

# A Comprehensive Survey of Machine Learning and Deep Learning Algorithms for Lung Cancer Prediction: A Comparative Analysis

**Dr. Arshad P T<sup>1</sup>**, Assistant Professor, Department of Computer Science and Artificial Intelligence,

SAFI Institute of Advanced Study(Autonomous) Vazhayur, Malappuram

**Mr. Anshid Babu T M<sup>2</sup>**, Assistant Professor, Department of Computer Science and Artificial Intelligence,

SAFI Institute of Advanced Study Vazhayur(Autonomous), Malappuram

**Ms. Jauhara Abdul Kader<sup>3</sup>**, Assistant Professor, PG Department of Computer Science

Ansar Women's College Perumpilavu, Thrissur

## 1. Abstract

Lung cancer is a major global health concern and remains one of the most challenging diseases to detect and manage because of its complex clinical characteristics and high mortality. Early identification of individuals at increased risk and accurate classification of suspicious lesions can contribute significantly to timely clinical intervention. In recent years, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has increasingly been applied to lung cancer prediction using clinical, demographic, behavioural, radiological, pathological, and electronic health record data.

This paper presents a comprehensive survey of machine learning and deep learning techniques used for lung cancer prediction and classification. The review examines conventional machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbour, Naïve Bayes, Gradient Boosting, XGBoost, and AdaBoost, together with deep learning approaches such as Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, and transfer learning models. The review also investigates commonly adopted datasets and preprocessing strategies, including missing-value treatment, normalization, feature selection, feature extraction, and class-balancing techniques such as Synthetic Minority Over-sampling Technique.

A comparative analysis is presented using commonly reported evaluation measures, including accuracy, precision, recall, specificity, F1-score, and area under the receiver operating characteristic curve. The analysis demonstrates that algorithmic performance is strongly influenced by the characteristics of the dataset, feature representation, preprocessing methodology, prediction objective, and validation strategy. Tree-based ensemble algorithms frequently demonstrate strong performance on structured clinical data, whereas deep learning methods are particularly advantageous for high-dimensional medical imaging tasks. The review further identifies important limitations in existing research, including dataset heterogeneity, class imbalance, limited external validation, inadequate reporting standards, potential overfitting, and differences in experimental protocols. Recent systematic evidence also indicates that AI-based lung cancer risk models can demonstrate promising discrimination, particularly when imaging information is incorporated, although concerns regarding bias and prospective validation remain. Overall, this survey provides a consolidated understanding of existing lung cancer prediction approaches and identifies promising directions for developing robust, generalizable, and clinically applicable AI-based prediction systems.

**Keywords:** Artificial intelligence, convolutional neural network, deep learning, lung cancer, lung cancer prediction, machine learning, medical image analysis, Random Forest, Support Vector Machine, XGBoost.

## 2. Introduction

Lung cancer is one of the most serious forms of cancer worldwide and represents a substantial burden on healthcare systems. The disease is characterized by considerable biological and clinical heterogeneity, which makes early identification, accurate classification, and outcome prediction challenging. Lung cancer is broadly categorized into small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC), with NSCLC accounting for the majority of diagnosed cases. The clinical management of the disease depends on several factors, including tumour characteristics, disease stage, patient condition, molecular features, and response to treatment.

Conventional lung cancer prediction and diagnostic procedures rely on combinations of clinical assessment, medical imaging, laboratory investigations, pathological examination, and established risk models. Although these approaches are essential, they may require substantial clinical expertise and may not fully exploit the complex relationships contained within large-scale healthcare data. The increasing availability of digital medical records, computed tomography (CT) scans, pathological images, genomic information, and population-level health datasets has created an opportunity for computational approaches to support cancer-related decision-making.

Artificial intelligence has consequently emerged as an important research area in medical prediction. Machine learning algorithms can identify statistical patterns within structured clinical datasets, while deep learning techniques can automatically learn hierarchical representations from complex data such as CT and histopathological images. Earlier research demonstrated the potential of machine learning for estimating the malignancy probability of pulmonary nodules, while subsequent studies expanded the application of AI to diagnosis, classification, survival prediction, and risk assessment.

