A Review of Advances in Computational Predictive Modeling for Lung Cancer Diagnosis

Bhoomika Modi[1], Dr. Divyangna Gandhi[2], Dr. Sheetal Pandya[3]

[1]Research Scholar, CE Department, IITE, Indus University, Ahmedabad, Gujarat, India.
Email: Bhoomikamodi.21.rs@indusuni.ac.in
[2]Assistant Professor, EC Department, IITE, Indus University, Ahmedabad, Gujarat, India.
Email: divyangnagandhi.ec@indusuni.ac.in
[3]Associate Professor, Ganpat University. Email: Sheetal.rkcet@gmail.com

Abstract

Lung cancer is one of the world's leading causes of cancer-related death due to the challenges associated with late-stage diagnosis and the difficulty associated with accurately identifying malignant lung lesions in a timely manner. As a result, there has been extensive development of various computational predictive modeling approaches that employ the latest advancements in computational and machine learning techniques (including deep learning, radiomics, etc.), as well as the use of multiple-modality data sources for accurate diagnosis of lung cancer via medical imaging, clinical data and biological data. This review provides a summary of recent research and developments surrounding computational predictive modeling of lung cancer diagnosis, specifically emphasizing the use of deep learning models, transformer networks and multi-modal frameworks due to their unique ability to automatically extract complex features, provide context globally and integrate multiple-source data for improved diagnostic accuracy. A systematic synthesis of the currently available methodologies, datasets and evaluation strategies will be presented by analyzing the advantages these approaches have in regard to being accurately scalable, transferable to the clinical setting, and applicable clinically. This review aims to provide new information concerning current strengths and weaknesses while providing insight into the most useful approaches available at this time for the early detection of lung cancer, thus assisting with better decision-making in clinical settings.

Keywords

Lung cancer diagnosis, computational predictive modeling, machine learning, deep learning, radiomics, convolutional neural networks, transformer models, multimodal learning, medical imaging, risk prediction.

How to cite this article: Modi B, Gandhi D, Pandya S. A Review of Advances in Computational Predictive Modeling for Lung Cancer Diagnosis. Int J Drug Deliv Technol. 2026;16(65s): 1311-1321. DOI: 10.25258/ijddt.16.65s.135

1. Introduction

As one of the greatest public health concerns worldwide, lung cancer accounts for a large percentage of cancer deaths globally. Improvements in medical imaging systems, screening, and treatment methods continue to be made, yet the overall survivorship associated with this disease remains low. The primary reason for this is late-stage diagnosis of the disease, which is associated with reduced treatment options and the effectiveness of those treatment options. Early detection of lung cancer has a significant impact on survival; however, early detection due to small differences in radiographic imaging or similarities between benign and malignant lesions makes them difficult to identify clinically. The standard clinical process of providing a diagnosis, through radiographic imaging and examination, places a significant reliance upon the evaluation performed by the physician in providing a diagnosis to the patient, which introduces a level of subjectivity, inter-observer variation and the potential for delays in providing a diagnosis. Therefore, there is an increasing need for reliable, objective, and data-driven diagnostic processes to assist in the timely and accurate diagnosis of lung cancer [1],[2].

1.1 Clinical Context and Diagnostic Challenges

Lung carcinoma is separated on a clinical basis into two major types: small cell (SCLC) and non-small cell (NSCLC). The majority of all lung cancer cases are NSCLC. Different histological types of NSCLC and SCLC demonstrate unique tumor biology, tumor progression, and imaging features; thus, these characteristics must be understood and taken into consideration when developing diagnostic and treatment strategies. Early detection of lung cancer is very important since it will dramatically improve the chance of a cure and the long-term survival of a patient. Accurately identifying small pulmonary nodules and determining whether they are benign or malignant is particularly challenging in routine patient evaluations. The primary reasons for this challenge are: variability in imaging studies performed, variation in scanning protocols, differences in patients' anatomy, and the inherent heterogeneity of tumors from one patient to another. Consequently, when small pulmonary nodules are not identified or not properly classified, there is often a delay in diagnosis and less effective treatment. Thus, there is a need for advanced computational methods to provide standardized, accurate, and scalable diagnostic assistance for a wide variety of clinical scenarios [3], [4], [5].

1.2 Role of Computational Predictive Modeling

By building on these clinical challenges, computational predictive modeling has become a revolutionary method for improving how we accurately and efficiently diagnose lung cancer. Through the use of machine learning, deep learning, and artificial intelligence, this model will analyze vast amounts of medical imaging and clinical data in a systematic data-driven way; unlike traditional methods that heavily rely upon the human interpretation of data, computational models have the ability to learn patterns, identify subtle differences between them, and learn about the different types of complex, relational connections that exist within the data to the point where these connections would not otherwise be apparent to clinicians.

