



## OPEN ACCESS

## EDITED BY

Pierpaolo Ferrante,  
National Institute for Insurance Against  
Accidents at Work (INAIL), Italy

## REVIEWED BY

Lamiaa L. M. Ebraheim,  
Zagazig University, Egypt  
Charles Maibvise,  
University of Hertfordshire, United Kingdom

## \*CORRESPONDENCE

Biquan Zhang  
✉ bq\_2003@163.com

RECEIVED 19 October 2025

REVISED 16 December 2025

ACCEPTED 24 December 2025

PUBLISHED 15 January 2026

## CITATION

Ma Y, Ye L, Pan J, Shen D, Wang Q, Song B,  
Shen Y, Zhu X, Chen F, Shi J, Ye Q, Qin S,  
Ren R, Luo X, Xu J, Zhao J, Zhu D, Zhou Q,  
Zhu Y and Zhang B (2026) Lung involvement  
percentage in patients with COVID-19 during  
the Omicron wave in China: a  
SHAP-explained machine learning study from  
a single center.  
*Front. Public Health* 13:1728282.  
doi: 10.3389/fpubh.2025.1728282

## COPYRIGHT

© 2026 Ma, Ye, Pan, Shen, Wang, Song, Shen,  
Zhu, Chen, Shi, Ye, Qin, Ren, Luo, Xu, Zhao,  
Zhu, Zhou, Zhu and Zhang. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Lung involvement percentage in patients with COVID-19 during the Omicron wave in China: a SHAP-explained machine learning study from a single center

Yuhang Ma, Li Ye, Jing Pan, Dongyuan Shen, Qiang Wang, Bin Song, Yiliang Shen, Xiaoqiang Zhu, Feng Chen, Jian Shi, Qin Ye, Siwei Qin, Rong Ren, Xin Luo, Jun Xu, Jianzhong Zhao, Dongxing Zhu, Qiujuan Zhou, Yiming Zhu and Biquan Zhang\*

Department of Radiology, Suzhou Hospital of Integrated Traditional Chinese and Western Medicine, Suzhou, China

**Background:** Following the lifting of China's stringent lockdown policy on December 7, 2022, COVID-19 cases surged in a pattern, creating unprecedented strain on healthcare systems. The Omicron variant, characterized by high transmissibility and rapid spread, led to a sharp rise in infections. Understanding its clinical impact—particularly on lung involvement percentage—is crucial for optimizing patient care under such outbreak conditions. This study aimed to assess the extent of lung involvement percentage during the outbreak and its major associations.

**Methods:** The hospital's daily computed tomography examination volume was quantified using artificial intelligence-based pulmonary inflammation analysis software and used as an indicator of epidemic intensity. Associations between lung involvement percentage and age, sex, and daily case counts were evaluated using GEE Logistic Regression, complemented by machine learning models. Model interpretation was performed using SHapley Additive exPlanations.

**Results:** GEE Logistic regression demonstrated that age was strongly associated with lung involvement (OR 1.0813, 95% CI 1.0703–1.0925,  $p < 0.0001$ ), while daily case counts also showed a small but significant independent association (OR 1.0033, 95% CI 1.0018–1.0047,  $p < 0.0001$ ). Sex exhibited only minimal association (OR 0.8098, 95% CI 0.6983–0.9391,  $p = 0.0053$ ). Complementary machine learning analyses, including gradient boosting, identified age as the dominant contributor, followed by daily case counts with a small effect and sex with minimal contribution. SHAP analysis provided interpretable insights into how each feature influenced model predictions at both global and individual levels.

**Conclusion:** During the Omicron surge, greater age and higher daily case counts were associated with higher lung involvement percentage. These associations highlight the relevance of demographic and epidemic factors in characterizing pulmonary findings during large-scale outbreaks.