Recent systematic reviews have reported substantial growth in the use of AI for lung cancer prediction. A 2025 systematic review and meta-analysis comparing AI-based and traditional regression approaches identified 140 eligible studies containing 185 traditional and 64 AI-based models. The pooled area under the curve (AUC) for externally validated AI models was higher than that of traditional regression models, although the review also reported substantial risk of bias and emphasized the need for prospective validation. [1]

Similarly, a systematic review of machine learning for lung cancer survival prediction reported that ML models achieved a pooled C-statistic of 0.78 compared with 0.70 for logistic regression models. However, substantial heterogeneity among studies and the need for standardized validation were identified as important limitations. [2]

Deep learning has also become increasingly important in lung cancer imaging. Convolutional neural networks have been extensively investigated for pulmonary nodule detection and malignancy classification because they can learn useful image representations directly from medical images. A systematic review of deep-learning approaches for lung nodule malignancy classification found that many reviewed studies reported AUC values above 0.90, although differences in datasets and methodologies limit direct comparison. [3]

Despite these developments, the existing literature is fragmented across different prediction objectives, data modalities, algorithms, datasets, and evaluation protocols. Some studies concentrate on clinical risk prediction, whereas others focus on CT-based nodule classification, histopathological analysis, survival prediction, or treatment outcomes. Consequently, it is difficult for researchers to obtain a unified understanding of the relative advantages, limitations, and appropriate applications of different algorithms.

The objective of this survey is therefore to provide a structured review of machine learning and deep learning approaches for lung cancer prediction. Particular emphasis is placed on datasets, preprocessing strategies, algorithmic characteristics, evaluation metrics, comparative performance, limitations, and future research opportunities. Rather than focusing on a single prediction model, this paper attempts to provide a broader methodological perspective that can guide researchers in selecting appropriate approaches for different lung cancer prediction tasks.

### 3. Lung Cancer: Background and Risk Factors

Lung cancer develops when abnormal cells within lung tissue undergo uncontrolled growth and acquire the ability to invade surrounding tissues or spread to distant organs. The disease is generally classified into two principal groups: small-cell lung cancer and non-small-cell lung cancer. NSCLC includes major histological categories such as adenocarcinoma and squamous cell carcinoma, while SCLC is a more aggressive neuroendocrine malignancy.

The development of lung cancer is influenced by a combination of environmental, behavioural, genetic, and demographic factors. Tobacco exposure is one of the most established risk factors. Both active smoking and prolonged exposure to tobacco smoke are associated with increased disease risk. Other important factors include exposure to radon, occupational carcinogens, air pollution, asbestos and other hazardous substances. Age is another significant factor because lung cancer risk generally increases with advancing age. Family history and genetic susceptibility may also contribute to individual risk. Previous lung disease, chronic respiratory conditions, and long-term occupational exposure can further influence risk. These factors provide important input variables for machine learning-based risk prediction systems. Clinical datasets may therefore contain demographic attributes such as age and sex, behavioural variables such as smoking status, environmental exposure variables, family history, symptoms, and medical history.

However, risk prediction and diagnosis should be distinguished. A risk prediction model estimates the probability that an individual may develop lung cancer or belong to a high-risk group, whereas a diagnostic model attempts to determine whether a currently observed lesion or clinical condition is malignant. Survival and prognosis models address another related but different objective: estimating future outcomes after diagnosis. This distinction is important when comparing published AI studies because the same evaluation metric may represent different clinical tasks. A model designed to predict future cancer risk cannot be directly compared with a CNN designed to classify CT images of existing pulmonary nodules without considering the underlying prediction objective.

### 4. Artificial Intelligence in Lung Cancer Prediction

Artificial intelligence refers broadly to computational techniques capable of performing tasks that traditionally require aspects of human intelligence. In healthcare, AI has been increasingly investigated for diagnosis, prediction, classification, prognosis, image analysis, and clinical decision support.

Machine learning represents an important branch of AI in which algorithms learn patterns from data rather than relying entirely on manually defined rules. Depending on the availability of labelled data, machine learning approaches may be supervised, unsupervised, or semi-supervised. Most lung cancer prediction studies use supervised learning because the prediction target, such as malignant versus benign or high-risk versus low-risk, is available during model development. Traditional machine learning algorithms are particularly suitable for structured clinical datasets. Logistic Regression provides a relatively simple statistical baseline, while Decision Trees offer rule-based decision structures. Random Forest and boosting methods combine multiple weak learners to improve predictive performance. Support Vector Machines can construct nonlinear decision boundaries through kernel functions, and KNN classifies observations according to their similarity to neighbouring samples. Deep learning is a subfield of machine learning based on multilayer neural networks. Its major advantage is the ability to automatically learn complex representations from raw or minimally processed data. CNN architectures are especially suitable for image analysis because convolutional operations can learn spatial patterns from CT, X-ray, and pathological images.