A wide range of diagnostic tasks, such as identifying pulmonary nodules, determining benign vs malignant tumours, studying disease progression, and matching patient-specific risk to outcomes, have been performed using these methodologies. This technique allows for greater accuracy and consistency when making decisions based on multidimensional data sets and has significantly reduced variability and helped clinicians manage complicated cases. For instance, automated systems may assist in examining large data sets, identifying potentially concerning cases, and supplying quantitative measures that complement clinical expertise. The rapid growth of digital ecosystems in healthcare has contributed to an even greater expansion of the use of predictive modelling methods. The growth of annotated data sets, improved computational infrastructures, and increased sophistication in algorithmic design have made it possible the creation of very robust and scalable models. As a result, computational predictive modeling methods are being more widely integrated into the diagnostic workflow as tools to assist clinicians by increasing diagnostic accuracy and efficiency, rather than replacing them [6], [7].

The movement toward more data-driven forms of diagnosis is a development that addresses many of the shortcomings associated with conventional methods. With their ability to connect different levels of analysis with the complexity of clinical data, computational models represent an excellent means of achieving more precise, consistent, and timely identification of lung cancer, leading to higher rates of success for patients diagnosed and treated for the disease, as well as improving overall success rates of healthcare delivery [8].

1.3 Organization of the Paper

In order to create a systematic framework for understanding the topic, the paper is structured as follows. The Literature Survey in Section 2 will provide an analysis of recent and relevant literature in chronological order that discusses methodologies or approaches, data sets, results, and future trends. The System Design in Section 3 explains how the taxonomy of computation techniques is integrated into a diagnostic framework and considers both advantages/disadvantages and practical implications of using them. Finally, Section 4 provides a summary of the main points as well as an overview of the overall impact and importance of computational predictive modeling on the diagnosis of lung cancer.

2. Literature Survey

To comprehend advances in computational predictive modeling of lung cancer diagnosis must be done through the literature in a systematic format. In this section, selected studies will be systematically reviewed for their methodology, database, outcomes and limitations. The literature will be presented chronologically, with the most recently published first to demonstrate a current view of research in the subject. In addition, the searches performed, the criteria for selection of studies, and expected emerging research directions will also be integrated into the literature analysis in order to maintain the analytical flow. Studies were selected based upon their relevance to the diagnostics of lung cancer, the use of some type of computational modeling technique, and lastly, the date of publication. Priority for consideration was given to studies published between 2022 to 2025; therefore, the review presented will provide a synopsis of the latest works in machine learning, deep learning and multi-modal frameworks within this number of years. There are a limited number of studies prior to 2022 included in this literature review; these studies have been included to provide a basis for understanding the early studies and thus the development of predictive modeling methodologies through the years. Studies related to

Treatment Planning or Drug Discovery / Development have not been included to promote a focused diagnostic perspective.

Pati et al. [9] provided a comprehensive overview of recent advances in computational biology for lung cancer, emphasizing the integration of biological insights with computational models to enhance disease understanding. Building on this, Wang et al. [10] introduced the Nodule-Plus R-CNN framework for 3D pulmonary nodule segmentation, demonstrating improved detection accuracy using deep learning, although computational cost remains a limitation. Similarly, Levi et al. [11] developed a machine learning model using electronic medical records, highlighting the significance of structured clinical data in improving prediction performance, especially in risk-based diagnostic systems. Expanding the perspective, Bouchaffra et al. [12] and Shariff et al. [13] explored deep learning and medical imaging techniques, demonstrating how convolutional neural networks significantly improve feature extraction and early detection capabilities. Jamil et al. [14] further surveyed computational diagnostic techniques, emphasizing the growing need for hybrid and integrated frameworks that combine multiple learning paradigms.

Huang et al. [15] and Yang et al. [16] analyzed broader applications of artificial intelligence in lung cancer diagnosis, highlighting both current advancements and future challenges, particularly in interpretability and clinical integration. Arshad et al. [17] extended this work by focusing on hybrid architectures combining CNNs with transformer-based models, demonstrating improved contextual understanding and diagnostic accuracy, although challenges related to explainability persist. Supporting the importance of computational infrastructure, Cai et al. [18] discussed available biological data resources and computational tools that facilitate research advancements, while Bertolaccini et al. [19] reviewed recent clinical and technological progress in lung cancer detection. In parallel, Osamika et al. [20] explored predictive analytics using big data, showing that large-scale datasets significantly enhance model robustness and scalability.