## KEYWORDS

advanced age, COVID-19, daily case counts, lung involvement percentage, machine learning, Omicron, outbreak, SHAP

## Introduction

Since the emergence of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the pandemic has posed a significant global health challenge. After multiple variant-driven waves worldwide, China lifted stringent lockdown measures in China on December 7, 2022, Omicron's high transmissibility led to a rapid increase in infections, exerting immense pressure on healthcare systems. Research to date has mainly explored the COVID-19 epidemiology and clinical features, as well as factors such as age, gender, and case counts.

### Age-related characteristics of COVID-19

Age is a critical predictor of COVID-19 severity and mortality. Previous studies have shown that increasing age is closely associated with declining immune function and the accumulation of underlying comorbidities, making older patients more susceptible to developing severe disease (1). However, in clinical practice, relying solely on age to assess individual risk may result in misclassification, potentially affecting resource allocation and patient management.

### Sex as a stratification factor

Numerous studies have demonstrated significant differences between males and females in clinical presentation and outcomes following SARS-CoV-2 infection, which may be attributed to sex-related variations in immune responses, hormone levels, and behavioral factors (2). Males generally exhibit lower innate and adaptive immune responses, including reduced CD4<sup>+</sup> T cell counts, diminished CD8<sup>+</sup> T cell cytotoxic activity, and decreased immunoglobulin production by B cells (3). Moreover, physiological responses to viral infections differ between sexes, with reports indicating that the immune system of females is approximately twice as robust as that of males (4). Sex-related differences in immune response efficiency are associated with disease outcomes, with males exhibiting a significantly higher risk of COVID-19-related mortality than females (hazard ratio 1.59) (5). Some researchers have linked these findings to genes located on the X chromosome (6). Estrogen can enhance the immunomodulatory activity of vitamin D, thereby improving infection outcomes (7). In contrast, male sex hormones may increase susceptibility to COVID-19 and worsen disease prognosis. First, they are thought to facilitate viral entry by upregulating angiotensin-converting enzyme 2 (ACE2) receptors, the entry point for SARS-CoV-2. Second, testosterone exhibits immunosuppressive effects, potentially dampening antibody responses. Males may benefit from T cell immune stimulants and anti-testosterone interventions, while estrogen could be utilized to reduce COVID-19 disease severity (8). A study conducted in China reported that males and females had similar susceptibility to infection. However, independent of age, infected males faced poorer clinical outcomes and a higher risk of mortality (9).

## Impact of epidemic intensity on COVID-19 severity

During periods of strict lockdown in China, sporadic COVID-19 cases generally exhibited mild disease severity. In contrast, under non-lockdown conditions, the severity of cases appeared to increase more markedly than anticipated, suggesting that the intensity of viral spread may significantly influence disease outcomes. It is hypothesized that higher transmission rates lead to increased viral loads, making patients more susceptible to severe complications such as acute respiratory distress syndrome (ARDS) and multiple organ failure; however, direct evidence supporting this hypothesis remains limited. Additionally, during outbreak peaks, shortages of healthcare resources—such as hospital beds and ICU capacity—may further exacerbate mortality risk (10, 11).

Despite these advances, prior studies have largely been conducted under conditions of strict containment or during earlier, less transmissible SARS-CoV-2 variants. Consequently, evidence describing the relationship between age, sex, daily case counts, and CT-based lung involvement during a large-scale outbreak without lockdown remains limited. In particular, few studies have examined how epidemic intensity, measured by daily case counts, may influence radiographic severity in real-world settings where healthcare systems are under substantial pressure. To address these gaps, the present study aims to characterize the distribution of CT-determined lung involvement percentage and to evaluate its association with age, sex, and daily case counts.

## Materials and methods

### Study design

This study is a retrospective case series of consecutive patients who underwent chest CT during the Omicron outbreak in Suzhou. Machine learning-based predictive modeling was applied alongside classical statistical analyses to evaluate associations between lung involvement and key patient- and population-level factors.

