Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Auxiliary Diagnosis of Pulmonary Nodules' Benignancy and Malignancy Based on Machine Learning: A Retrospective Study

Wanling Wang[1], Bingqing Yang[2], Huan Wu[1], Hebin Che[1], Yue Tong[1], Bozun Zhang[1], Hongwu Liu[3,*], Yuanyuan Chen[1,*]

[1]Medical Innovation Research Department of PLA General Hospital, Beijing, People's Republic of China; [2]Goodwill Hessian Health Technology Co. Ltd, Beijing, People's Republic of China; [3]Department of Pulmonary and Critical Care Medicine, the Seventh Medical Center of Chinese PLA General Hospital, Beijing, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yuanyuan Chen; Hongwu Liu, Email charry135@163.com; liuhw1005@163.com

Background: Lung cancer, one of the most lethal malignancies globally, often presents insidiously as pulmonary nodules. Its nonspecific clinical presentation and heterogeneous imaging characteristics hinder accurate differentiation between benign and malignant lesions, while biopsy's invasiveness and procedural constraints underscore the critical need for non-invasive early diagnostic approaches.

Methods: In this retrospective study, we analyzed outpatient and inpatient records from the First Medical Center of Chinese PLA General Hospital between 2011 and 2021, focusing on pulmonary nodules measuring 5–30mm on CT scans without overt signs of malignancy. Pathological examination served as the reference standard. Comparative experiments evaluated SVM, RF, XGBoost, FNN, and Atten_FNN using five-fold cross-validation to assess AUC, sensitivity, and specificity. The dataset was split 70%/30%, and stratified five-fold cross-validation was applied to the training set. The optimal model was interpreted with SHAP to identify the most influential predictive features.

Results: This study enrolled 3355 patients, including 1156 with benign and 2199 with malignant pulmonary nodules. The Atten_FNN model demonstrated superior performance in five-fold cross-validation, achieving an AUC of 0.82, accuracy of 0.75, sensitivity of 0.77, and F1 score of 0.80. SHAP analysis revealed key predictive factors: demographic variables (age, sex, BMI), CT-derived features (maximum nodule diameter, morphology, density, calcification, ground-glass opacity), and laboratory biomarkers (neuroendocrine markers, carcinoembryonic antigen).

Conclusion: This study integrates electronic medical records and pathology data to predict pulmonary nodule malignancy using machine/deep learning models. SHAP-based interpretability analysis uncovered key clinical determinants. Acknowledging limitations in cross-center generalizability, we propose the development of a multimodal diagnostic systems that combines CT imaging and radiomics, to be validated in multi-center prospective cohorts to facilitate clinical translation. This framework establishes a novel paradigm for early precision diagnosis of lung cancer.

Keywords: pulmonary nodules, benignancy, malignancy, machine learning, risk factors

Introduction

Since 2013, PM2.5 has been classified by the International Agency for Research on Cancer (IARC) as a carcinogen that elevates the risk of lung cancer. While pollution levels have decreased in some regions, they continue to rise in rapidly industrializing Asian countries, presenting a significant global public health challenge.[1,2] The lungs, being directly exposed to ambient air pollutants, bear the brunt of these adverse health impacts.[3] Lung cancer is the deadliest cancer worldwide, often presenting as lung nodules in the early stages.[4,5] Lung cancer poses a serious threat to human health and will continue to significantly affect human life in the future. Although the mortality rate of lung cancer is high, early

Received: 17 January 2025
Accepted: 14 June 2025
Published: 27 June 2025

diagnosis can significantly extend patient survival.[6] Traditional biopsy is considered the gold standard for determining lung tumor malignancy.[7] However, this technique requires invasive surgery to extract tissue for analysis and is time-consuming. Due to the temporal and spatial heterogeneity of tumors, this method is limited, making it difficult for patients to undergo multiple biopsies. Furthermore, pulmonary nodules, which are defined as round or irregular lesions less than or equal to 30 mm in diameter, can be either benign or malignant.[8] If malignant pulmonary nodules go untreated, they may progress to lung cancer, subsequently having a considerable negative impact on patient survival.[9] Therefore, developing more effective and less destructive methods to diagnose malignant lung nodules remains an urgent clinical challenge.[10]

Pulmonary nodules (PN) refer to focal, round, high-density solid or subsolid shadows in the lungs with a diameter of ≤ 30 mm. They are characterized by increased density on imaging, can be single or multiple, and do not accompany lung collapse, pleural effusion, or enlarged mediastinal lymph nodes. PN can be further classified into benign pulmonary nodules (BPN) and malignant pulmonary nodules (MPN) based on their characteristics. BPN refers to nodules caused by benign lesions such as inflammation, vascular malformations, or hyperplasia in the lungs, while MPN refers to nodules that present as primary lung tumors or metastatic tumors. In recent years, with advancements in imaging technology and increased awareness of physical examinations, the detection rate of PN has significantly increased. The majority of patients exhibits no related clinical symptoms and are only discovered through physical examinations whereas regular follow-ups should be conducted for benign nodules. Evaluation strategies involve estimating the probability of malignancy and utilizing advanced imaging techniques.[11] the diagnosis of PN still largely relies on the clinical physician's expertise, with emerging radiomic frameworks providing assistance in differentiating nodule types.[12]

With the continuous advancement of deep learning, many scholars have employed deep learning networks to study the benign and malignant nature of pulmonary nodules. Nibali et al[13] proposed a classification model based on residual networks, which achieves the benign and malignant diagnosis of pulmonary nodules through lung CT images. Hyunsuk Yoo et al[14] utilized chest X-ray images as data and constructed a classification model based on convolutional neural networks to evaluate its performance in detecting pulmonary nodules and lung cancer. KV Venkadesh et al[15] employed low-dose screening CT as data and based on convolutional neural networks to assess the risk of malignant tumors. However, the aforementioned deep learning methods are based on a single image data modality and do not include text data such as patients' medical records and laboratory test results. Furthermore, clinical CT images are difficult to obtain and often require professional annotation and image segmentation, making their application in practical clinical settings challenging. To address these limitations, this study aims to develop a multimodal artificial intelligence (AI) framework that integrates clinical text data (eg, medical histories, laboratory findings) with imaging features to improve the accuracy and accessibility of PN classification. Unlike existing approaches, our method reduces reliance on resource-intensive image preprocessing and leverages heterogeneous real-world clinical data to better mimic diagnostic workflows. The proposed model is designed to assist clinicians in distinguishing benign and malignant nodules more efficiently, particularly in settings with limited imaging expertise or infrastructure. By validating this approach using data from the Chinese General Hospital, including solitary and multiple PN cases, we seek to identify critical risk factors and establish a robust predictive tool that bridges the gap between radiomics and clinical decision-making.

This study utilizes actual clinical data from the Chinese General Hospital, including cases of solitary and multiple pulmonary nodules, aiming to explore the risk factors of pulmonary nodules and construct a predictive model for their benign and malignant outcomes, thereby providing a theoretical basis for determining the nature of pulmonary nodules in clinical practice. The model demonstrates robust performance in scenarios with limited radiological expertise and infrastructure, ensuring reliable predictions without requiring advanced imaging techniques or specialized equipment.

Materials

Patients

This study is based on electronic medical record data of outpatients and inpatients from January 2011 to December 2021 at the Chinese General Hospital. Inclusion criteria: 1) patients who underwent lung CT examination for the first time with a maximum nodule diameter between 5–30 mm and no evidence of malignant tumors in the lung

CT examination; 2) patients who had clear pathological results after the first lung CT examination suggesting pulmonary nodules. Exclusion criteria: 1) patients with a history of lung surgery; 2) patients who developed malignant tumors before the first lung CT examination suggesting pulmonary nodules; 3) patients who had malignant tumors between the first lung CT examination suggesting pulmonary nodules and the pathological diagnosis of benign nodules; 4) patients who had a history of malignant tumors between the first lung CT examination suggesting pulmonary nodules and the pathological diagnosis of malignant nodules.