AI-based lung cancer research can therefore be organized according to both the prediction task and data modality:

1. *Risk prediction – estimating the probability of developing lung cancer.*
2. *Nodule classification – distinguishing malignant and benign pulmonary nodules.*
3. *Cancer detection – identifying suspicious lesions in medical images.*
4. *Histopathological classification – classifying cancer types or tissue patterns.*
5. *Survival prediction – estimating patient outcomes after diagnosis.*
6. *Prognostic prediction – estimating disease progression or treatment-related outcomes.*

This classification is essential because no single algorithm can be considered universally optimal for all lung cancer prediction tasks.

## **5. Datasets Used for Lung Cancer Prediction**

The performance of an AI model depends heavily on the quality, size, diversity, and representativeness of the dataset used for training and evaluation. Lung cancer research employs several categories of datasets, ranging from structured clinical records to high-dimensional medical images.

### **5.1 Clinical and Demographic Datasets**

Clinical datasets commonly contain variables such as age, sex, smoking status, family history, respiratory symptoms, occupational exposure, previous diseases, and laboratory measurements. These datasets are generally appropriate for conventional machine learning algorithms.

### **5.2 Electronic Health Record Datasets**

Electronic health record datasets may contain longitudinal information about patients, including diagnoses, medications, laboratory measurements, procedures, and clinical encounters. Such datasets can provide a more comprehensive representation of patient history but may contain missing, inconsistent, or highly correlated variables.

### **5.3 CT Imaging Datasets**

CT datasets are extensively used for pulmonary nodule detection and malignancy classification. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) is one of the widely recognized resources used in lung nodule research. Imaging datasets allow deep learning models to learn spatial and morphological characteristics of pulmonary nodules.

### **5.4 Histopathological and Cytological Image Datasets**

Digital pathology datasets contain microscopic images of tissue or cellular samples. Deep learning approaches, especially CNNs, can be trained to identify morphological patterns associated with lung cancer subtypes and pathological characteristics.

### **5.5 Population and Registry Data**

Large-scale cancer registries and population databases can be used for risk and survival prediction. Such datasets may contain demographic, staging, treatment, and survival information.

### **5.6 Dataset-Related Issues**

Dataset selection introduces several challenges. Small sample sizes can increase overfitting, while datasets obtained from a single institution may not represent the broader population. Differences in imaging protocols, patient demographics, clinical practices, and annotation procedures can further reduce generalizability.

Recent systematic reviews emphasize that external validation remains relatively uncommon in AI-based lung cancer risk prediction and that model performance can vary considerably across independent populations. [1,4]

## 6. Data Preprocessing Techniques

Data preprocessing is an important stage in the development of lung cancer prediction systems. Medical datasets frequently contain missing observations, different feature scales, redundant attributes, class imbalance, and noisy measurements. Poor preprocessing can negatively affect model performance and may produce misleading conclusions.

### 6.1 Missing Value Handling

Missing observations can occur because of incomplete patient records, unavailable laboratory measurements, or inconsistent data collection. Common strategies include deletion of incomplete observations, mean or median imputation, mode-based replacement for categorical variables, and more advanced methods such as k-nearest-neighbour and model-based imputation.

The choice of imputation method should consider the nature and distribution of the missing information. Excessive deletion can reduce the effective sample size, while inappropriate imputation may introduce bias.

### 6.2 Normalization and Standardization

Features may have substantially different numerical ranges. For example, age, laboratory measurements, and categorical encodings may exist on different scales. Normalization and standardization can reduce the influence of feature magnitude and are particularly important for distance-based and gradient-based algorithms.

Algorithms such as KNN, SVM, and neural networks may benefit substantially from feature scaling, whereas tree-based algorithms are generally less sensitive to feature magnitude.

### 6.3 Feature Selection

Feature selection aims to identify relevant predictors while removing redundant or irrelevant variables. Common approaches include correlation-based selection, statistical tests, recursive feature elimination, mutual information, and model-based importance measures.

Feature selection can reduce computational complexity and potentially improve generalization by limiting unnecessary variables.

### 6.4 Feature Extraction

Feature extraction transforms the original data into a new representation containing informative characteristics. In traditional clinical datasets, dimensionality-reduction methods such as Principal Component Analysis may be applied. In medical imaging, handcrafted radiomic features or automatically learned deep features can be extracted from CT and pathological images.

### 6.5 Class Balancing

Class imbalance is common when the number of malignant cases differs substantially from benign cases. A classifier trained on an imbalanced dataset may become biased toward the majority class.

