j.jtho.2018.11.023

[13] Sun, Xx., Yu, Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. Acta Pharmacol Sin 36, 1219–1227 (2015). https://doi.org/10.1038/aps.2015.92.

[14] Edsjö A, Holmquist L, Geoerger B, Nowak F, Gomon G, Alix-Panabières C, Staaf J, Ploeger C, Lassen U, Le Tourneau C, Lehtiö J. Precision cancer medicine: Concepts, current practice, and future developments. Journal of Internal Medicine. 2023 Oct;294(4):455-81.

[15] Punetha A, Kotiya D. Advancements in oncoproteomics technologies: Treading toward translation into clinical practice. Proteomes. 2023 Jan 10;11(1):2.

[16] Lee M. Deep learning techniques with genomic data in cancer prognosis: a comprehensive review of the 2021–2023 literature. Biology. 2023 Jun 21;12(7):893.

[17] Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. Electronics. 2023 Apr 10;12(8):1789.

[18] González-Castro L, Chávez M, Duflot P, Bleret V, Martin AG, Zobel M, Nateqi J, Lin S, Pazos-Arias JJ, Del Fiol G, López-Nores M. Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records. Cancers. 2023 May 13;15(10):2741.

[19] Yaqoob A, Musheer Aziz R, verma NK. Applications and techniques of machine learning in cancer classification: A systematic review. Human-Centric Intelligent Systems. 2023 Dec;3(4):588-615.

[20] Lotter W, Hassett MJ, Schultz N, Kehl KL, Van Allen EM, Cerami E. Artificial Intelligence in Oncology: Current Landscape, Challenges, and Future Directions. Cancer Discovery. 2024 May 1;14(5):711-26.

[21] Ahmad A, Imran M, Ahsan H. Biomarkers as biomedical bioindicators: approaches and techniques for the detection, analysis, and validation of novel Biomarkers of diseases. Pharmaceutics. 2023 May 31;15(6):1630.

[22] Li C, Wang H, Jiang Y, Fu W, Liu X, Zhong R, Cheng B, Zhu F, Xiang Y, He J, Liang W. Advances in lung cancer screening and early detection. Cancer biology & medicine. 2022 May 5;19(5):591.

[23] Quinn ZL, Barta JA, Johnson JM. Molecular lung cancer: How targeted therapies and personalized medicine are re-defining cancer care. The American Journal of the Medical Sciences. 2022 Oct 1;364(4):371-8.

[24] Zwierenga F, van Veggel BA, van den Berg A, Groen HJ, Zhang L, Groves MR, Kok K, Smit EF, Hiltermann TJ, de Langen AJ, van der Wekken AJ. A comprehensive overview of the heterogeneity of EGFR exon 20 variants in NSCLC and (pre) clinical activity to currently available treatments. Cancer Treatment Reviews. 2023 Sep 19:102628.

[25] Villalobos P, Wistuba II. Lung Cancer Biomarkers. Hematol Oncol Clin North Am. 2017 Feb;31(1):13-29.

[26] Jett, J., et al. (2014). Audit of the autoantibody test, EarlyCDT®-Lung, in 1600 patients: An evaluation of its performance in routine clinical practice. Lung Cancer, 83, 51–55.

[27] Chapman, C. J., et al. (2011). Immunobiomarkers in small cell lung cancer: Potential early cancer signals. Clinical Cancer Research, 17(5), 1474–1480.

[28] Arya, S., & Bhansali, S. (2011). Lung cancer and its early detection using biomarker-based biosensors. Chemical Reviews, 111, 6783–6809

[29] Okamura, K., et al. (2013). Diagnostic value of CEA and CYFRA 21-1 tumor markers in primary lung cancer. Lung Cancer, 80, 45–49.

[30] Kim, H., et al. (2011). Plasma ProGRP concentration is sensitive and specific for discriminating small cell lung cancer from nonmalignant conditions or non-small cell lung cancer. Journal of Korean Medical Science, 26, 625–630.

[31] Zamay, G. S., et al. (2016). DNA aptamers for characterization of histological structure of lung adenocarcinoma. Molecular Therapy – Nucleic Acids, 6, 150–162.