### Setting

The study was conducted at Suzhou Hospital of Integrated Traditional Chinese and Western Medicine, a hospital serving a mixed urban-rural population. The abandonment of China's zero-COVID policy on December 7, 2022, coincided with the spread of the highly transmissible Omicron variant. This combination led to a marked increase in daily CT volume during the late 2022 and early 2023.

### Participants

All consecutive chest CT scans performed at the hospital during the study period were screened. Inclusion criteria: (1) chest CT performed between December 14, 2022, and January 10, 2023; (2) exclusion of pneumonia caused by other infections; (3) only the first CT per patient was included.

## Variables

**Outcome:** Lung involvement percentage, defined as the percentage of lesioned lung, quantified using AI-based lung segmentation.

**Positivity of lung involvement,** defined as a binary variable indicating the presence or absence of lung lesions, derived from lung involvement percentage using a predefined threshold.

**Factor:** Age, sex, and the daily case counts of chest CT examinations performed at the hospital.

## Data sources/measurement

Each case was labeled as positive based on radiologist report-derived findings for preliminary exploratory analysis. Lung involvement percentage was quantified using AI-based segmentation software, which automatically segmented lung parenchyma and generated a continuous measure of involved lung tissue for subsequent modeling. AI-derived lung involvement maps were independently reviewed by two radiologists to identify and exclude false-positive findings caused by respiratory motion artifacts, dependent atelectasis, and obvious chronic pulmonary lesions that could be misclassified as COVID-19 pneumonia.

## Statistical methods

Associations were examined using generalized estimating equations (GEE) logistic regression. To address potential correlation among patients scanned on the same day, we specified an exchangeable working correlation structure and treated scanning day as the clustering variable. This approach provides consistent estimates of regression coefficients while adjusting standard errors for within-cluster correlation. Odds ratios (ORs) with 95% confidence intervals (CIs) were derived from model coefficients to quantify the strength and uncertainty of associations.

For modeling purposes, the continuous lung involvement percentage was converted into a binary outcome using a threshold optimized based on the balance between F1 score, precision, and recall. Predictive modeling was performed with logistic regression (LR), support vector machine (SVM), random forest (RF), gradient boosting (GB), decision tree (DT), naïve Bayes (NB), k-nearest neighbors (KNN), and multilayer perceptron (MLP) models. All models were trained and evaluated using grouped 5-fold cross-validation, with grouping defined by scanning day to prevent information leakage across folds. Model interpretation was conducted using SHapley Additive exPlanations (SHAP). Performance metrics included accuracy (ACC), F1 score, area under the receiver operating characteristic curve (ROC AUC), and area under the precision–recall curve (PR AUC). Differences before and after adding daily case counts were tested by the DeLong and Z tests. A  $p$ -value  $< 0.05$  was considered statistically significant.

## Results

### Participants

After applying inclusion and exclusion criteria (see flow diagram, Figure 1), 10,397 unique patients were included in the analysis. The

cohort comprised 4,630 males and 5,767 females, with mean age  $48.96 \pm 18.91$  years.

## Characteristics and lung involvement percentage

Characteristics are summarized in Table 1. Kernel density estimation plots of lung involvement percentage stratified by sex (Figure 2A) show that most patients had minimal lung involvement, with a small proportion exhibiting higher percentages, indicating a right-skewed distribution.

Age-stratified positivity proportions based on radiologist reports (Figure 2B) showed a progressively higher proportion of positive cases in older age groups. Date-stratified radiologist report-based positivity proportions (Figure 2C) exhibited temporal patterns characterized by higher positivity during periods with increased daily case counts. Similarly, the date-stratified mean lung involvement percentage (Figure 2D) showed a comparable temporal trend, with higher mean involvement observed on days with higher case counts.