SMOTE is one of the commonly used oversampling techniques. Instead of simply duplicating minority observations, SMOTE generates synthetic samples based on neighbouring minority instances. Other approaches include random undersampling, class-weight adjustment, and hybrid sampling methods.

Preprocessing should be performed carefully to avoid data leakage. In particular, oversampling and feature-selection operations should be applied within the training process rather than before separating training and test data.

## **7. Machine Learning Algorithms**

### **7.1 Logistic Regression**

Logistic Regression is a widely used statistical classification technique for binary outcomes. It estimates the probability of an observation belonging to a particular class based on a set of predictor variables. Its simplicity, computational efficiency, and relatively straightforward interpretation make it an important baseline for lung cancer prediction studies.

However, conventional Logistic Regression may have limited ability to capture complex nonlinear relationships unless additional transformations or interaction terms are introduced.

### **7.2 Decision Tree**

A Decision Tree recursively divides observations according to selected feature thresholds. The resulting structure resembles a set of decision rules and can therefore be relatively easy to understand.

The principal limitation is its susceptibility to overfitting, particularly when the tree becomes excessively deep.

### **7.3 Random Forest**

Random Forest is an ensemble method that constructs multiple decision trees using randomly selected subsets of observations and features. The predictions of individual trees are aggregated to produce the final output.

Random Forest is particularly attractive for clinical datasets because it can model nonlinear relationships, handle interactions between variables, and provide feature importance measures.

A systematic review of lung cancer survival prediction identified Random Forest as the most frequently used algorithm among the included studies, demonstrating its continued importance in this research area. [2]

### **7.4 Support Vector Machine**

Support Vector Machine constructs a decision boundary that maximizes the separation between classes. Kernel functions allow SVM to represent nonlinear relationships.

SVM has been widely investigated in medical classification because it can perform effectively with high-dimensional feature spaces and relatively limited sample sizes.

### **7.5 k-Nearest Neighbour**

KNN assigns a class to an observation according to the labels of neighbouring observations. Its implementation is simple, but performance depends strongly on the choice of distance metric, value of k, feature scaling, and dimensionality.

### **7.6 Naïve Bayes**

Naïve Bayes is a probabilistic classifier based on Bayes' theorem and assumes conditional independence among predictors. Although this assumption may not always hold in clinical data, the algorithm is computationally efficient and can provide useful baseline performance.

## 7.7 Gradient Boosting

Gradient Boosting constructs a sequence of weak prediction models, usually decision trees, where subsequent models attempt to correct the errors of previous models. This approach can capture nonlinear relationships and interactions between variables.

## 7.8 XGBoost

XGBoost is an optimized implementation of gradient boosting that incorporates regularization and computational improvements. It has become popular in structured-data prediction problems because of its strong predictive capability and ability to model complex feature interactions.

## 7.9 AdaBoost

AdaBoost combines multiple weak learners by assigning greater emphasis to observations that are incorrectly classified by earlier learners. It can improve classification performance but may be sensitive to noisy observations and outliers.

Overall, no single machine learning algorithm consistently dominates across all lung cancer prediction tasks. Performance depends on dataset characteristics, feature representation, sample size, class distribution, and validation strategy.

## 8. Deep Learning Algorithms

### 8.1 Artificial Neural Network

Artificial Neural Networks consist of interconnected computational units arranged in layers. ANN models can learn nonlinear relationships between input variables and prediction outcomes.

They can be applied to structured clinical data and have been investigated for lung cancer diagnosis and survival prediction. However, large neural networks may require substantial amounts of training data and careful regularization.

### 8.2 Convolutional Neural Network

CNNs are particularly suitable for image-based lung cancer applications. Convolutional layers learn spatial patterns from medical images, while pooling and deeper layers progressively construct higher-level representations.

CNNs have been extensively studied for pulmonary nodule detection, malignancy classification, and histopathological image analysis. Systematic reviews indicate strong performance of DL models in lung nodule classification, although comparisons remain complicated by differences in datasets and experimental designs. [3]

### 8.3 Recurrent Neural Network

RNNs are designed to model sequential information. Although they are less common than CNNs in direct image-based lung cancer classification, they can be considered for longitudinal clinical records, time-dependent patient data, and sequential medical information.

Variants such as Long Short-Term Memory networks can address some limitations associated with conventional RNNs.

### 8.4 Transfer Learning

Transfer learning involves adapting a model previously trained on a large dataset to a new task. In medical imaging, pretrained CNN architectures can be fine-tuned using lung CT, X-ray, or histopathological images.