Based on the above inclusion criteria, we counted all the visits of each patient who had undergone a lung CT examination suggesting pulmonary nodules with a diameter between 5–30mm and did not suggest malignant tumors in the database. Then, the first visit that met the criteria was considered the starting point of the study for each patient. A total of 89,829 patients were included, of which 83,969 patients did not have clear pathological results after the study's starting point. Additionally, 25 patients with a history of lung surgery were excluded, as well as 1243 patients with any diagnosis of malignant tumors before the first lung CT examination suggesting nodules. Furthermore, to ensure that the cohort consisted predominantly of single-nodule patients, 309 patients who developed malignant tumors between the first lung CT examination suggesting nodules and the pathological diagnosis of benign nodules were excluded, as well as 113 patients who had a history of malignant tumors between the first lung CT examination suggesting nodules and the pathological diagnosis of malignant nodules. Finally, 815 patients with a time interval greater than 180 days between the first lung CT examination suggesting nodules and the pathological diagnosis were also excluded. The specific recruitment is illustrated in Figure 1.

Statistical Analyses

Utilizing the robust Python package TableOne[16], we conducted statistical analysis of clinical features between malignant and benign pulmonary nodules. For continuous variables such as age and maximum nodule diameter, we calculated the mean ± standard deviation for both malignant and benign nodule groups. For categorical variables such as gender and the

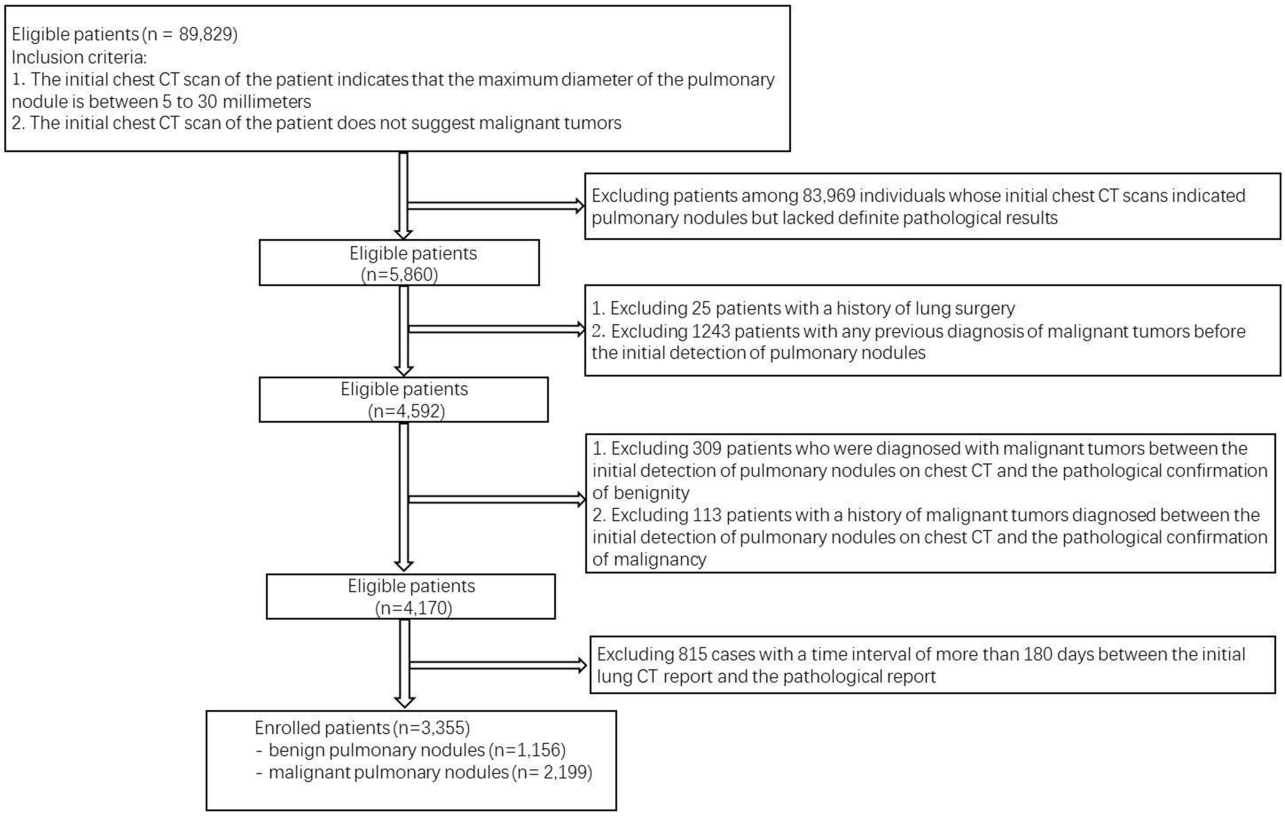

Figure 1 Overview of Study Design and Participant Enrollment.

 https://doi.org/10.2147/JMDH.S518166

presence of ground-glass opacity nodules, we calculated the percentage of each category within the malignant and benign nodule groups. Additionally, we calculated p-values for all variables to assess significant differences between the two groups.

The time distribution between the initial lung CT scan indicating pulmonary nodules and the subsequent pathology report is depicted in Figure 2. From the graph, it is evident that following the initial lung CT scan, pathology examinations are typically conducted within 60 days for approximately 83.70% of patients and within 90 days for approximately 91.45% of patients. Approximately 8.55% of patients undergo pathology examinations after a period exceeding 90 days.

Label Definition

By collecting patient data for each visit based on patient identification numbers, we determined the time at which lung nodules were first detected during the initial lung CT examination. Subsequently, if the pathological examination report included conclusions such as pulmonary malignant tumors, pulmonary cancer, pulmonary atypical adenomatous hyperplasia, or pulmonary lymphoma, the patient was classified as having malignant pulmonary nodules, with a total of 2199 such patients. The remaining patients whose pathological reports did not indicate malignant tumors were classified as having benign lung nodules, amounting to a total of 1156 patients.

Figure 3 depicts the distribution of benign and malignant nodules among all patients across different nodule diameter ranges, along with the proportion of malignant nodules. It can be observed that as the diameter of pulmonary nodules increases, the number of patients gradually decreases. Additionally, due to the requirement in this study for patients to have definitive pathological results following the initial indication of pulmonary nodules on CT scans, patients at higher risk of malignant nodules are more likely to be recommended for pathology examinations by clinical practitioners. Therefore, among the 3355 patients included in the study, the number of patients with malignant nodules exceeds the number of patients with benign nodules within each diameter range of nodules.

We categorized the patients we included in this study into the benign nodule patient group and the malignant nodule patient group based on pathological diagnosis. Within the malignant nodule group, diagnoses were classified into types such as adenocarcinoma, small cell lung cancer, squamous cell carcinoma, large cell lung cancer, adenosquamous carcinoma, and in situ adenocarcinoma. Within the benign nodule group, diagnoses were classified into types such as hemangioma, pneumonia, tuberculosis, pulmonary granuloma, leiomyoma, and bronchiolar epithelial hyperplasia. As shown in