According to Sweet et al. [21], there was a comparative analysis of machine learning techniques and the overall result showed that although ensemble models gave a better overall performance than individual classifiers based on accuracy (how well they performed) and robustness (consistency of performance), ensemble models are very dependant (heavily reliant) upon the quality of the features they were using. Gupta and Gupta [22] commented that the efficiency of conventional machine learning was good when applied to structured datasets, but conventional machine learning techniques would have limitations when applied to more complicated datasets (imaging data). Li et al. [23] examined the wide-ranging applicability of AI and machine learning into disease diagnosis, treatment, and prognosis, thus showing that AI's versatility can be applied to various areas within a health care system. Gayap and Ahloufi [24] stated that CNNs (convolutional neural networks) extracted hierarchical characteristics from CT scans more effectively, producing superior diagnostic accuracy, and the issues associated with this are the high compute resources needed, as well as the low interpretability of CNNs. Furthermore, Alavinejad et al. [25] mentioned the value of connecting artificial intelligence with smart nanomedicine techniques for lung cancer and pointed out that AI can potentially enhance diagnostic–treatment linkage and the development of personal health care strategy system. However, clinical validation (proving it either works or does not) has been limited, making it a very difficult area to establish as another confirmed AI technique in clinical practice.

Dritsas and Trigka [26] highlighted the role of traditional machine learning in lung cancer risk prediction, forming the foundation for modern predictive models, while Rao et al. [27] proposed an AI-based early detection system, achieving promising results but facing limitations in dataset size and generalizability. Mamun et al. [28] demonstrated that ensemble learning significantly improves classification performance compared to individual models. Asif et al. [29] further emphasized advancements in machine learning and deep learning in medical diagnostics, noting improvements in performance alongside challenges such as computational cost and data imbalance. Tortora et al. [30] introduced a multimodal learning framework integrating imaging and pathological data, demonstrating improved disease characterization and predictive performance, though integration complexity remains a concern. Sun et al. [31] showed that incorporating clinical variables into predictive models enhances decision-making accuracy, while Guan et al. [32] leveraged big data analytics to improve scalability and robustness of diagnostic systems, although issues related to data standardization and real-world deployment persist.

2.1 Search Strategy and Study Selection

A structured literature search strategy was applied to gather current and thorough analyses of different computational predictive modeling methods for lung cancer diagnosis. Academically verified databases and scientifically recognized repositories were utilized to identify relevant research based on whether they contained information regarding the diagnostic use of machine learning, deep learning, or other forms of artificial intelligence and included articles dated 2022-2025 to include the most recent advances in the field. Some earlier studies were also selected to give a background to the most current methods of research.

After filtering for relevant publications based on search terms including lung cancer diagnostic, machine learning, deep learning, radiomics, and multimodal learning, a strict definition of the inclusion criteria was created to ensure that only studies that developed or tested a computational model specifically for diagnostic use were included in the review. This included nodule detection, classification, and prediction. Any study that focused on areas unrelated to diagnostic use, such as treatment planning or drug development, was excluded so that studies would remain clearly defined to the scope of this review. To be included in this review, studies that reported performance measures, experimental validation, or comparative analysis were given priority above studies that did not report those and were therefore less methodologically sound or practically relevant.

2.2 Emerging Trends and Observations

The existing literature clearly shows that there has been a substantial shift in research interests away from the traditional machine learning methods that have been employed for many years, towards more advanced and integrated computational frameworks. Deep learning models, especially convolutional neural networks, have rapidly gained momentum in health care applications and have become the de facto standard for diagnosing lung cancer via medical imaging, due to their ability to perform automated extraction of complex hierarchical features from the medical imaging data. The movement toward these advanced models represents the increasing demand for models capable of addressing the limitations of traditional high-dimensional models.

The development of "hybrid" architectures that are based on convolutional neural networks and transformer-based models is also a major trend. These hybrid models attempt to utilize the advantages of local feature extraction and global contextual understanding to enhance their performance in diagnosis. Hybrid models are especially useful in complex scenarios where both spatial and contextual information are required for correct classification. Furthermore, there is an increasing emphasis on multimodal learning frameworks that bring together different types of artefacts such as imaging, clinical records, and pathological data in one integrated framework in order to provide a more complete picture of disease characteristics. Although the use of multimodal learning models has been associated with improvements in robustness and reductions in uncertainty, the integration, alignment, and computational complexity of multimodal learning models create challenges for future development.

Recent research emphasizes that models should scale and work well also in implementing processes for clinical workflows through large datasets, as well as emphasizing interpretability so clinicians can trust systems, use the models and incorporate into day-to-day patient care. The emerging trends indicate the transition away from performance-based research to research that focuses on usability in clinical settings has reflected the maturity of this space and its readiness for implementation across real-world health care systems.