[32] Howard, B., et al. (2003). Identification and validation of a potential lung cancer serum biomarker detected by matrix-assisted laser desorption/ionization-time of flight spectra analysis. Proteomics, 3, 1720–1724.

[33] Yilmaz D, Sharp PS, Main MJ, Simpson PB. Advanced molecular imaging for the characterisation of complex medicines. Drug Discovery Today. 2022 Jun 1;27(6):1716-23.

[34] Chauvie S, Mazzoni LN, O'Doherty J. A Review on the Use of Imaging Biomarkers in Oncology Clinical Trials: Quality Assurance Strategies for Technical Validation. Tomography. 2023 Oct 17;9(5):1876-902.

[35] Purkayastha K, Dhar R, Pethusamy K, Srivastava T, Shankar A, Rath GK, Karmakar S. The issues and challenges with cancer biomarkers. Journal of Cancer Research and Therapeutics. 2023 Jan 1;19(Suppl 1):S20-35.

[36] Armakolas A, Kotsari M, Koskinas J. Liquid biopsies, novel approaches and future directions. Cancers. 2023 Mar 3;15(5):1579.

[37] Laguna JC, García-Pardo M, Alessi J, Barrios C, Singh N, Al-Shamsi HO, Loong H, Ferriol M, Recondo G, Mezquita L. Geographic differences in lung cancer: focus on carcinogens, genetic predisposition, and molecular epidemiology. Therapeutic Advances in Medical Oncology. 2024 Mar;16.

[38] Guo H, Zhang J, Qin C, Yan H, Liu T, Hu H, Tang S, Tang S, Zhou H. Biomarker-targeted therapies in non–small cell lung cancer: current status and perspectives. Cells. 2022 Oct 12;11(20):3200.

[39] Johnson M, Garassino MC, Mok T, Mitsudomi T. Treatment strategies and outcomes for patients with EGFR-mutant non-small cell lung cancer resistant to EGFR tyrosine kinase inhibitors: focus on novel therapies. Lung Cancer. 2022 Aug 1;170:41-51.

[40] Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell and Tissue Research. 2023 Oct;394(1):17-31.

[41] Feng H, Yang B, Wang J, Liu M, Yin L, Zheng W, Yin Z, Liu C. Identifying malignant breast ultrasound images using ViT-patch. Applied Sciences. 2023 Mar 9;13(6):3489.

[42] Espinoza JL, Dong LT. Artificial Intelligence Tools for Refining Lung Cancer Screening. Journal of Clinical Medicine. 2020; 9(12):3860. https://doi.org/10.3390/jcm9123860

[43] Dubey C, Shukla N, Kumar D, Singh AK, Dwivedi VK. Breast cancer modeling and prediction combining machine learning and artificial neural network approaches. In2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) 2022 Nov 4 (pp. 119-124).

[44] Wankhade Y, Toutam S, Thakre K, Kalbande K, Thakre P. Machine learning approach for breast cancer prediction: A review. In2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) 2023 May 4 (pp. 566-570).

[45] Liu H, Liu M, Li D, Zheng W, Yin L, Wang R. Recent advances in pulse-coupled neural networks with applications in image processing. Electronics. 2022 Oct 11;11(20):3264.

[46] Mahto R, Ahmed SU, Rahman RU, Aziz RM, Roy P, Mallik S, Li A, Shah MA. A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. BMC bioinformatics. 2023 Dec 15;24(1):479..

[47] Gu Y, Wang M, Gong Y, Li X, Wang Z, Wang Y, Jiang S, Zhang D, Li C. Unveiling breast cancer risk profiles: a survival clustering analysis empowered by an online web application. Future Oncology. 2023 Dec 1;19(40):2651-67.

[48] Holdsworth J, Scapicchio M. What is deep learning? 2024 June 17; Retrieved from: https://www.ibm.com/topics/deep-learning

[49] Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., & Tse, D. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25, 954-961.

[50] Liu Y, Zhang H, Xu Y, Liu YZ, Al-Adra DP, Yeh MM, Zhang Z. Five Critical Gene-Based Biomarkers With Optimal Performance for Hepatocellular Carcinoma. Cancer Informatics. 2023 Aug;22:11769351231190477.

[51] Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics. 2022 Jun 27;2:927312.