## GEE Logistic regression analysis

GEE Logistic regression showed that age was strongly associated with lung involvement (OR 1.0813 per year increase, 95% CI 1.0703–1.0925,  $p < 0.0001$ ), while daily case counts also showed a modest but significant association (OR 1.0033 per 1-case increase, 95% CI 1.0018–1.0047,  $p < 0.0001$ ). Sex exhibited a minimal association, with females exhibiting lower odds of lung involvement compared with males (OR 0.8098, 95% CI 0.6983–0.9391,  $p = 0.0053$ ).

## Predictive performance of machine learning models

For predictive modeling, age, sex and the inclusion or exclusion of daily case counts were used as independent variables, while binary lung involvement percentage served as the dependent variable. The binary classification threshold (0.4%) was determined based on the intersection of F1 score, precision, and recall (Figure 3). ACC, F1 score, ROC AUC, and PR AUC are summarized in Table 2.

Across the eight evaluated models, logistic regression demonstrated consistently strong and stable performance both before and after the inclusion of daily case counts. In contrast, the impact of adding daily case counts varied across machine learning models. While modest improvements in discrimination were observed for certain models (e.g., SVM and KNN), several tree-based and neural network models exhibited a decline in performance under grouped five-fold cross-validation. DeLong tests identified statistically significant changes in ROC AUC for SVM, RF, DT, KNN and MLP, whereas PR AUC comparisons using Z-tests showed significant differences for most models.

After incorporating daily case counts as an independent variable, ROC and PR curves of the models are shown in Figure 4. Confusion matrices summarizing the classification results of the eight machine learning models are presented in Figure 5, with darker colors