Table 1: Comparative Analysis of Recent Studies

Sr. No	Study (Year)	Method & Modality	Dataset & Task	Performance & Key Insight	Limitation
1	Arshad et al. (2025) [17]	AI (CNN + Transformer)	Multi-dataset; Diagnosis	Hybrid models improve contextual learning	Limited explainability
2	Cai et al. (2022) [18]	Computational Biology	Multi-source data; Analysis	Provides foundational computational insights	Limited direct diagnostic modeling
3	Bertolaccini et al. (2024) [19]	Clinical + AI	Clinical studies; Diagnosis	Advances in treatment-linked diagnostics	Limited computational validation
4	Osamika et al. (2023) [20]	Predictive Analytics1	Big data; Risk prediction	Big data improves prediction capability	Data inconsistency
5	Sweet et al.	Ensemble ML	Tabular	Ensemble	Feature

	(2024) [21]		data; Prediction	improves robustness	dependency
6	Gupta et al. (2023) [22]	ML model	Structured dataset	Effective classification in controlled data	Not suitable for imaging
7	Li et al. (2022) [23]	ML/DL	Multi-domain; Diagnosis	Integrated ML improves prediction	Generalization issues
8	Gayap et al. (2024) [24]	CNN (Imaging)	CT scans; Detection	CNN improves feature extraction accuracy	Requires large labelled data
9	Alavinejad et al. (2025) [25]	AI + Nanomedicine	Biomedical data	AI enhances diagnostic-treatment link	Limited clinical validation
10	Dritsas et al. (2022) [26]	ML baseline	Risk dataset	Foundational predictive modeling	Lower performance
11	Rao et al. (2023) [27]	AI-based model	Clinical dataset; Prediction	Good classification accuracy	Limited generalization

3. System Design

As we reflect on what we learned from the literature review, it is now necessary to analyze how the computational predictive models are designed and employed in real-world lung cancer diagnostic systems. While the previous section provided an analysis of past research and their associated methodologies, this section will provide a consolidated overview of how each of those approaches fits into a systematic format. The ultimate goal here is bridging the gap between the theoretical advances and their implementation in practice by explaining the overall system design, computational technique taxonomy, dataset used, as well as the technical characteristics of each type of model.

A typical workflow for computationally predicting lung cancer diagnosis is a series of systematic steps that combine many steps, including: data collection, preparation, extraction of features from the data, training the model, and finally making a prediction. Each step has an impact on how accurate and reliable the diagnostic system will be. In addition to the algorithm used to build the diagnostic system, the quality of the input data used to train the model, as well as the capacity to combine various forms of information, will impact the success of these systems.

There are many segments in this section organized in order to give an overall, clear picture. The first segment of this section covers the entire system framework, including a complete list of steps to follow in creating an overall diagnostic model from start to finish. Next will be a listing of the many computational structures, e.g., Machine Learning, Deep Learning, Multimodal, etc., describing their characteristics; these models will be further divided into their various approaches. Then, listed are several types of datasets that are typically used when training classification or regression models; this indicates the importance of data in creating predictive models. Finally, we will compare all the computational methods noted above and review their strengths, weaknesses, and practical applications for each of these models when used in a practical way within our society today.

3.1 Overall System Framework

In order to translate the theoretical advancements in the literature into feasible solutions, it is essential to have knowledge and understanding of how computational predictive models are utilized in a structured diagnostic environment. Most of the current methods for diagnosing lung cancer utilize a well-defined workflow, comprising multiple phases that will allow for the conversion of raw medical data into clinically useful predictions. The diagnostic process itself typically starts with the capture of data, where raw input data is collected from numerous sources such as CT scans, clinical records, or multimodal datasets. CT imaging is the most frequently utilized imaging technique, due to the impact of the discovery of pulmonary nodules being evident and the ability of CT to identify fine and non-fine structural details of lung tissues [24], [32]. Based on the advanced diagnostic frameworks, the diagnostic process may also include clinical and pathologic data to help improve the diagnostic accuracy and provide a more comprehensive evaluation.

Data acquisition is followed by another step called the "pre-processing" stage, which is very important because it will improve the quality of the data and help make samples more consistent across different types of input (data). Common tasks completed during pre-processing are: noise removal; normalization; increasing the contrast; and rescaling of images. Pre-processing is essential to reduce variability introduced into the data from differences in imaging conditions and prepares all of the data collected to be ready for further analysis. The use of inconsistent pre-processing techniques will have a significant impact on both model performance and reproducibility, based on very recent studies [21],[31].

Many systems develop segmentations of suspicious lung regions or nodules after pre-processing. Afterwards, the system will identify the location of an ROI. This is critical in imaging as the model can then only focus on relevant areas and ignore the background noise of the scan. Advanced deep learning models like CNNs are frequently leveraged for automated segmentations, thereby improving the efficiency and accuracy of the identification of potential abnormalities [24]. Once segmentations are completed and relevant areas are identified, feature extraction and feature representation can commence. With regard to traditional machine learning approaches, handcrafted features such as textural features, shape features, and intensity features can be extracted from the ill structure and used for a classification task. With regard to deep learning models, they are capable of automatically learning hierarchical representations of the complex patterns that exist in the raw data; therefore, a much better representation of the underlying signal is created. The automated process of identifying and extracting features has led to substantial improvements in diagnostic performance in recent years [21], [32].