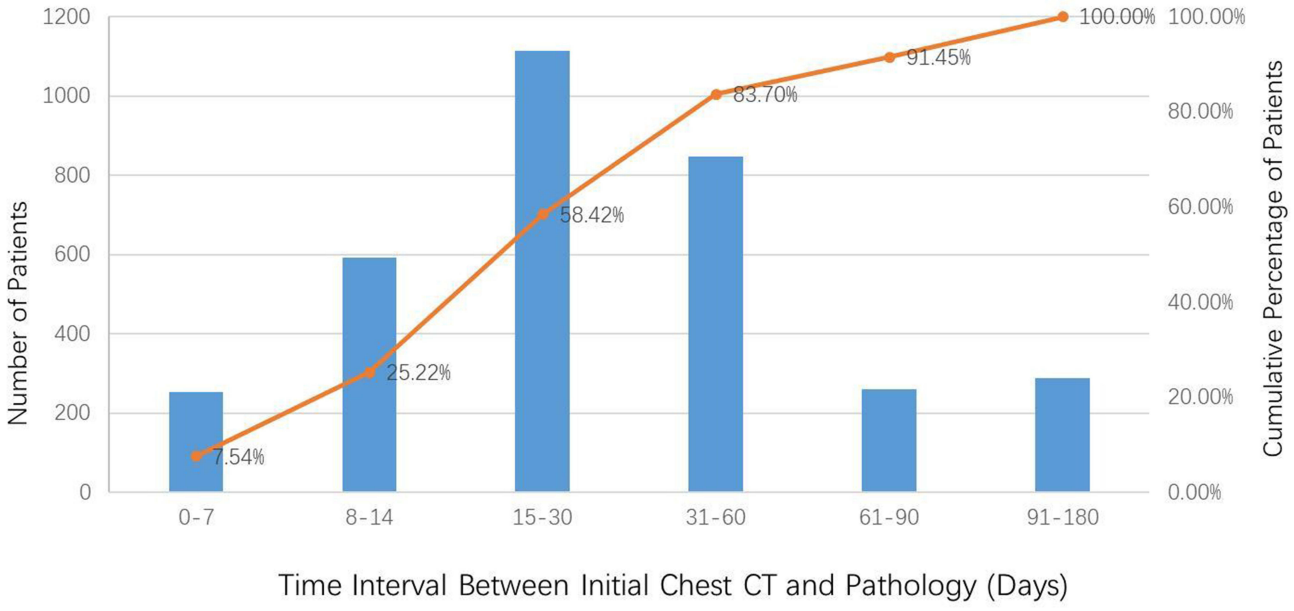

Figure 2 Time Interval Distribution Between Chest CT and Pathological Examination.

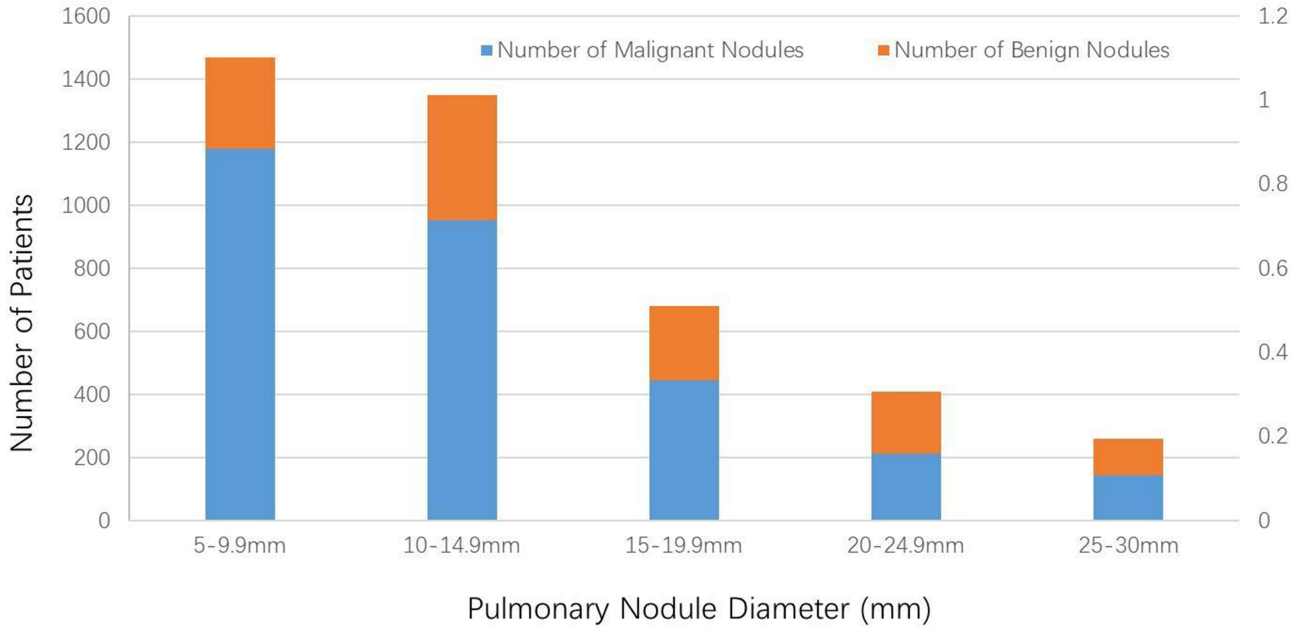

Figure 3 Pulmonary Nodule Size Distribution.

Figures 4 and 5, adenocarcinoma was the most common type in the malignant nodule group, accounting for nearly 1832 cases, far exceeding other types. Within the benign nodule group, the most common diagnosis was bronchiolar epithelial hyperplasia, with a total of 219 cases. This was followed by hemangioma, pneumonia, and pulmonary granuloma, each with approximately 50 cases, whereas the number of other pathological diagnoses was relatively small.

Methods

Inclusion of Features

We collected complete clinical data and relevant examination and laboratory test data for all enrolled patients, including gender, age (calculated as the difference between lung CT time and date of birth), BMI, carcinoembryonic antigen levels,

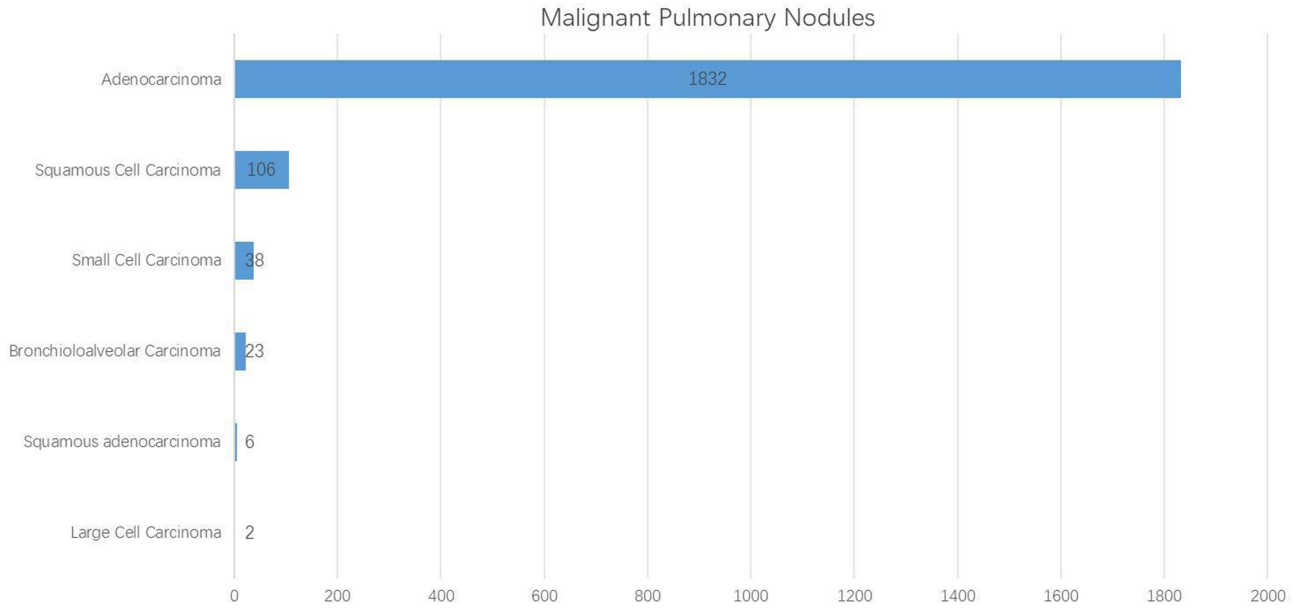

Figure 4 Distribution of Pathological Diagnosis of Malignant Pulmonary Nodules.