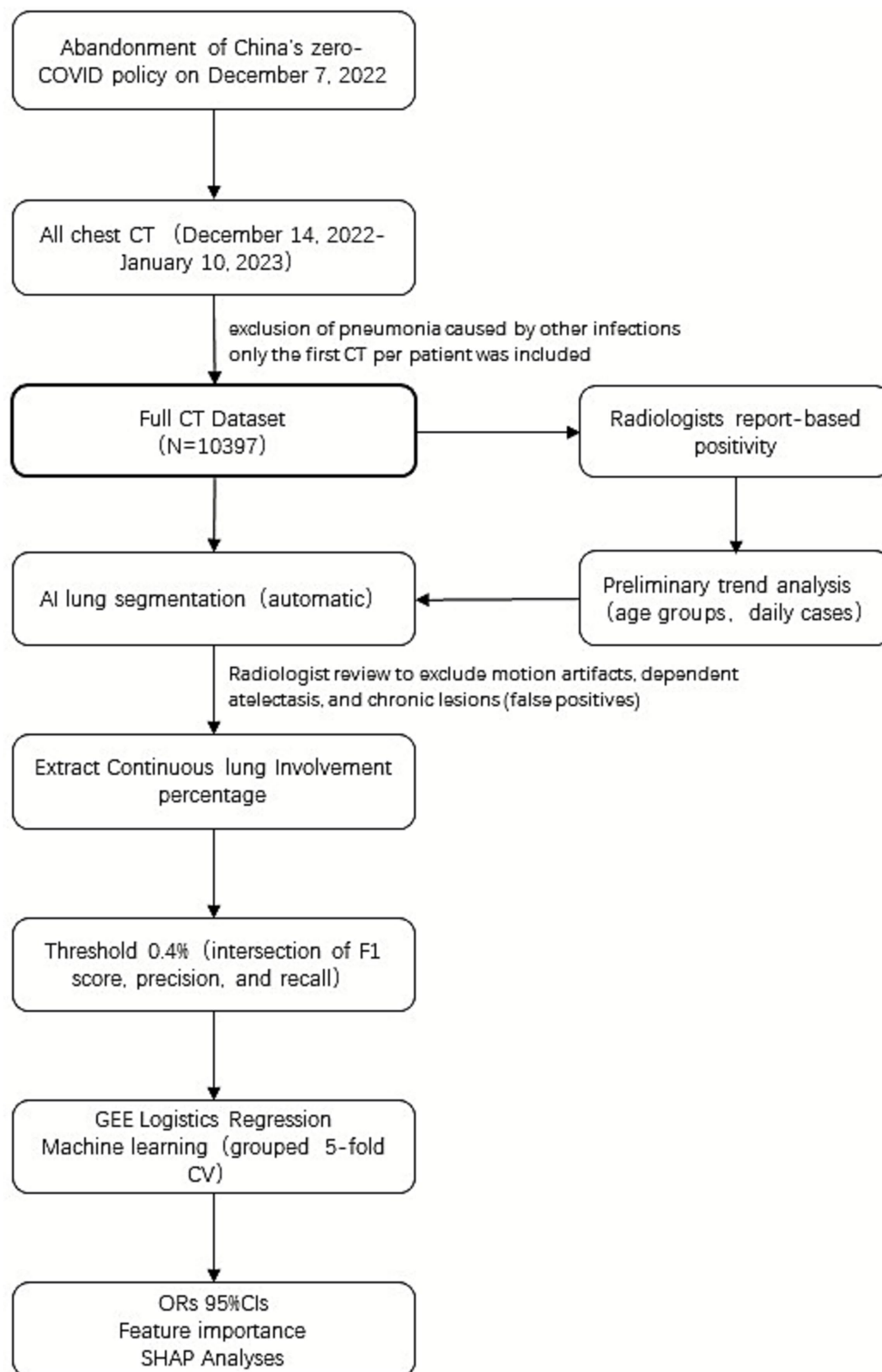


FIGURE 1  
Flow diagram of data selection and analyses.

indicating higher counts. Density plots of predicted probabilities for non-involvement (Class 0, green) and lung involvement (Class 1, red) across the eight classifiers are shown in Figure 6, illustrating the distributional separation between the two classes. Predicted positivity

rates stratified by age group are shown in Figure 7. Observed positivity rates (red bars) increased progressively with age, particularly among individuals aged over 60 years, where predictions from all models showed close agreement with observed values. In

TABLE 1 Characteristics of all patients.

Variable	Total (N = 10,397)	Survivors (N = 10,381)	Deaths (N = 16)	Deaths (%)	p-value
Age (years, mean ± SD)	48.9 ± 18.9	48.9 ± 18.9	79.0 ± 10.0	0.15%	<0.0001
Age group 0–20, n (%)	283 (2.7%)	283 (2.7%)	0 (0.0%)	0.0%	
Age group 21–30, n (%)	1,521 (14.6%)	1,521 (14.6%)	0 (0%)	0.0%	
Age group 31–40, n (%)	2,474 (23.8%)	2,474 (23.8%)	0 (0%)	0.0%	
Age group 41–50, n (%)	1,584 (15.2%)	1,584 (15.2%)	0 (0%)	0.0%	
Age group 51–60, n (%)	1,629 (15.7%)	1,628 (15.7%)	1 (6.25%)	0.06%	
Age group 61–70, n (%)	1,213 (11.7%)	1,212 (11.7%)	1 (6.25%)	0.06%	
Age group 71–80, n (%)	978 (9.4%)	973 (9.37%)	5 (32.25%)	0.51%	
Age group 80 + n (%)	715 (6.88%)	706 (6.8%)	9 (56.25%)	1.26%	
Female, n (%)	5,764 (55.5%)	5,759 (55.5%)	5 (31.2%)	0.09%	0.0893
Male, n (%)	4,627 (44.5%)	4,616 (44.5%)	11 (68.8%)	0.24%	
Lung involvement percentage (%), mean ± SD	2.2 ± 7.1	2.1 ± 6.6	69.1 ± 14.5		<0.0001
Lung involvement percentage = 0, n (%)	6,175 (59.4%)	6,175 (59.4%)	0 (0.0%)		<0.0001