Model training as well as generating predictions occurs with the extracted features through the use of computational algorithms and the creation of a diagnostic output. Depending on the method, this may involve machine learning classifiers, deep neural networks or hybrid models that use many methods. An example of output from this process would include classification outputs (benign/malignant), probability of classification and risk prediction for clinicians to use in making decisions regarding treatment options. The final outcome of the system will provide a diagnostic output, which may be incorporated into the clinical workflow to support radiologists and other health care professionals in their daily practice. In more advanced systems, the diagnostic output may be combined with visualisation tools or decision support systems to aid in providing clearer and easier-to-use systems in clinical practice [30].

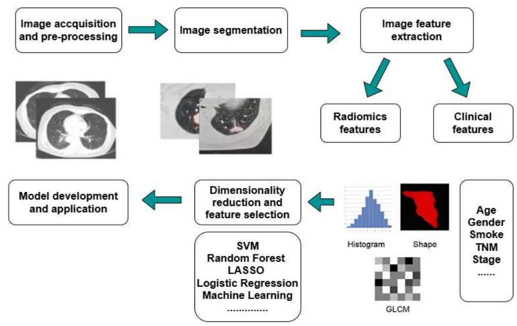

Figure 1: Workflow of lung cancer diagnostic system from data acquisition to prediction [24], [32]

The computational lung cancer diagnosis system is shown in Figure 1, with all the steps from data collection through prediction listed in order. Data is collected using scanned images of the lung after they have been enhanced through image pre-processing, and then the images are separated into individual areas of interest (AOIs) using image segmentation. Features are extracted from each AOI, which are a combination of radiomics (texture, shapes, intensity) and/or statistical (i.e., patient-specific) features. Once completed, the features are reduced in dimensionality and selected based on the best characteristics. Finally, the features generated are used in several different techniques, including Support Vector Machines (SVM), Random Forests, and Logistic Regression, for the purposes of classification and prediction. In summary, each step demonstrates how multiple computational algorithms can be used to increase the accuracy of the diagnostic process and to assist health care providers in making more informed clinical decisions [24], [32].

3.2 Taxonomy of Computational Methods

An overall framework for lung cancer diagnosis will help us understand the various types of computation used in that process. If we categorize computation by type of learning or the way it processes the data that we have available to it, we can create a taxonomy that allows us to organize the many different techniques that exist and understand how each technique contributes to some portion of lung cancer diagnosis.

Computational Methods can be thought of as either traditional machine learning methods or next-generation deep learning-based models. Traditional Machine Learning Methods are heavily reliant on manually defining domain-specific features such as texture, shape, intensity, etc. from medical images or clinical data, and then training classifiers, such as SVMs, decision trees and ensemble methods using the derived features. Many studies have shown that traditional methods were able to achieve moderate success in structured datasets, especially for risk prediction and classification tasks, but these methods have limited utility for complex imaging patterns [21], [22].

To address these constraints, deep learning models, specifically convolutional neural networks (CNNs), have emerged as the primary solution for lung cancer diagnosis. The ability of CNNs to automatically extract hierarchical feature representations from raw imaging data removes the requirement for manual feature extraction. The automatic extraction of hierarchical features provides the advantage of enabling CNNs to identify complex spatial patterns and subtle variations in lung/cancer nodules via increased accuracy for diagnosis in areas such as detection, segmentation, and classification [24],[32]. New transformer architectures, in conjunction with attention mechanisms, have also been produced, which complements CNNs' ability to effectively represent hierarchical feature structures and representations, by providing an additional way for models to analyze the global context of a medical image. This is particularly useful for diagnosing patients when there are long-term dependencies between different areas of their medical records. There have also been concepts put forth to create hybrid architectures, which utilize CNNs in conjunction with transformer architectures, thereby harnessing both local feature-based extraction and a global context of a medical image [17],[29].

Furthermore, radiomics-based approaches are another important category of methods, in which multiple quantitative features can be extracted from medical images of a patient's tumor to describe the characteristics of their tumor. Then, these features may be used to train machine-learning models for classification or prediction tasks. Radiomics can provide more interpretability than deep learning methods; however, it is also sensitive to variability in the image acquisition and pre-processing protocols, which limits reproducibility [24]. Recently, multimodal-model learning frameworks have emerged as a valuable approach, which integrate multiple types of information (e.g., imaging, patient medical history, and pathologic data) to create multimodal models that provide a comprehensive understanding of a cancer patient's disease by creating complementary information about the patient's tumor. Research has shown that the use of multimodal approaches can provide substantially improved diagnostic accuracy over traditional univariate methods; however, multimodal approaches present challenges related to data integration and computational cost [30], [32].