https://doi.org/10.2147/JMDH.S518166

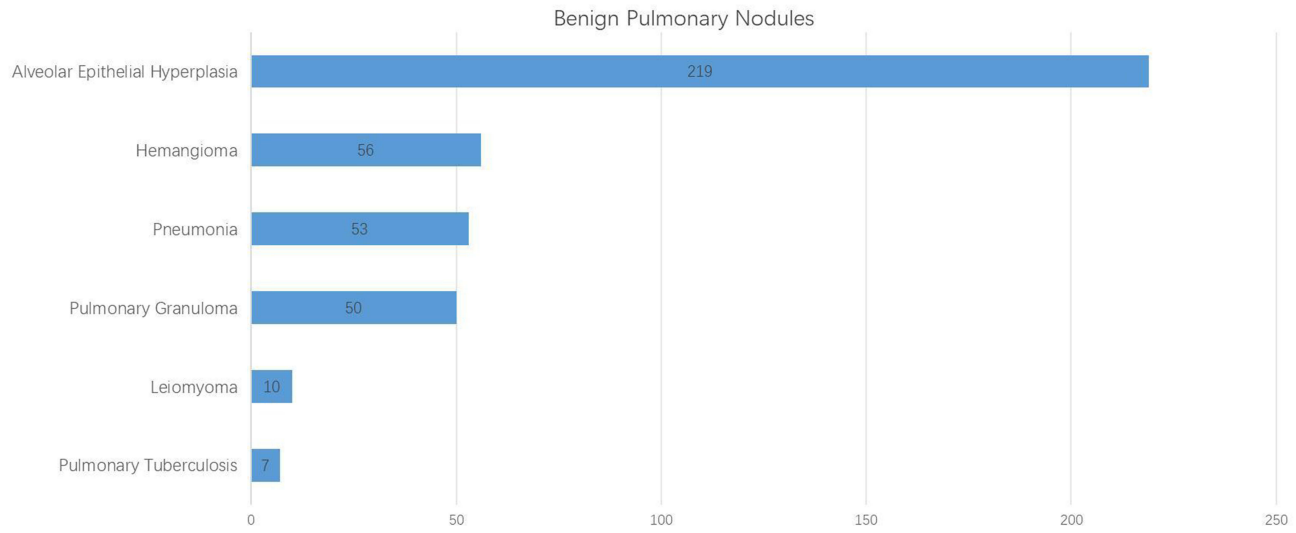

Figure 5 Distribution of Pathological Diagnosis of Benign Pulmonary Nodules.

specific neuroendocrine markers, cytokeratin 19 fragment determination, and squamous cell carcinoma-related antigen determination. Numerical variables were imputed using SimpleImputer (strategy="median") to handle missing values, ie, employing the mean for imputation. Additionally, we considered imaging features such as maximum diameter, shape, density, calcification status, clarity of abnormal margin, ground-glass opacity nodules, presence of spiculation, presence of lobulation, presence of cavitation, presence of vascular convergence sign, and presence of pleural indentation when evaluating the nodules. Model performance was evaluated using AUC, accuracy, sensitivity, and specificity.[5,17,18]

Feature Preprocessing

In identifying the types of features, all features were categorized into discrete features and continuous features. For discrete features, the One-Hot encoding method was used, where each dimension of the feature has values of 0 or 1, representing the presence or absence of the feature. All missing discrete features were considered absent and filled with 0. For continuous features, outliers were detected using the 3σ principle based on the normal distribution, with μ±3σ as the detection boundary. Values exceeding this boundary were regarded as outliers and treated as missing values. To mitigate potential bias from class imbalance, a simple undersampling operation was applied to the training set prior to model training. During model training, a 5-fold cross-validation strategy was employed to ensure robust evaluation of the model's performance. Finally, to reduce the negative impact of different feature scales and variances, the features were standardized using MaxAbs normalization[19]. To guarantee the reproducibility of the method, we fixed the random seed in all experimental steps, including data splitting, undersampling, and model initialization.

Different sections of electronic medical records may contain the same entity information, but they represent different clinical significance in clinical practice. For example, a disease mentioned in the chief complaint and present illness sections represents the patient's recent condition, whereas the same disease mentioned in the past medical history section represents the patient's previous condition. The impact on the judgment of the benign or malignant nature of pulmonary nodules varies. Therefore, in this study, features were concatenated based on sections to differentiate the same entities in different sections. Additionally, to address the differences in how different doctors describe symptoms, signs, and disease names, we standardized different entities that actually refer to the same object based on the British Medical Journal (BMJ) Best Practice knowledge base.[20]

Feature Selection

The current structured healthcare data contains abundant information, resulting in numerous features after applying One-Hot encoding to discrete-valued variables. Among these features, some may represent noise or have little to no relevance

to the current task. Therefore, it is necessary to filter out features to some extent, eliminating irrelevant ones, which can reduce computational complexity, enhance the model's predictive inference efficiency, reduce overfitting, and improve model generalizability.

Regularized linear models are powerful tools for feature understanding and selection. L1 regularization can produce sparse models, which are highly useful for selecting feature subsets. However, compared to L2 regularization, L1 is less user-friendly for data interpretation because the coefficients for features other than the selected high-quality ones tend to zero. Consequently, this study employs the Stability Selection method,[21] which merges bootstrap resampling with a feature selection algorithm based on L1 regularization (Lasso). The key concept involves repeatedly running the feature selection algorithm on different data and feature subsets. A particular feature's significance is tallied based on its selection frequency—that is, the number of times it is deemed important divided by the number of subsets in which it is tested—thus yielding a consolidated result for feature selection. The specific implementation process is illustrated in Figure 6.

Algorithm

We first performed information extraction on the unstructured data within the radiological findings and electronic health records using Natural Language Processing (NLP) techniques. Specifically, we employed a pre-trained Chinese BERT model (albert-base-Chinese-cluecorpussmall) for end-to-end processing of clinical texts: The multi_compose function was first used to complete text cleaning, sentence segmentation, synonym normalization and stop word filtering. Then, we utilized HuggingFace's BertTokenizer to encode each medical case into input_ids and attention_mask. Subsequently, we constructed a BERT + fully connected classifier structure in the NET class, and fine-tuned the model with cross-entropy loss. Finally, the [CLS] vector or pooled output of each sample was directly used for classification to validate the incremental value of NLP features in downstream prediction tasks. Based on the extracted information, we applied stable feature selection methods to identify salient factors that significantly impact the diagnosis of pulmonary nodule benignity or malignancy. Subsequently, we utilized and compared various machine learning models and deep learning models, specifically including models such as Support Vector Machine(SVM), Logistic Regression(LR), the probability statistics-based Naive Bayes, Random Forest which employs a Bagging strategy and aggregates multiple decision trees, as well as GBDT, XGBoost, and LightGBM based on a Boosting strategy. In terms of deep learning, the Feedforward Neural Network (FNN) was employed. The FNN model structure used in this study is shown in Figure 7.

Model Explanation

We employed the SHapley Additive exPlanations (SHAP) method[22] to investigate the correlation between features and model predictions and to provide corresponding feature explanations. SHAP is a framework based on additive feature attribution methods that computes SHAP values for each feature of every sample. These values reflect the extent and direction of a feature's influence on the model's predictive outcome.