[52] Al-Tashi Q, Saad MB, Muneer A, Qureshi R, Mirjalili S, Sheshadri A, Le X, Vokes NI, Zhang J, Wu J. Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. International journal of molecular sciences. 2023 Apr 24;24(9):7781.

[53] Abbasian MH, Ardekani AM, Sobhani N, Roudi R. The role of genomics and proteomics in lung cancer early detection and treatment. Cancers. 2022 Oct 20;14(20):5144.

[54] Zhang B, Shi H, Wang H. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. Journal of multidisciplinary healthcare. 2023 Dec 31:1779-91.

[55] Li Z, Zeng T, Zhou C, Chen Y, Yin W. A prognostic signature model for unveiling tumor progression in lung adenocarcinoma. Frontiers in Oncology. 2022 Nov 1;12:1019442.

[56] Ladbury C, Amini A, Govindarajan A, Mambetsariev I, Raz DJ, Massarelli E, Williams T, Rodin A, Salgia R. Integration of artificial intelligence in lung cancer: Rise of the machine. Cell Reports Medicine. 2023 Feb 21;4(2).

[57] Jain N, Jana PK. LRF: A logically randomized forest algorithm for classification and regression problems. Expert Systems with Applications. 2023 Mar 1;213:119225.

[58] Scornet E. Trees, forests, and impurity-based variable importance in regression. InAnnales de l'Institut Henri Poincare (B) Probabilites et statistiques 2023 Feb (Vol. 59, No. 1, pp. 21-52). Institut Henri Poincaré.

[59] Ahmadi A, Mohammadnejadi E, Razzaghi-Asl N. Gefitinib derivatives and drug-resistance: A perspective from molecular dynamics simulations. Computers in Biology and Medicine. 2023 Sep 1;163:107204.

[60] Guido R, Ferrisi S, Lofaro D, Conforti D. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. Information. 2024 Apr 19;15(4):235.

[61] Mahto R, Ahmed SU, Rahman RU, Aziz RM, Roy P, Mallik S, Li A, Shah MA. A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. BMC bioinformatics. 2023 Dec 15;24(1):479.

[62] Chaudhary JK, Sharma H, Tadiboina SN, Singh R, Khan MS, Garg A. Applications of Machine Learning in Viral Disease Diagnosis. In2023 10th International Conference on Computing for Sustainable Global Development (INDIACom) 2023 Mar 15 (pp. 1167-1172).

[63] Maleki N, Niaki ST. An intelligent algorithm for lung cancer diagnosis using extracted features from Computerized Tomography images. Healthcare Analytics. 2023 Nov 1;3:100150.

[64] Shafat Z, Ahmed MM, Almajhdi FN, Hussain T, Parveen S, Ahmed A. Identification of the key miRNAs and genes associated with the regulation of non-small cell lung cancer: a network-based approach. Genes. 2022 Jun 29;13(7):1174.

[65] Thanoon MA, Zulkifley MA, Mohd Zainuri MA, Abdani SR. A review of deep learning techniques for lung cancer screening and diagnosis based on CT images. Diagnostics. 2023 Aug 8;13(16):2617.

[66] Delzell, D.A.P., Magnuson, S., Peter, T., Smith, M., & Smith, B.J. (2019). Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Frontiers in Oncology*, 9, 1393.

[67] Schwyzer, M., Ferraro, D.A., Muehlematter, U.J., Curioni-Fontecedro, A., Huellner, M.W., von Schulthess, G.K., & Burger, I.A. (2018). Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks—Initial results. *Lung Cancer*, 126, 170-173.

[68] Sun, Y., Li, C., Jin, L., Gao, P., Zhao, W., Ma, W., & Tan, M. (2020). Radiomics for lung adenocarcinoma manifesting as pure ground-glass nodules: Invasive prediction. *European Radiology*, 30, 3650-3659.

[69] Feng, B., Chen, X., Chen, Y., Li, Z., Hao, Y., Zhang, C., & Liao, Y. (2019). Differentiating minimally invasive and invasive adenocarcinomas in patients with solitary sub-solid pulmonary nodules with a radiomics nomogram. Clinical Radiology, 74, 570.e1-570.e11.