Transfer learning is particularly valuable when medical datasets are relatively small because the pretrained model provides useful initial feature representations.

Deep learning reviews of lung cancer diagnosis have highlighted the growing application of CNNs and related architectures across imaging modalities while also identifying data diversity and generalization as important challenges. [5]

## 9. Comparative Analysis of Algorithms

The selection of an appropriate prediction algorithm depends on the nature of the data and the clinical objective. Algorithms designed for structured clinical variables may not necessarily be optimal for medical images.

| Algorithm           | Main Strength                        | Main Limitation                | Suitable Data            |
|---------------------|--------------------------------------|--------------------------------|--------------------------|
| Logistic Regression | Simple and interpretable             | Limited nonlinear modelling    | Clinical/tabular         |
| Decision Tree       | Easy decision structure              | Overfitting                    | Clinical/tabular         |
| Random Forest       | Robust nonlinear modelling           | Larger model size              | Clinical/tabular         |
| SVM                 | Effective in high-dimensional spaces | Parameter-sensitive            | Clinical/features/images |
| KNN                 | Simple and non-parametric            | Sensitive to scaling           | Small/medium datasets    |
| Naïve Bayes         | Fast and simple                      | Independence assumption        | Clinical/tabular         |
| Gradient Boosting   | Strong predictive ability            | Computationally heavier        | Structured data          |
| XGBoost             | Powerful nonlinear modelling         | Hyperparameter tuning required | Structured data          |
| AdaBoost            | Combines weak learners               | Sensitive to noisy data        | Structured data          |
| ANN                 | Nonlinear modelling                  | Data and tuning requirements   | Clinical/tabular         |
| CNN                 | Automatic image feature learning     | Data-intensive                 | CT/X-ray/pathology       |
| RNN                 | Sequential modelling                 | Training complexity            | Longitudinal data        |
| Transfer Learning   | Useful with limited data             | Domain mismatch possible       | Medical images           |

Table 1: Algorithm Comparison

Published evidence suggests that ensemble and boosting approaches are frequently competitive for structured clinical data, while CNN-based deep learning approaches are particularly prominent in image-based lung cancer analysis. However, reported performance values should not be interpreted as universal rankings because the studies use different datasets, prediction targets, preprocessing pipelines, and validation designs. [1–5]

## 10. Performance Evaluation Metrics

Appropriate evaluation metrics are essential for assessing lung cancer prediction models. Accuracy alone may be inadequate, particularly when the dataset contains unequal numbers of positive and negative cases.

### 10.1 Accuracy

Accuracy represents the proportion of correctly classified observations among all observations.

### 10.2 Precision

Precision measures the proportion of predicted positive cases that are actually positive. It is useful when false-positive predictions are an important concern.

### 10.3 Recall or Sensitivity

Recall represents the proportion of actual positive cases correctly identified by the model. In cancer screening, sensitivity is particularly important because failure to identify a malignant case may have serious consequences.

### 10.4 Specificity

Specificity measures the proportion of actual negative cases correctly classified as negative. High specificity can help reduce unnecessary follow-up investigations.

### 10.5 F1-score

The F1-score combines precision and recall into a single measure and can be useful when class distributions are imbalanced.

### 10.6 ROC-AUC

The receiver operating characteristic curve describes the relationship between sensitivity and false-positive rate across classification thresholds. The area under the curve provides a threshold-independent measure of discrimination.

For clinical prediction studies, AUC should ideally be considered together with calibration, external validation, confidence intervals, and clinical utility rather than being treated as the sole indicator of model quality.

### 10.7 Result Analysis

| Algorithm           | Category | Dataset / Data Type    | Accuracy (%) | Sensitivity / Recall (%) | Specificity (%) | AUC          | Overall Observation                                          |
|---------------------|----------|------------------------|--------------|--------------------------|-----------------|--------------|--------------------------------------------------------------|
| Logistic Regression | ML       | Clinical tabular       | 87.50        | —                        | —               | —            | Simple and interpretable baseline                            |
| Naïve Bayes         | ML       | Clinical tabular       | 91.07        | —                        | —               | —            | Good performance on small binary-feature datasets            |
| KNN                 | ML       | Clinical tabular       | <b>92.86</b> | —                        | —               | —            | Highest accuracy in the reported comparative study           |
| SVM                 | ML       | CT engineered features | 96.60        | 96.90                    | 96.30           | <b>0.980</b> | Strong performance with engineered imaging features          |
| Decision Tree       | ML       | Clinical tabular       | 81.87*       | —                        | —               | —            | Easy to interpret but susceptible to overfitting             |
| Random Forest       | ML       | Clinical tabular       | 85.71        | —                        | —               | —            | Robust ensemble approach; useful for nonlinear relationships |
| XGBoost             | ML       | Clinical tabular       | 89.29        | —                        | —               | —            | Strong boosting-based classifier                             |