In this experiment, the absolute magnitude of SHAP values for each sample and feature reflects the degree of influence of that feature on the model's prediction outcome, while the sign of SHAP values indicates the direction of its

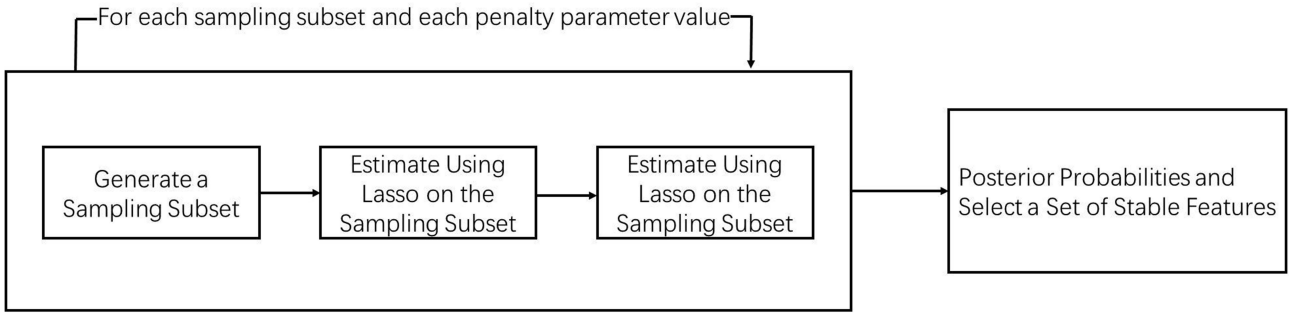

Figure 6 Stability Feature Selection Results.

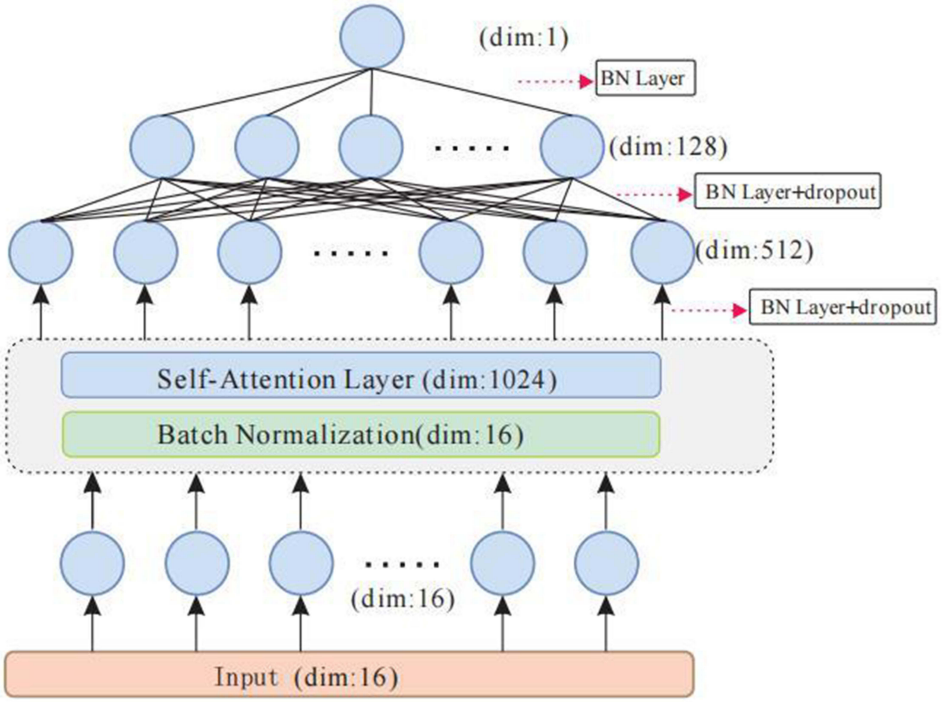

Figure 7 Architecture of the FNN with Self-Attention Mechanism.

influence. Specifically, when the SHAP value is greater than 0, it indicates that the feature supports the model's prediction of a higher risk of malignant lung nodules. Conversely, when the SHAP value is less than 0, it signifies the opposite. In recent years, SHAP has gradually been applied to the interpretation of predictive models.[23,24]

Combining the aforementioned content, our overall workflow is illustrated in Figure 8. Initially, historical medical records were included in the study based on predefined inclusion criteria. Subsequently, unstructured data underwent Natural Language Processing (NLP) tokenization to convert it into structured data. Following this, a feature set was curated based on the structured data, and various processes such as discretization, standardization, outlier detection, and vectorization were applied to the feature set. Next, we compared various machine learning and deep learning algorithms to train the most optimal predictive model. Finally, leveraging the optimal predictive model, we employed the SHAP method to explain the factors influencing the predictions.

Result

Feature Selection Results

This study employed a stable feature selection method, wherein the least useful features have scores close to zero. Consequently, features with scores close to zero were excluded. The excluded features included: cytokeratin 19 fragment determination, squamous cell carcinoma-related antigen determination, etc.

Experimental Results

Training was conducted on multiple comparison models, and they were evaluated using the 5-Fold cross-validation method. The average results of the 5-Fold experiment are presented in Table S1. It was observed that the accuracy of ensemble algorithms based on the Boosting strategy, as well as neural network algorithms, was significantly higher than those of LR (Logistic Regression), SVM (Support Vector Machine), and Naive Bayes. The accuracy of RForest (Random Forest), based on the Bagging strategy, was lower than that of LR but higher than those of both SVM and Naive Bayes. Among them, the accuracy and sensitivity of Atten_FNN (Attention-enhanced Feedforward Neural Network) were the highest, and the precision and specificity of RForest are the greatest. However, the sensitivity and F1 score of the RForest

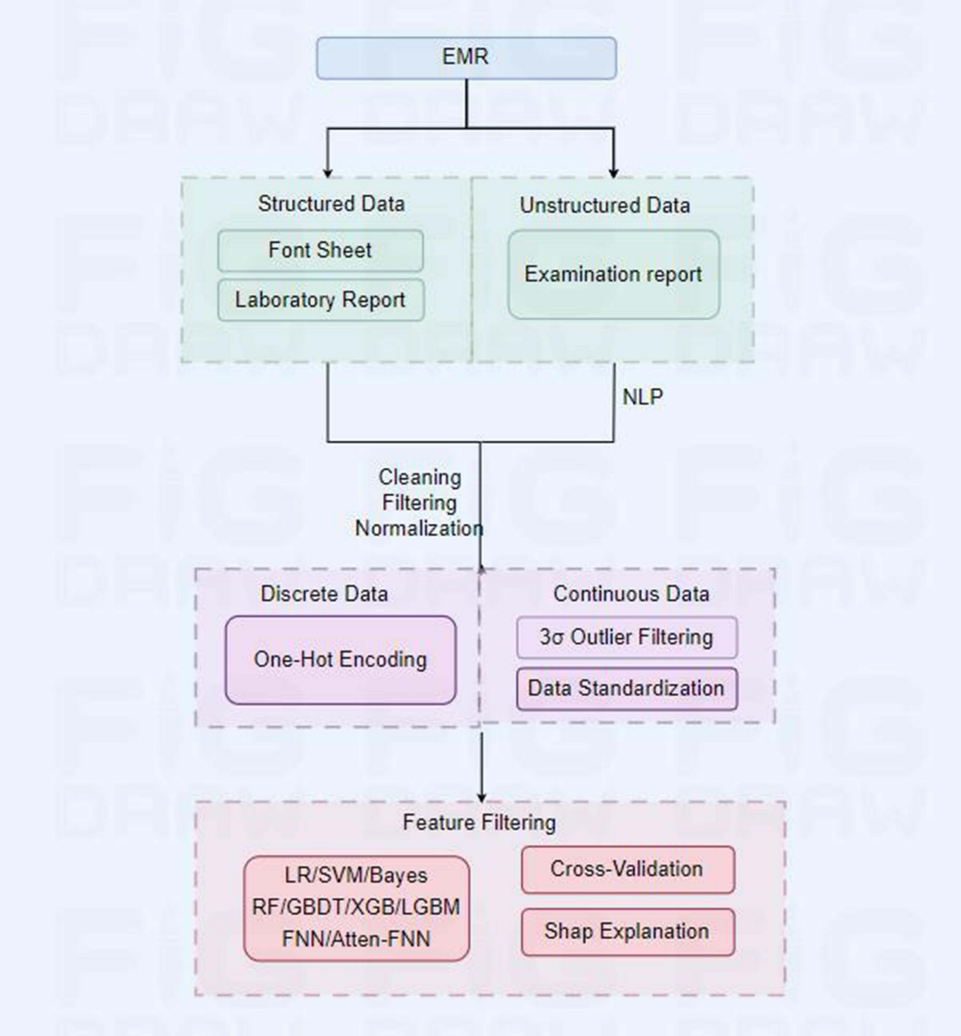

Figure 8 Workflow for Model Development and Evaluation.

algorithm were noticeably lower than those for Atten_FNN. Considering that this research is aimed at the early detection of malignant risks in lung nodules, which requires relatively high sensitivity, an overall evaluation of the experimental indicators led to the selection of Atten_FNN as the best practice model for predicting lung nodules.