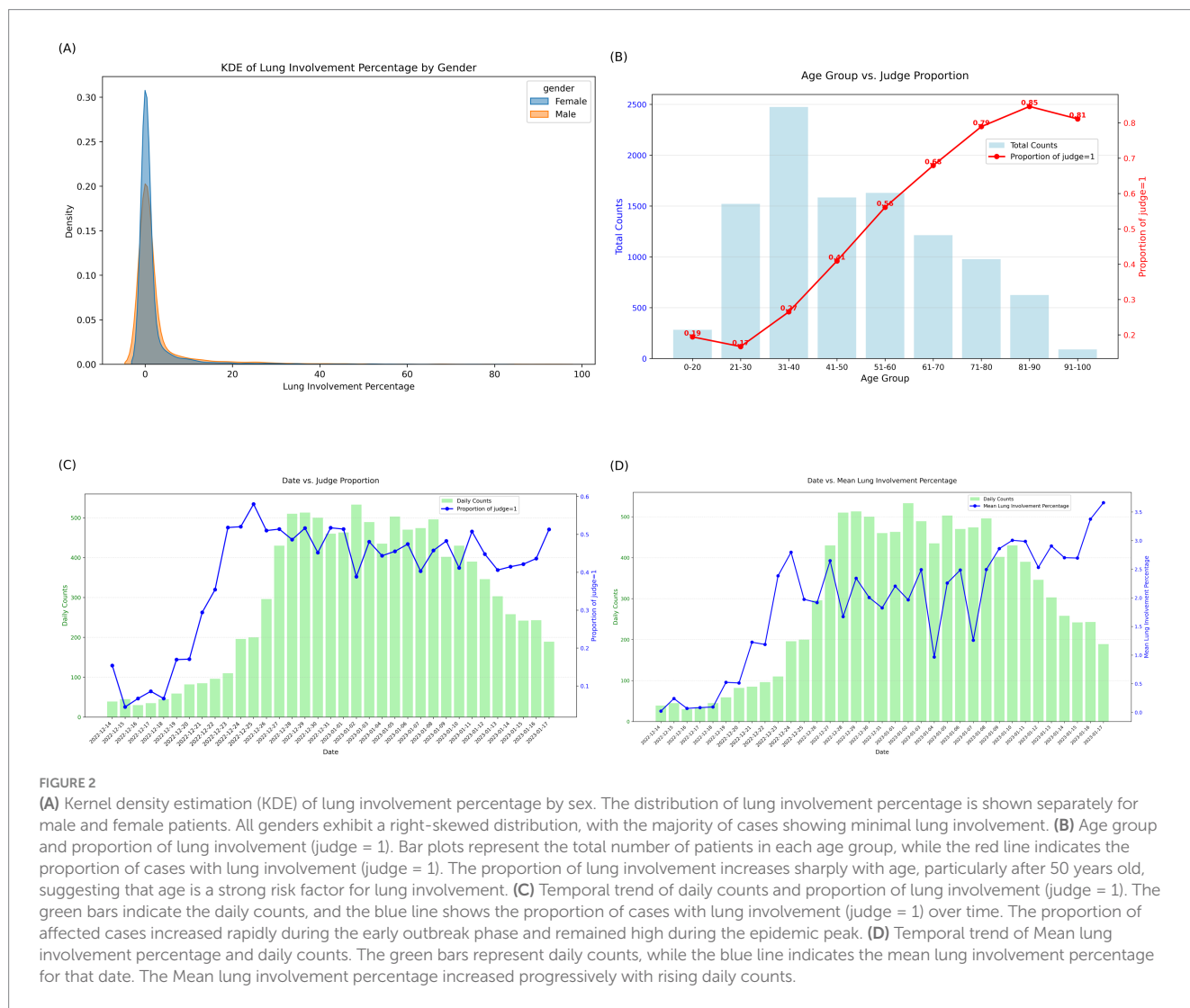