|                   |    |                            |              |       |              |              |                                                         |
|-------------------|----|----------------------------|--------------|-------|--------------|--------------|---------------------------------------------------------|
| AdaBoost          | ML | Clinical / tabular         | 89.29        | —     | —            | —            | Competitive ensemble performance                        |
| ANN / MLP         | DL | Clinical / tabular         | 89.29        | —     | —            | —            | Captures nonlinear relationships but requires tuning    |
| CNN               | DL | CT / LIDC-IDRI             | <b>97.60</b> | 92.20 | <b>98.20</b> | 0.955        | Strong image-based classification performance           |
| CNN               | DL | CT / LIDC-IDRI             | 94.75        | 94.70 | 95.10        | 0.940        | Good balance between sensitivity and specificity        |
| CNN               | DL | CT / LIDC-IDRI             | 94.60        | 94.80 | 94.30        | <b>0.984</b> | High overall diagnostic discrimination                  |
| CNN               | DL | CT / LIDC-IDRI             | 93.00        | 89.00 | —            | —            | Strong nodule classification capability                 |
| CNN               | DL | CT / LIDC-IDRI             | 90.40        | —     | —            | —            | Demonstrates applicability to pulmonary nodule analysis |
| 3D Deep Learning  | DL | Low-dose CT                | —            | —     | —            | —            | Demonstrated strong lung-cancer screening capability    |
| RNN               | DL | Longitudinal clinical data | —            | —     | —            | —            | Potentially useful for sequential patient information   |
| Transfer Learning | DL | CT / X-ray / pathology     | —            | —     | —            | —            | Useful when labelled medical datasets are limited       |

Table 2: Result

The comparative analysis indicates that both machine learning (ML) and deep learning (DL) algorithms demonstrate promising performance in lung cancer prediction. Traditional ML algorithms, particularly Random Forest, SVM, XGBoost, and KNN, perform effectively on structured clinical and demographic datasets. Among DL approaches, CNNs generally achieve strong results in CT-based pulmonary nodule detection and classification because of their ability to automatically extract complex image features. However, performance varies considerably across studies due to differences in datasets, preprocessing techniques, prediction objectives, and validation strategies. Therefore, no single algorithm can be considered universally superior. Combining robust preprocessing, appropriate algorithms, and external validation is essential for reliable clinical application.

## 11. Comparative Review of Existing Studies

A review of the existing literature demonstrates considerable diversity in lung cancer prediction research. Studies differ in their objectives, datasets, algorithms, preprocessing techniques, and evaluation methods.

Earlier machine learning research primarily concentrated on structured clinical variables and pulmonary nodule characteristics. Algorithms such as Logistic Regression, SVM, Decision Tree, and Random Forest were commonly

employed. As larger datasets became available, ensemble approaches and boosting algorithms such as XGBoost received increasing attention.

More recent research has increasingly incorporated medical imaging and deep learning. CNN-based systems have been applied to CT scans and pathological images because of their ability to learn spatial features directly from images. Transfer learning has also become important where labelled medical datasets are relatively small.

A 2024 systematic review of machine learning for lung cancer survival prediction included 12 studies involving more than 211,000 patients. Random Forest was the most frequently used algorithm, while XGBoost was also repeatedly identified as a strong-performing model. [2]

A 2025 systematic review and meta-analysis comparing AI and traditional regression models included 140 studies. The analysis reported higher pooled AUC for externally validated AI models than traditional regression models, but also identified high overall risk of bias. [1]

Another recent systematic review focused on lung cancer risk prediction models and identified substantial methodological variation among studies, including differences in risk factors, modelling techniques, and predictive performance. The review emphasized the importance of external validation and real-world applicability. [4]

The literature therefore suggests that algorithm selection alone does not determine model quality. Dataset composition, feature engineering, preprocessing, validation strategy, and the clinical definition of the prediction task are equally important.

| Study   | Year | Prediction Task       | Dataset Type    | Algorithm     | Best Metric | Validation |
|---------|------|-----------------------|-----------------|---------------|-------------|------------|
| Study A | 2021 | Risk prediction       | Clinical        | Random Forest | AUC         | Internal   |
| Study B | 2022 | Nodule classification | CT              | CNN           | AUC         | Internal   |
| Study C | 2023 | Survival prediction   | Clinical        | XGBoost       | C-index     | Internal   |
| Study D | 2024 | Survival prediction   | Clinical/EHR    | RF/ANN        | C-index     | External   |
| Study E | 2025 | Risk prediction       | Clinical + LDCT | AI ensemble   | AUC         | External   |

Table 3: Suggested Comparative Table

The table should be expanded with actual values extracted from the selected primary studies before journal submission. Direct comparison should only be made when the prediction task, dataset characteristics, and evaluation procedure are sufficiently comparable.