[70] Avanzo, M., Stancanello, J., Pirrone, G., & Sartor, G. (2020). Radiomics and deep learning in lung cancer. Strahlentherapie und Onkologie, 196(10), 879-887.

[71] Aydin, N., Çelik, Ö., Aslan, A.F., Odabaş, A., & Şahin, M.C. (2021). Detection of lung cancer on computed tomography using artificial intelligence applications developed by deep learning methods and the contribution of deep learning to the classification of lung carcinoma. Current Medical Imaging, 17(9), 1137-1141.

[72] Teramoto, A., Tsukamoto, T., Kiriyama, Y., & Fujita, H. (2017). Automated classification of lung cancer types from cytological images using deep convolutional neural networks. Biomedical Research International, 2017, 4067832.

[73] Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature Medicine, 24(10), 1559-1567.

[74] Flores-Fernández, J.M., Herrera-López, E.J., Sánchez-Llamas, F., Rojas-Calvillo, A., Cabrera-Galeana, P.A., Leal-Pacheco, G., González-Palomar, M.G., Femat, R., & Martínez-Velázquez, M. (2012). Development of an optimized multi-biomarker panel for the detection of lung cancer based on principal component analysis and artificial neural network modeling. Expert Systems with Applications, 39(10), 10851-10856.

[75] Chen, B.T., Chen, Z., Ye, N., Mambetsariev, I., Fricke, J., Daniel, E., Wang, G., Wong, C.W., Rockne, R.C., & Colen, R.R. (2020). Differentiating peripherally located small cell lung cancer from non-small cell lung cancer using a CT radiomic approach. Frontiers in Oncology, 10, 593.

[76] Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Re, C., Rubin, D.L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communications, 7, 12474.

[77] Coroller, T.P., Agrawal, V., Narayan, V., Hou, Y., Grossmann, P., Lee, S.W., Mak, R.H., & Aerts, H.J. (2016). Radiomic phenotype features predict pathological response in non-small cell lung cancer. Radiotherapy and Oncology, 119(3), 480-486.

[78] Kureshi, N., Abidi, S.S., & Blouin, C. (2016). A predictive model for personalized therapeutic interventions in non-small cell lung cancer. IEEE Journal of Biomedical and Health Informatics, 20(2), 424-431.

[79] Tian, P., He, B., Mu, W., Liu, K., Liu, L., Zeng, H., Liu, Y., Jiang, L., Zhou, P., & Huang, Z. (2021). Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. Theranostics, 11(5), 2098-2107.

[80] Liu, C., Liu, X., Wu, F., Xie, M., & Feng, Y. (2018). Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: A feasibility study. Journal of Medical Internet Research, 20(5), e11087.

[81] Kim, M.S., Park, H.Y., Kho, B.G., Park, C.K., Oh, I.J., Kim, Y.C., Kim, S., Yun, J.S., Song, S.Y., & Na, K.J. (2020). Artificial intelligence and lung cancer treatment decision: Agreement with the recommendation of multidisciplinary tumor board. Translational Lung Cancer Research, 9(2), 507-514.

[82] Dercle, L., Fronheiser, M., Lu, L., Du, S., Hayes, W., Leung, D.K., Roy, A., Wilkerson, J., Guo, P., Fojo, A.T., & et al. (2018). Identification of non-small cell lung cancer sensitive to systemic cancer therapies using radiomics. Clinical Cancer Research, 26(8), 2151-2162.

[83] Zhang, G., Cao, Y., Zhang, J., Ren, J., Zhao, Z., Zhang, X., Li, S., Deng, L., & Zhou, J. (2021). Predicting EGFR mutation status in lung adenocarcinoma: Development and validation of a computed tomography-based radiomics signature. American Journal of Cancer Research, 11(3), 546.

[84] Mu, W., Jiang, L., Zhang, J., Shi, Y., Gray, J.E., Tunali, I., Gao, C., Sun, Y., Tian, J., Zhao, X., & et al. (2020). Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nature Communications, 11, 5228.

[85] Mathios, D., Johansen, J.S., Cristiano, S., Medina, J.E., Phallen, J., Larsen, K.R., et al. (2021). Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nature Communications, 12, 5060.