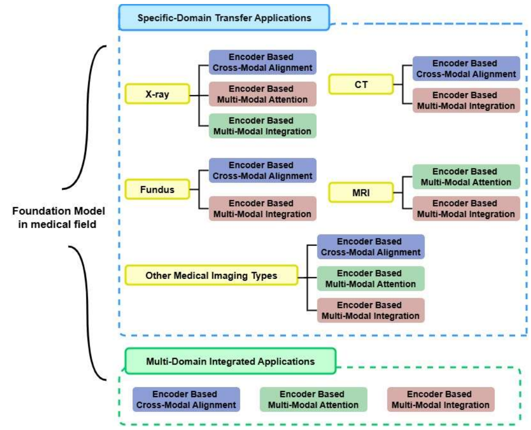

Figure 2: Taxonomy of computational methods used in lung cancer diagnosis [17], [24], [32]

Figure 2 displays a classification scheme (taxonomy) of methods that utilize computer-based computation systems for the diagnosis of lung cancer, and is organized according to the method's learning paradigm type and application domain. It illustrates how foundational models in medical imaging interact using different types of radiography (e.g., X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Fundus (or retinal) imaging) through computer-based computational frameworks. It also categorizes various types of computer-based computational alignment (cross-modal functional alignment) that utilize encoder architectures and use cross-modal multi-modal attention mechanisms, as well as encoder/decoder integration techniques into the frameworks that support multi-domain integrated systems (combining different data sources to improve diagnostic performance). The extended definition assists in providing a comprehensive description of how today's computer-based computational methods are organized in order to provide accurate and scalable means to diagnose lung cancer [17], [24], [32].

3.3 Dataset Description

High-quality and accurately labelled datasets are essential for the successful development and assessment of predictive models used when diagnosing lung cancer. The functionality of machine learning and deep learning models depends heavily on the datasets that were utilized throughout the training and evaluation process, how diversified they were, how large their datasets were, etc. Consequently, it is vital to understand the attributes of commonly used datasets in order to determine the validity and/or applicability of predictive models.

1. The **LIDC-IDRI dataset** includes thoracic CT scans and considerable detail in the annotations made by several radiologists. The LIDC-IDRI lung cancer

data has been used to identify, locate and classify pulmonary nodules as well as develop benchmarks against which deep learning-based identification and classification models can be evaluated. As well, data from multiple annotators can help inform researchers about variability in diagnoses; however, variability between different annotators can cause some inconsistencies. The LIDC-IDRI lung cancer data is available from The Cancer Imaging Archive. (Accessed: April 2026) [10], [24].

Available at: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

2. The **LUNA16 dataset** is a specially curated subset of the LIDC-IDRI dataset and was created for the purpose of benchmark testing lung nodule detection algorithms. The annotations and evaluation protocols for this dataset are standardized in order to allow different computational techniques to be compared, especially deep learning-based detection systems. This dataset is a popular means for evaluating sensitivity and false-positive rates in nodule detection tasks. The LUNA16 dataset originated from the LUNA16 challenge platform. (Accessed: April 2026) [5], [24].

   Available at: https://luna16.grand-challenge.org/

3. The **Kaggle Lung Cancer Dataset** is one of the most popularly used publicly available datasets containing labelled CT scan data for classification and predictive modelling tasks. The Kaggle Lung Cancer Dataset is commonly used in machine learning competitions and experiments because of its well-structured format and ease of access. As such, this dataset facilitates the development and evaluation of classification models and allows for comparisons between different models and algorithms. However, the dataset has a relatively small number of records with respect to size and diversity and, therefore, may limit the generalization of any resulting classification models. This dataset was provided to you through the Kaggle platform. (Accessed: April 2026) [21], [33].

   Available at: https://www.kaggle.com/code/mohamedgobara/lung-cancer-98-8-custom-cnn-model/input

3.4 Technical Analysis of Methods

A comparative technical analysis of major computational predictive modeling approaches used in lung cancer diagnosis is essential for understanding their effectiveness, advantages, and practical limitations in clinical applications. Different methodologies, including traditional machine learning, deep learning, transformer-based architectures, radiomics, and multimodal learning frameworks, vary significantly in terms of feature extraction capability, computational complexity, interpretability, and diagnostic performance. The below table presents the comparative analysis of major computational methods, highlighting their key characteristics, advantages, limitations, and associated references used in lung cancer diagnostic systems.