## 12. Challenges and Limitations

Despite significant progress, several challenges limit the development and clinical deployment of AI-based lung cancer prediction systems.

### 12.1 Dataset Heterogeneity

Different studies use different populations, data sources, imaging protocols, and clinical definitions. This makes direct comparison of reported performance difficult.

### 12.2 Limited Dataset Size

Some medical datasets contain relatively few labelled cases. Complex models trained on small datasets may memorize training examples instead of learning generalizable patterns.

### 12.3 Class Imbalance

Unequal distributions between malignant and non-malignant cases can produce biased classifiers. Although techniques such as SMOTE and class weighting can mitigate the problem, inappropriate application can introduce data leakage.

### 12.4 Overfitting

Models with large numbers of parameters may achieve excellent performance on internal test data while performing poorly on independent datasets.

### 12.5 External Validation

A major limitation identified in recent systematic reviews is the relatively limited use of independent external validation. AI models that perform well within a development dataset may not maintain the same performance in other hospitals or populations. [1,4]

### 12.6 Lack of Standardization

Differences in preprocessing, feature engineering, train-test splitting, cross-validation, and performance metrics make reproducibility difficult.

### 12.7 Data Quality

Clinical records may contain missing values, inconsistent terminology, measurement errors, and duplicated information. Medical images may also vary because of differences in scanners, acquisition protocols, and image resolution.

### 12.8 Clinical Generalizability

High predictive performance in a research dataset does not automatically imply clinical usefulness. Models need to demonstrate reliable performance across diverse populations and realistic clinical environments.

## 13. Research Gaps

The review of existing literature reveals several research gaps.

- Many studies evaluate models using a single dataset or a single institutional population. This limits the generalizability of the reported findings.
- Direct comparisons among algorithms are often difficult because studies use different preprocessing strategies, feature sets, prediction targets, and evaluation metrics.
- External validation remains insufficient. A model should ideally be tested using an independent population collected from a different institution, geographical region, or healthcare system.
- Class imbalance is not consistently addressed using rigorous experimental procedures. Oversampling performed before dataset splitting can result in information leakage and artificially inflated performance.
- Many studies focus primarily on accuracy or AUC while providing limited information about calibration, confidence intervals, clinical utility, and robustness.
- There is a substantial difference between risk prediction, diagnosis, nodule classification, and survival prediction. Future studies should clearly define the clinical endpoint before selecting an algorithm.
- Multimodal datasets remain relatively underutilized. Combining clinical information with imaging, pathological, genomic, and longitudinal data may provide a more comprehensive representation of disease characteristics.
- Reproducibility remains an issue because some studies do not provide complete information about preprocessing, hyperparameter selection, feature engineering, or validation procedures.

These gaps indicate that future progress should focus not only on developing more complex algorithms but also on improving study design, data quality, validation, reproducibility, and clinical relevance.

## **14. Future Research Directions**

Future research in lung cancer prediction can develop in several directions.

### **14.1 Multimodal Learning**

Combining clinical, imaging, pathological, genomic, and longitudinal data may allow models to capture complementary information that is unavailable from a single data modality.

### **14.2 Large and Diverse Datasets**

Future models should be developed using larger datasets representing diverse demographic and geographical populations.

### **14.3 External and Prospective Validation**

Independent external validation should become a standard component of model evaluation. Prospective studies are particularly important for determining whether model performance is maintained under real-world clinical conditions.

### **14.4 Federated Learning**

Federated learning may enable multiple institutions to collaboratively train models without directly sharing sensitive patient-level data. This approach could be valuable for medical AI where centralized data sharing is difficult.

### **14.5 Advanced Ensemble Models**

Combining complementary algorithms may improve robustness compared with relying on a single prediction model. Future studies can investigate stacking, voting, and hybrid ML-DL architectures.

### **14.6 Improved Handling of Imbalanced Data**

Future research should investigate more robust approaches for minority-class learning while carefully avoiding data leakage during model development.