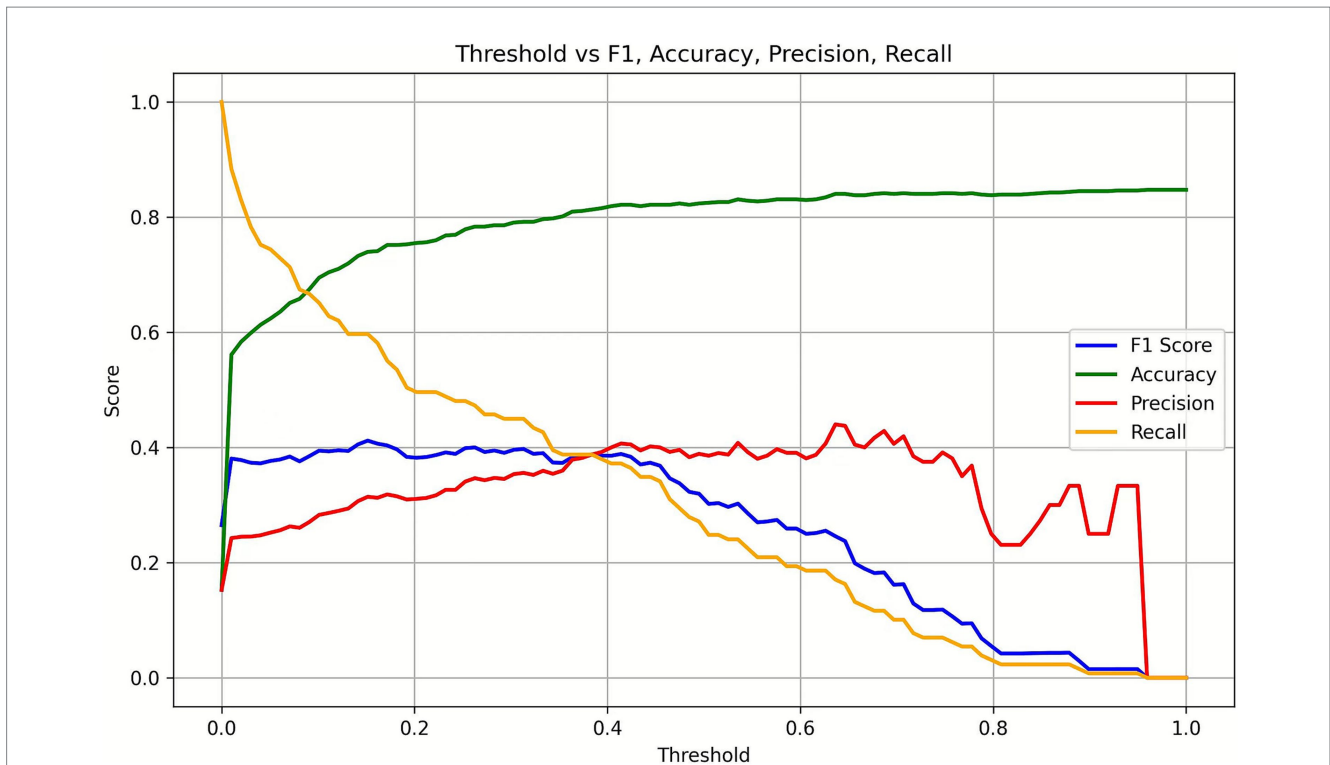