Table 2: Technical Analysis of Methods

Method Category	**Technique / Model**	**Key Characteristics**	**Advantages**	**Limitations**	**Related References**
Traditional Machine Learning	SVM, Decision Tree, Random Forest, Ensemble Learning	Uses handcrafted features extracted from medical images and structured clinical datasets	Computationally efficient, interpretable, suitable for structured data and risk prediction tasks	Limited capability in handling complex spatial imaging patterns and high-dimensional image features	[21], [22], [26]
Deep Learning	Convolutional Neural Networks (CNNs), 3D	Automatically learns hierarchical feature representations directl	High diagnostic accuracy, effective in detection, segmentation, and	Requires large annotated datasets and high computational	[24], [32]

	CNNs	y from imaging data	classification of lung nodules	resources	
Transformer-Based Models	Vision Transformers (ViTs), Attention-Based Architectures	Captures global contextual relationships and long-range dependencies within medical images	Improved contextual understanding and enhanced diagnostic performance	Computationally intensive and requires extensive training data	[17], [29]
Hybrid Architectures	CNN + Transformer Models	Combines local feature extraction with global contextual learning	Better feature representation and improved classification accuracy	Increased model complexity and reduced interpretability	[17], [29]
Radiomics-Based Methods	Texture, Shape, and Intensity Feature Analysis	Extracts quantitative imaging biomarkers from medical images	Provides explainable and clinically interpretable diagnostic features	Sensitive to image acquisition variability and preprocessing differences	[24]
Multimodal Learning Frameworks	Imaging + Clinical + Pathological Data Integration	Integrates multiple heterogeneous data sources into a unified diagnostic framework	Improved robustness, predictive accuracy, and comprehensive disease characterization	Challenges in data integration, alignment, heterogeneity, and computational complexity	[30], [32]

3.5 Advantages and limitations of the different computational analysis methods

Evaluating the strengths and weaknesses of different computational methods used to diagnose lung cancer from a technical perspective is needed to evaluate the usefulness of these methods for diagnosing lung cancer. Each technique has advantages and disadvantages that may make it less than ideal for actual use.

Traditional machine learning models are advantageous due to their simplicity, lower computational requirements, and higher interpretability. These models are particularly effective when working with structured datasets and well-defined features. However, their dependence on manual feature extraction limits their ability to handle complex imaging data, resulting in comparatively lower performance in detection and classification tasks [21], [22].

On the other hand, deep learning models, especially convolutional neural networks, provide significant improvements in diagnostic accuracy by automatically learning hierarchical features from raw data. They are highly effective in analyzing medical images and detecting subtle patterns that may not be visible through traditional methods. Despite their superior performance, these models require large annotated datasets and high computational resources, and they often lack interpretability, making them difficult to deploy in clinical environments without additional explanation mechanisms [24], [32].

Transformer-based and hybrid models further enhance performance by incorporating attention mechanisms that capture global contextual relationships. These models are particularly useful in complex diagnostic scenarios where both local and global information are critical. However, their implementation is computationally intensive and requires extensive training data, which can limit their practical adoption [17], [29].

Finally, multimodal learning approaches offer the most comprehensive diagnostic capability by integrating multiple data sources such as imaging and clinical information. These models improve robustness and predictive performance by leveraging complementary data. However, challenges related to data integration, heterogeneity, and lack of standardized datasets remain significant obstacles, affecting scalability and reproducibility [30], [32].

In summary, while deep learning and multimodal approaches currently demonstrate the highest performance in lung cancer diagnosis, their practical implementation is constrained by data availability, computational complexity, and interpretability issues. Therefore, future research must focus on developing balanced models that achieve high accuracy while ensuring scalability, transparency, and clinical usability.

4. Conclusion

This review paper presents a comprehensive analysis of recent advancements in computational predictive modeling for lung cancer diagnosis, with a focus on machine learning, deep learning, and multimodal approaches. The study highlights that deep learning models, particularly convolutional neural networks, have significantly improved diagnostic accuracy by enabling automatic feature extraction from medical imaging data. Furthermore, emerging approaches such as transformer-based architectures and multimodal frameworks demonstrate enhanced capability in capturing complex disease patterns by integrating diverse data sources. The analysis also reveals that while these advanced models offer superior performance, challenges related to data availability, model interpretability, computational complexity, and lack of standardization continue to limit their widespread clinical adoption. The role of high-quality datasets and structured system frameworks has been identified as a key factor influencing model reliability and generalizability. Overall, this review contributes to the field by providing a structured understanding of existing computational techniques, datasets, and diagnostic workflows, thereby offering valuable insights for researchers and practitioners. It emphasizes the need for developing balanced and scalable predictive models that not only achieve high accuracy but also ensure transparency and practical applicability in real-world healthcare settings.