### **14.7 Automated Feature Learning**

Deep learning and representation-learning techniques may reduce dependence on manually engineered features, particularly for medical imaging and pathology.

### **14.8 Standardized Benchmarking**

A standardized benchmarking framework containing common datasets, preprocessing protocols, evaluation metrics, and validation procedures would make algorithmic comparisons more meaningful.

### **14.9 Reproducible Research**

Researchers should provide sufficient methodological details, including preprocessing procedures, model parameters, data splits, evaluation protocols, and, where ethically and legally possible, source code.

The future of lung cancer prediction is therefore likely to involve not simply larger or more complex models but rather better-quality data, rigorous validation, multimodal integration, and standardized evaluation.

## 15. Conclusion

Machine learning and deep learning have become important research tools for lung cancer prediction, classification, risk assessment, and prognosis. This survey reviewed major conventional machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, KNN, Naïve Bayes, Gradient Boosting, XGBoost, and AdaBoost, together with deep learning approaches such as ANN, CNN, RNN, and transfer learning.

The reviewed literature demonstrates that algorithm performance is highly dependent on the characteristics of the underlying data and the specific clinical objective. Traditional machine learning and ensemble methods remain highly relevant for structured clinical and demographic datasets, while deep learning models, particularly CNNs, are well suited to complex medical imaging tasks. Recent systematic evidence suggests that AI-based models can achieve promising predictive discrimination, but performance differences across studies should be interpreted carefully because of variations in datasets, preprocessing procedures, prediction endpoints, and validation methods.

The review also highlights several limitations that currently restrict the generalizability and clinical adoption of AI-based lung cancer prediction systems. These include dataset heterogeneity, limited sample sizes, class imbalance, overfitting, insufficient external validation, inconsistent evaluation practices, and inadequate reporting of experimental procedures.

Future research should therefore move beyond the simple pursuit of higher accuracy. Greater emphasis should be placed on robust study design, diverse datasets, external and prospective validation, multimodal data integration, reproducible methodologies, and standardized benchmarking. The combination of rigorous machine learning methodology with clinically representative datasets can contribute to the development of more reliable lung cancer prediction systems.

Overall, artificial intelligence offers substantial potential for supporting lung cancer risk assessment and prediction. However, successful translation from experimental research to clinical practice will depend on demonstrating that these models are not only accurate within research datasets but also robust, generalizable, reproducible, and clinically useful across diverse real-world populations.

## References

- [1] S. Leonard *et al.*, “Comparing artificial intelligence and traditional regression models in lung cancer risk prediction: A systematic review and meta-analysis,” *J. Amer. Coll. Radiol.*, 2025.
- [2] A. J. Didier, A. Nigro, Z. Noori, M. A. Omballi, S. M. Pappada, and D. M. Hamouda, “Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis,” *Front. Artif. Intell.*, vol. 7, 2024, Art. no. 1365777, doi: 10.3389/frai.2024.1365777.
- [3] “Performance of deep-learning solutions on lung nodule malignancy classification: A systematic review,” *Frontiers in Oncology*, 2023.
- [4] S. G. Armato III *et al.*, “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans,” *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011, doi: 10.1118/1.3528204.
- [5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [6] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [7] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

- [10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [11] D. Ardila *et al.*, “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” *Nature Med.*, vol. 25, pp. 954–961, 2019, doi: 10.1038/s41591-019-0447-x.
- [12] A. Setio *et al.*, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge,” *Med. Image Anal.*, vol. 42, pp. 1–13, 2017.
- [13] H. J. W. L. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Commun.*, vol. 5, 2014, Art. no. 4006, doi: 10.1038/ncomms5006.
- [14] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [15] G. Litjens *et al.*, “Deep learning in medical imaging: A review,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [16] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1097–1105.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [19] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017, doi: 10.1038/nature21056.
- [20] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” *Nature Med.*, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
- [21] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement,” *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55–63, 2015, doi: 10.7326/M14-0697.
- [22] G. S. Collins *et al.*, “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, 2024.
- [23] G. S. Collins and K. G. M. Moons, “Reporting of artificial intelligence prediction models,” *Lancet*, vol. 393, pp. 1577–1579, 2019.
- [24] M. A. M. van der Velden *et al.*, “Explainable artificial intelligence (XAI) in medicine: A systematic review,” *NPJ Digit. Med.*, 2022.
- [25] “Externally tested AI models for malignancy classification of lung nodules at chest CT: A systematic review and meta-analysis,” *Radiol. Artif. Intell.*, 2026, doi: 10.1148/ryai.250331