Model Explanation

According to the SHAP method, we performed feature explanation on the Atten_FNN model. Figure 9 lists 16 features with varying degrees of impact on the model, with feature importance calculated by the mean of the absolute values of all SHAP values for each feature. From the figure, it is evident that the feature with the greatest influence on the prediction of the malignancy of lung nodules is the ground-glass nodule, followed by age, the largest nodule diameter, and nodule shape (near round), among others.

Figure 10 illustrates the distribution of the top 16 features' impact on model predictions. The horizontal axis shows SHAP values, where SHAP values > 0 indicate support for malignant pulmonary nodules predictions and SHAP values < 0 indicate support for benign pulmonary nodules predictions. Furthermore, the larger the absolute value of the SHAP score, the greater the influence of the feature on the prediction outcome. The vertical axis represents a temperature bar, illustrating the magnitude of each feature's value, where red indicates larger feature values and blue indicates smaller feature values. Combining the SHAP values on the horizontal axis with the temperature bar on the vertical axis allows for a clearer understanding of each feature's impact direction and magnitude on the prediction outcome across different feature values.

https://doi.org/10.2147/JMDH.S518166

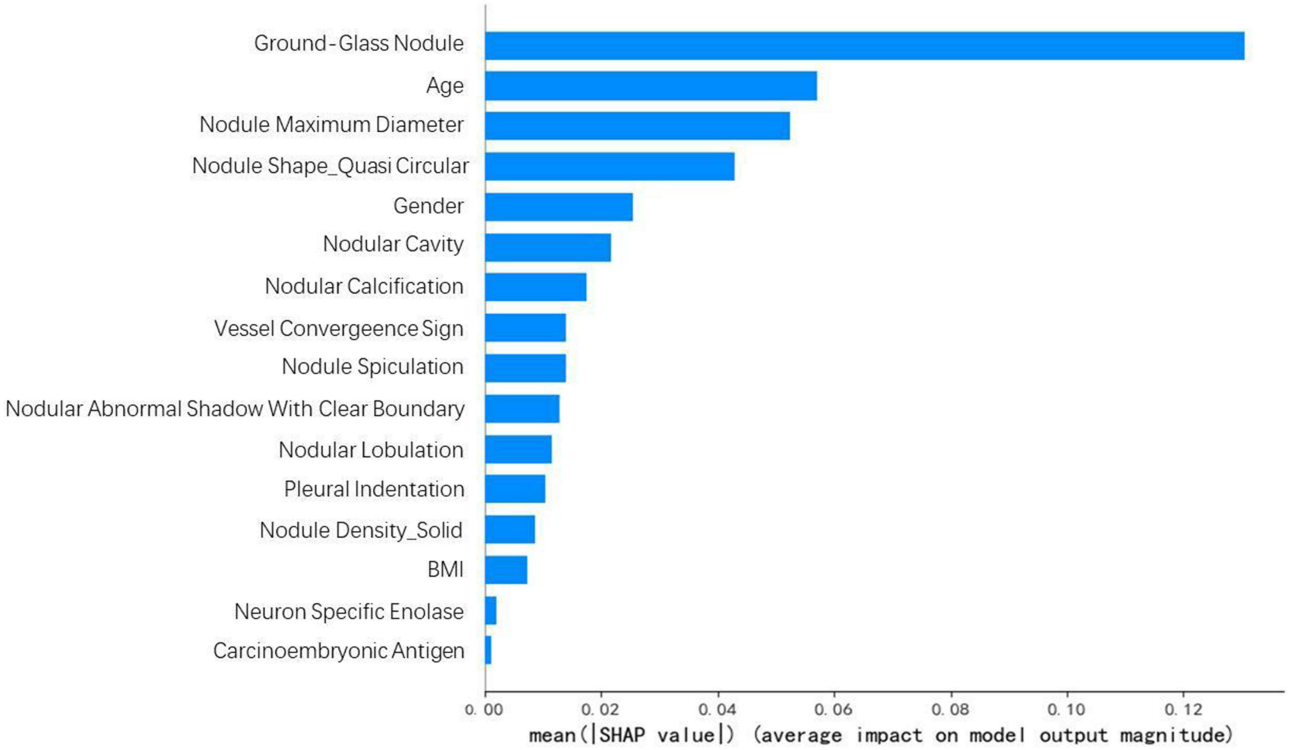

Figure 9 SHAP Analysis: Contribution of Top 16 Features to Atten_FNN Predictions.

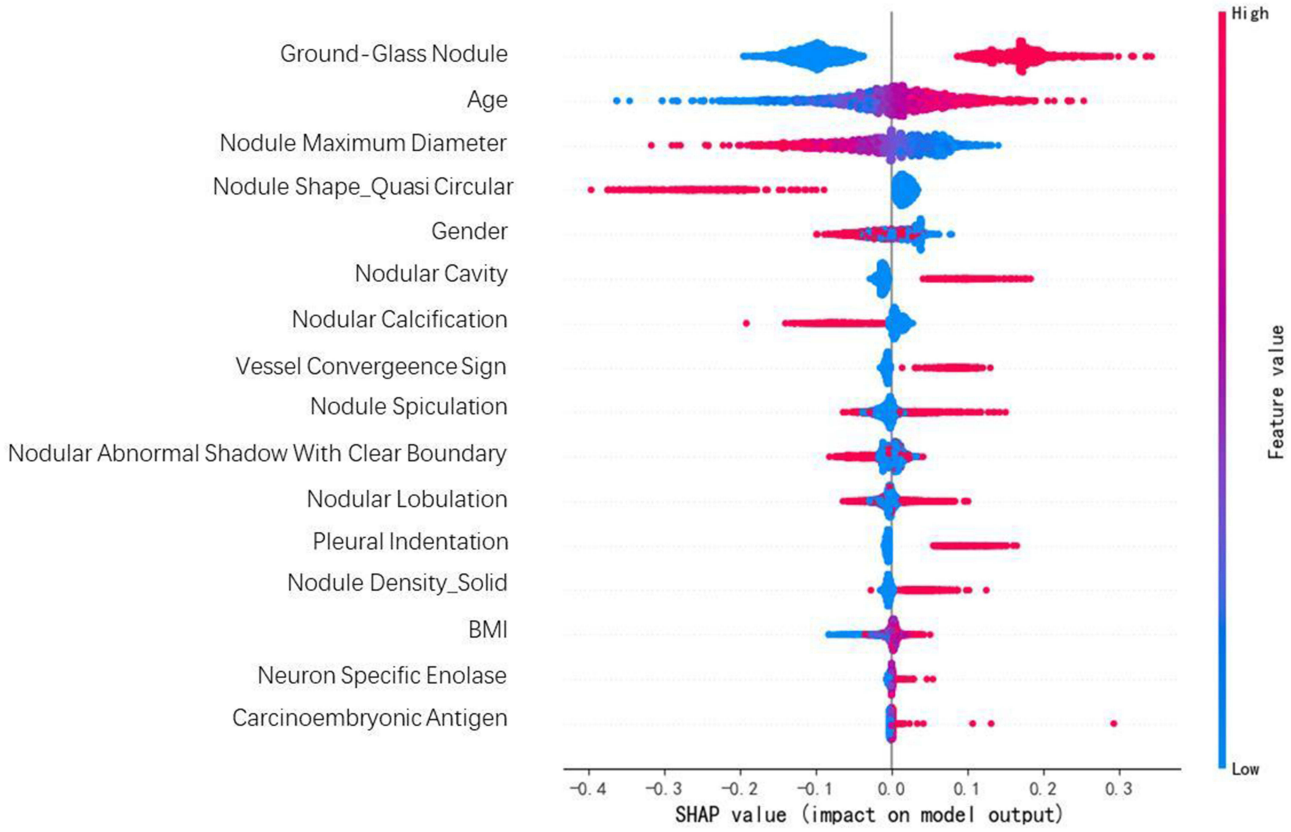

Figure 10 SHAP Analysis: Feature Impact on Model Output.