[86] Chabon, J.J., Hamilton, E.G., Kurtz, D.M., Esfahani, M.S., Moding, E.J., Stehr, H., et al. (2020). Integrating genomic features for non-invasive early lung cancer detection. Nature, 580, 245–251.

[87] Liang, W., Zhao, Y., Huang, W., Gao, Y., Xu, W., Tao, J., et al. (2019). Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). Theranostics, 9(9), 2056–2070.

[88] Whitney, D.H., Elashoff, M.R., Porta-Smith, K., Gower, A.C., Vachani, A., Ferguson, J.S., et al. (2015). Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. *BMC Medical Genomics*, 8(1), 18.

[89] Podolsky, M.D., Barchuk, A.A., Kuznetcov, V.I., Gusarova, N.F., Gaidukov, V.S., Tarakanov, S.A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, 17(2), 835-838.

[90] McWilliams, A., Tammemagi, M.C., Mayo, J.R., Roberts, H., Liu, G., Soghrati, K., et al. (2013). Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med., 369, 910–919.

[91] Van Riel, S.J., Ciompi, F., Wille, M.M.W., Dirksen, A., Lam, S., Scholten, E.T., et al. (2017). Malignancy risk estimation of pulmonary nodules in screening CTs: comparison between a computer model and human observers. PLoS One, 12, e0185032.

[92] Kriegsmann, M., Casadonte, R., Kriegsmann, J., Dienemann, H., Schirmacher, P., Kobarg, J.H., et al. (2016). Reliable entity subtyping in non-small cell lung cancer by matrix-assisted laser desorption/ionization imaging mass spectrometry on formalin-fixed paraffin-embedded tissue specimens. Mol Cell Proteomics, 15, 3081–3089.

[93] Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., & Mollura, D. (2016). Characterization of lung nodule malignancy using hybrid shape and appearance features. In Medical image computing and computer-assisted intervention (pp. 662–670). Springer, Cham.

[94] Hussein, S., Cao, K., Song, Q., & Bagci, U. (2017). Risk stratification of lung nodules using 3D CNN-based multi-task learning. In Information processing in medical imaging (pp. 249–260). Springer, Cham.

[95] Khosravan, N., Celik, H., Turkbey, B., Jones, E.C., Wood, B., & Bagci, U. (2019). A collaborative computer-aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. Med Image Anal., 51, 101–115.

[96] Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., et al. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med Image Anal., 26, 195–202.

[97] Venkadesh, K.V., Setio, A.A.A., Schreuder, A., Scholten, E.T., Chung, K.M., Wille, M.M.W., et al. (2021). Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. Radiology, 300, 438–447.

[98] Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med., 25, 954–961.

[99] AbdulJabbar, K., Raza, S.E.A., Rosenthal, R., Jamal-Hanjani, M., Veeriah, S., Akarca, A., et al. (2020). Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. Nat Med., 26, 1054–1062.

[100] Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. J Thorac Oncol., 13, S562.

[101] Lin, C.H., Lin, C.J., Li, Y.C., & Wang, S.H. (2021). Using generative adversarial networks and parameter optimization of convolutional neural networks for lung tumor classification. Appl Sci., 11, 480.

[102] Foser S, Maiese K, Digumarthy SR, Puig-Butille JA, Rebhan C. Looking to the future of early detection in cancer: liquid biopsies, imaging, and artificial intelligence. Clinical Chemistry. 2024 Jan;70(1):27-32.

[103] Baltatzis V. *Deep Learning for Lung Cancer Detection: An Analysis of the Effects of Imperfect Data and Model Biases* (Doctoral dissertation, King's College London, 2023).

[104] Ge Y, Nie Q, Huang Y, Liu Y, Wang C, Zheng F, Li W, Duan L. Beyond prototypes: Semantic anchor regularization for better representation learning. InProceedings of the AAAI Conference on Artificial Intelligence 2024 Mar 24 (Vol. 38, No. 3, pp. 1887-1895).

[105] Bishop CM, Bishop H. Deep learning: Foundations and concepts. Springer Nature; 2023 Nov 1.

[106] Räz T. ML interpretability: Simple isn't easy. Studies in history and philosophy of science. 2024 Feb 1;103:159-67.

[107] Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Computers in Biology and Medicine. 2023 Oct 4:107555.