5. References

1. Cancer.Net. *Lung Cancer – Non-Small Cell: Risk Factors and Prevention*. Available at: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/risk-factors-and-prevention (Accessed: April 2026).
2. Indian Cancer Society. *Annual Report 2019–2020*. Available at: https://www.indiancancersociety.org/pdf/ics-annual-report-19-20.pdf (Accessed: April 2026).
3. Bhuiyan M.S., et al. "Advancements in early detection of lung cancer in public health: A comprehensive study utilizing machine learning algorithms and predictive models." *Journal of Computer Science and Technology Studies*, 6(1), 2024, pp. 113–121.
4. Alavilli S.K. "Integrating computational drug discovery with machine learning for enhanced lung cancer prediction." *Journal of Current Science*, 11(4), 2023.
5. Tawfeek M.A., et al. "Enhancing the efficiency of lung cancer screening: Predictive models utilizing deep learning from CT scans." *Neural Computing and Applications*, 37(17), 2025, pp. 11459–11477.
6. Ibrahim J.A., Sharma, A. "Utilizing high computational ML model for lung cancer prediction with accurate data analysis." *AIP Conference Proceedings*, Vol. 2855, 2023.
7. Chen, Y., et al. "LDNNET: Towards robust classification of lung nodule and cancer using lung dense neural network." *IEEE Access*, Vol. 9, 2021.
8. Qureshi R., et al. "Computational methods for analysis and prediction of EGFR-mutated lung cancer drug resistance." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1), 2022, pp. 238–255.
9. Pati P., et al. "Current advances in computational biology of lung cancer." 2022.
10. Wang W., et al. "Nodule-Plus R-CNN and deep self-paced active learning for 3D instance segmentation of pulmonary nodules." *IEEE Access*, Vol. 7, 2019.
11. Levi M., et al. "Machine learning computational model to predict lung cancer using electronic medical records." *Cancer Epidemiology*, 92, 2024, 102631.
12. Bouchaffra D., et al. "Advancement in lung cancer diagnosis: A comprehensive review of deep learning approaches." 2024.
13. Shariff V.A.H.U.D.D.I.N., et al. "Advances in lung cancer diagnosis with medical

imaging and deep learning techniques." *Journal of Theoretical and Applied Information Technology*, 101(17), 2023.

14. Jamil D., et al. "A comprehensive survey of computational techniques for lung cancer diagnosis and prediction." 2025.
15. Huang D., et al. "Artificial intelligence in lung cancer: Current applications, future perspectives, and challenges." *Frontiers in Oncology*, 14, 2024.
16. Yang D., et al. "Advances in artificial intelligence applications in lung cancer." *Frontiers in Oncology*, 14, 2024.
17. Arshad M.F., et al. "Artificial intelligence and machine learning in lung cancer." *Cancers*, 17(24), 2025.
18. Cai L., et al. "Lung cancer computational biology and resources." *Cold Spring Harbor Perspectives in Medicine*, 12(2), 2022.
19. Bertolaccini L., et al. "Recent advances in lung cancer research." *Updates in Surgery*, 76(6), 2024.
20. Osamika D., et al. "Predictive analytics for chronic respiratory diseases using big data." 2023.
21. Sweet M.M.R., et al. "Comparative analysis of machine learning techniques for lung cancer prediction." 2024.
22. Gupta S., Gupta, M.K. "Computational model for malignant mesothelioma diagnosis." *The Computer Journal*, 66(1), 2023.
23. Li Y., et al. "Machine learning for lung cancer diagnosis, treatment, and prognosis." *Genomics, Proteomics & Bioinformatics*, 20(5), 2022.
24. Gayap H.T., Akhloufi, M.A. "Deep machine learning for medical diagnosis: Lung cancer detection." *BioMedInformatics*, 4(1), 2024.
25. Alavinejad M., et al. "Smart nanomedicines powered by AI in lung cancer." *Medical Oncology*, 42(5), 2025.
26. Dritsas E., Trigka, M. "Lung cancer risk prediction with machine learning models." *Big Data and Cognitive Computing*, 6(4), 2022.
27. Rao K.B.V., et al. "Early lung cancer prediction by AI-inspired algorithm." *IEEE UPCON*, 2023.
28. Mamun M., et al. "Lung cancer prediction using ensemble learning." *IEEE AIIoT Congress*, 2022.
29. Asif S., et al. "Advancements in machine learning in medical diagnostics." *Archives of Computational Methods in Engineering*, 2025.
30. Tortora M., et al. "RadioPathomics: Multimodal learning in NSCLC." *IEEE Access*, 2023.
31. Sun T., et al. "Risk prediction model for lung infection after chemotherapy." *Frontiers in Oncology*, 2024.
32. Guan P., et al. "Big data analytics on lung cancer diagnosis." *IEEE/ACM Transactions*, 21(4), 2023.
33. Kaggle Dataset. *Lung Cancer CNN Dataset*. Available at: https://www.kaggle.com/code/mohamedgobara/lung-cancer-98-8-custom-cnn-model/input (Accessed: April 2026).