For example, features such as ground-glass opacity nodules, nodular cavitation, nodular spiculation, nodular lobulation, pleural indentation, and solid density nodules typically suggest a higher likelihood of malignant pulmonary nodules in patients. Conversely, nodules exhibiting a near-round shape, calcification, and clear abnormal margins usually indicate a higher likelihood of benign pulmonary nodules. Additionally, it was observed that female patients are more likely to have malignant pulmonary nodules than male patients. Moreover, as age, BMI, specific neural isoenzyme, and carcinoembryonic antigen levels increase, the likelihood of malignant pulmonary nodules also gradually increases.

Discussion

Artificial intelligence and deep learning (DL) are increasingly recognized for their potential in early epidemic disease diagnosis. Traditional models, limited by their inability to capture complex dependencies and low interpretability, often produce inaccurate predictions.[5] In contrast, DL models such as FNN and Atten_FNN, used by Haizhe Jin and SS Ghosal, demonstrate superior performance.[25] Among various DL architectures, FNN-based computer-aided diagnosis demonstrates superior performance.[26] Inspired by the successful application of attention mechanisms in medical imaging tasks, an enhanced Atten_FNN model was proposed.[27–31]This model employs attention mechanisms to automatically learn input-output relationships and quantify similarities via softmax-normalized inner products.[32,33] These advanced models hold significant prognostic value for early prediction and classification of pulmonary nodules.[34]

This study established a large database of 89,829 patients with pulmonary nodules under strict inclusion and exclusion criteria. All participants had nodules measuring 5–30 mm in diameter detected during their initial lung CT scan, with pathological confirmation. As shown in Table 1, the MPN group was older and had a higher female proportion, while the benign group exhibited a significantly larger average maximum nodule diameter. Features such as ground-glass opacity(GGO), cavitation, vascular convergence, solid density, lobulation, and pleural indentation were significantly more prevalent in the malignant group, whereas calcification, indistinct margins, and round shape were more common in the benign group. Incorporating these clinical features enhances our predictive model's ability to comprehensively assess patients'health status and disease risk.

Table 1 Baseline Characteristics of Patients with MPN and BPN

Characteristics	MPN(n=2199)	BPN(n=1156)	P_value
Demographic data			
Age, mean (SD)	54.91 (10.98)	51.23 (11.53)	<0.001
BMI, mean (SD)	24.36 (3.24)	24.62 (3.39)	0.040
Male, n (%)	890 (40.5)	645 (55.8)	<0.001
Lab results			
Carcinoembryonic Antigen, mean (SD)	4.68 (25.78)	2.16 (2.37)	0.008
Neuron Specific Enolase, mean (SD)	12.70(9.61)	11.98 (4.60)	0.058
Cellular Protein 19 Fragment, mean (SD)	7.86 (186.55)	2.47 (3.27)	0.460
Squamous Cell Carcinoma Associated Antigen, mean (SD)	0.95 (1.33)	0.92 (2.68)	0.717
Chest CT examination			
Nodule Maximum Diameter, mean (SD)	12.70 (5.75)	14.89 (6.23)	<0.001
Nodular Abnormal Shadow With Clear Boundary, n (%)	513(23.3)	378 (32.7)	<0.001
Ground-Glass Nodule, n (%)	1427 (64.9)	245 (21.2)	<0.001
Nodule Spiculation, n (%)	487(22.1)	233 (20.2)	0.197
Nodular Lobulation, n (%)	659 (30.0)	266(23.0)	<0.001
Nodular Cavity, n (%)	329 (15.0)	107(9.3)	<0.001
Vascular bundle sign, n (%)	330 (15.0)	50(4.3)	<0.001
Pleural Indentation, n (%)	182 (8.3)	48 (4.2)	<0.001
Nodular Calcification, n (%)	301 (13.7)	214 (18.5)	<0.001
Nodule Density_Solid, n (%)	316 (14.4)	121 (10.5)	0.002
Nodule Shape_Quasi Circular, n (%)	140 (6.4)	208 (18.0)	<0.001

The study employed stability-based feature selection methods (Figure 6) to identify key predictors of pulmonary nodule malignancy, including imaging features (ground-glass nodules, diameter, shape, cavitation, calcification) and clinical factors (age, gender, BMI, carcinoembryonic antigen levels). Experiments compared multiple machine learning algorithms (SVM, LR, Bayes, RForest, GBDT, XGBoost, LightGBM) and deep learning models (FNN andAtten_FNN). AlthoughXGBoost achieved competitive performance, Atten_FNN balanced complexity and interpretability by quantifying feature contributions via attention weights. Selected as the optimal model for its high sensitivity (reducing missed diagnoses) and specificity (minimizing misdiagnoses), Atten_FNN enables accurate early clinical intervention, improving survival rates in lung cancer patients.

The study employed SHAP analysis to interpret model predictions, clarifying the distributional impact of feature values on predictions and their specific influences. The most influential feature for predicting pulmonary nodule malignancy was the presence of GGO, age, maximum nodule diameter, and round-shaped. GGOs strongly indicated malignant nodules, while round-shaped nodules were associated with benign diagnoses. Gender analysis revealed higher malignancy likelihood in female patients. Increased biomarkers such as age, BMI, and elevated carcinoembryonic antigen levels were also associated with elevated malignant risk (Figure 10). SHAP visualizations provided in-depth feature impact interpretation, enhancing model transparency and offering robust clinical decision-making support.

This study, based on electronic medical records and machine learning modeling, developed a predictive model for predicting the early malignant risk of pulmonary nodules. The model demonstrated high accuracy, assisting clinicians in promptly performing pathological examinations and making rapid diagnostic and treatment decisions for patients with pulmonary nodules. Additionally, it helped reduce the number of pathological examinations required for patients with benign nodules. Future prospective studies could further assess the model's impact on clinical practice, including its influence on physician behavior and its value in predicting malignant risk and improving patient outcomes.

This single-center study is limited by institutional-specific clinical workflows and potential patient selection bias, which may affect generalizability and warrant external validation. As an exploratory phase focused on feature selection and discriminative performance, we recognize the critical need for calibration analysis to align predicted probabilities with observed frequencies and avoid misleading biopsy decisions. Subsequent work will incorporate calibration metrics to enhance clinical reliability. The exclusion of chest CT images data due to accessibility constraints will be addressed by developing EMR + CT–based models for accuracy comparison by developing EMR + CT–based models for accuracy comparison.

Limitation

Firstly, this study is a single-center investigation, limited by the specific clinical diagnosis and treatment protocols of the hospital and potential patient selection bias, which may restrict the applicability of the results to other medical institutions. Therefore, further external validation studies are necessary to confirm the generalizability of the findings.

Secondly, due to the challenges in obtaining imaging data, this study did not incorporate chest CT images, potentially leading to the loss of valuable patient information. Future research could explore predictive models for the benign and malignant classification of pulmonary nodules based on both electronic medical records and CT images, to compare the predictive accuracy with the model used in this study.

Lastly, due to the retrospective nature of this study, the impact of the predictive model on clinical practice was not evaluated, as it falls outside the scope of this research. Future studies could assess the influence of the predictive model on clinicians' clinical behavior through prospective studies, as well as its value in improving the prediction of malignant risk of pulmonary nodules and patient prognosis.

Data Sharing Statement

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

The approval number of the ethics committee of the Chinese PLA General Hospital is S2022-203-01.

Consent for Publication

No published individual participant data were reported that would require consent from the participants.

Declarations

The studies involving human participants were reviewed and approved by the ethics committee of the Chinese People's Liberation Army General Hospital. The ethics committee waived the requirement of written informed consent for participation because this study was a retrospective analysis of patient medical records, which did not involve additional interventions or risks to the participants, and obtaining individual consent was impracticable. All patient data were anonymized and handled with strict confidentiality to protect privacy. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Author Contributions

All authors made significant contributions to the work of the report, whether in conception, research design, execution, data acquisition, analysis and interpretation, or in all of these areas. Participated in the drafting, revision or critical review of the manuscript; gave final approval of the version to be published; agreed to submit the article to this journal; And agreed to be responsible for all aspects of the work.

Funding

There is no funding to report.

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Loomis D, Grosse Y, Lauby-Secretan B, et al. The carcinogenicity of outdoor air pollution. *Lancet Oncol*. 2013;14(13):1262–1263. doi:10.1016/S1470-2045(13)70487-X
2. Pei Z, Wu M, Zhu W, et al. Associations of long-term exposure to air pollution with prevalence of pulmonary nodules: a cross-sectional study in Shijiazhuang, China. *Ecotoxicol Environ Saf*. 2023;262:115311.
3. International Agency for Research on Cancer. Global cancer burden growing, amidst mounting need for services. *Saudi Med J*. 2024;45(3):326–327.
4. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
5. Warkentin MT, Al-Sawaihey H, Lam S, et al. Radiomics analysis to predict pulmonary nodule malignancy using machine learning approaches. *Thorax*. 2024;79(4):307–315. doi:10.1136/thorax-2023-220226
6. Pettit NR, Horner D, Freeman S, Rieger K. Pilot intervention to increase uptake of lung cancer screening through the emergency department. *Am J Emerg Med*. 2024;79:157–160. doi:10.1016/j.ajem.2024.02.019
7. Sim YT, Goh YG, Dempsey MF, Han S, Poon FW. PET-CT evaluation of solitary pulmonary nodules: correlation with maximum standardized uptake value and pathology. *Lung*. 2013;191(6):625–632. doi:10.1007/s00408-013-9500-6
8. Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: a review. *JAMA*. 2022;327(3):264–273. doi:10.1001/jama.2021.24287
9. Huang S, Yang J, Shen N, Xu Q, Zhao Q. Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. *Semin Cancer Biol*. 2023;89:30–37. doi:10.1016/j.semcancer.2023.01.006
10. Li Y, Jiang G, Wu W, et al. Multi-omics integrated circulating cell-free DNA genomic signatures enhanced the diagnostic performance of early-stage lung cancer and postoperative minimal residual disease. *eBioMedicine*. 2023;91:1.
11. Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: a comprehensive review on current approach and management. *Ann Thorac Med*. 2019;14(4):226–238. doi:10.4103/atm.ATM_110_19
12. Tietz E, Müller-Franzes G, Zimmermann M, et al. Evaluation of pulmonary nodules by radiologists vs. radiomics in stand-alone and complementary CT and MRI. *Diagnostics*. 2024;14(5):483. doi:10.3390/diagnostics14050483
13. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. *Int J Comput Assist Radiol Surg*. 2017;12(10):1799–1808. doi:10.1007/s11548-017-1605-6
14. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. *JAMA Network Open*. 2020;3(9):e2017135. doi:10.1001/jamanetworkopen.2020.17135
15. Venkadesh KV, Setio AAA, Schreuder A, et al. Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. *Radiology*. 2021;300(2):438–447. doi:10.1148/radiol.2021204433
16. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: an open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1(1):26–31. doi:10.1093/jamiaopen/ooy012

17. Zhao Z, Gui J, Yao A, Le NQK, Chua MCH. Improved prediction model of protein and peptide toxicity by integrating channel attention into a convolutional neural network and gated recurrent units. *ACS omega*. 2022;7(44):40569–40577. doi:10.1021/acsomega.2c05881
18. Kha Q-H, Tran T-O, Nguyen V-N, Than K, Le NQK, Le NQK. An interpretable deep learning model for classifying adaptor protein complexes from sequence information. *Methods*. 2022;207:90–96. doi:10.1016/j.ymeth.2022.09.007
19. Shalabi LA, Shaaban Z, Kasasbeh B. Data mining: a preprocessing engine. *J Comput Sci*. 2006;2(9):735–9.
20. Tao L, Zhang C, Zeng L, et al. Accuracy and effects of clinical decision support systems integrated with bmj best practice–aided diagnosis: interrupted time series study. *JMIR Med Inform*. 2020;8(1):e16912. doi:10.2196/16912
21. Meinshausen N, Bühlmann P. Stability Selection. *J R Stat Soc Ser B*. 2010;72(4):417–473. doi:10.1111/j.1467-9868.2010.00740.x
22. Van den Broeck G, Lykov A, Schleich M, Suciu D. On the tractability of SHAP explanations. *J artif intell res*. 2022;74:851–886.
23. Jiang Y, Zhao Q, Guan J, Wang Y, Chen J, Li Y. Analyzing prehospital delays in recurrent acute ischemic stroke: insights from interpretable machine learning. *Patient Educ Couns*. 2024;123:108228. doi:10.1016/j.pec.2024.108228
24. Wen X, Xie Y, Wu L, Jiang L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid Anal Prev*. 2021;159:106261. doi:10.1016/j.aap.2021.106261
25. Ghosal SS, Sarkar I, El Hallaoui I. Lung nodule classification using convolutional autoencoder and Clustering Augmented Learning Method (CALM). In: *HSDM@ WSDM*. 2020.
26. Iqbal SN, Qureshi A, Li J, Mahmood T. On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Arch Comput Methods Eng*. 2023;30(5):3173–3233. doi:10.1007/s11831-023-09899-9
27. Lu X, Chang EY, Hsu C-N, Du J, Gentili A. Multi-classification study of the tuberculosis with 3D CBAM-ResNet and EFFICIENTNET. In: *CLEF (Working Notes)*. 2021.
28. Nawshad MA, Shami UA, Sajid S, Fraz MM. Attention based residual network for effective detection of covid-19 and viral pneumonia. 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2); 2021: IEEE.
29. Sangeroki BA, Cenggoro TW. A fast and accurate model of thoracic disease detection by integrating attention mechanism to a lightweight convolutional neural network. *Procedia Comput Sci*. 2021;179:112–118. doi:10.1016/j.procs.2020.12.015
30. Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint, arXiv*. 2021;2021:1.
31. Almahasneh M, Xie X, Paiement A. Attentnet: fully convolutional 3d attention for lung nodule detection. *SN Comput Sci*. 2025;6(3):292. doi:10.1007/s42979-025-03799-4
32. Cao C, Yang S, Li M, Li C. CircSSNN: circRNA-binding site prediction via sequence self-attention neural networks with pre-normalization. *BMC Bioinf*. 2023;24(1):220. doi:10.1186/s12859-023-05352-7
33. Zeng Y, Chen X, Luo Y, Li X, Peng D. Deep drug-target binding affinity prediction with multiple attention blocks. *Briefings Bioinf*. 2021;22(5):bbab117. doi:10.1093/bib/bbab117
34. Apostolopoulos ID, Papathanasiou ND, Apostolopoulos DJ, Papandrianos N, Papageorgiou EI. A multi-modal machine learning methodology for predicting solitary pulmonary nodule malignancy in patients undergoing PET/CT examination. *Big Data Cogn Comput*. 2024;8(8):85. doi:10.3390/bdcc8080085