FIGURE 2

(A) Kernel density estimation (KDE) of lung involvement percentage by sex. The distribution of lung involvement percentage is shown separately for male and female patients. All genders exhibit a right-skewed distribution, with the majority of cases showing minimal lung involvement. (B) Age group and proportion of lung involvement (judge = 1). Bar plots represent the total number of patients in each age group, while the red line indicates the proportion of cases with lung involvement (judge = 1). The proportion of lung involvement increases sharply with age, particularly after 50 years old, suggesting that age is a strong risk factor for lung involvement. (C) Temporal trend of daily counts and proportion of lung involvement (judge = 1). The green bars indicate the daily counts, and the blue line shows the proportion of cases with lung involvement (judge = 1) over time. The proportion of affected cases increased rapidly during the early outbreak phase and remained high during the epidemic peak. (D) Temporal trend of Mean lung involvement percentage and daily counts. The green bars represent daily counts, while the blue line indicates the mean lung involvement percentage for that date. The Mean lung involvement percentage increased progressively with rising daily counts.

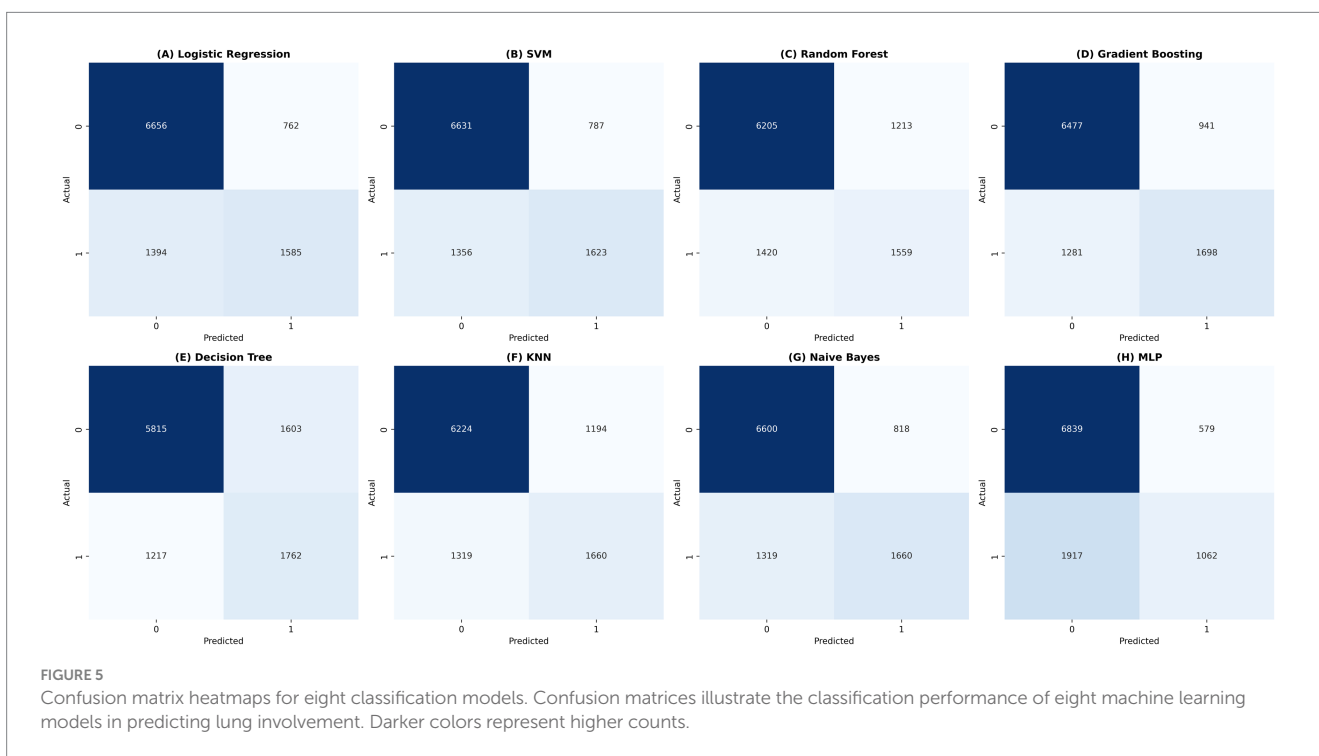
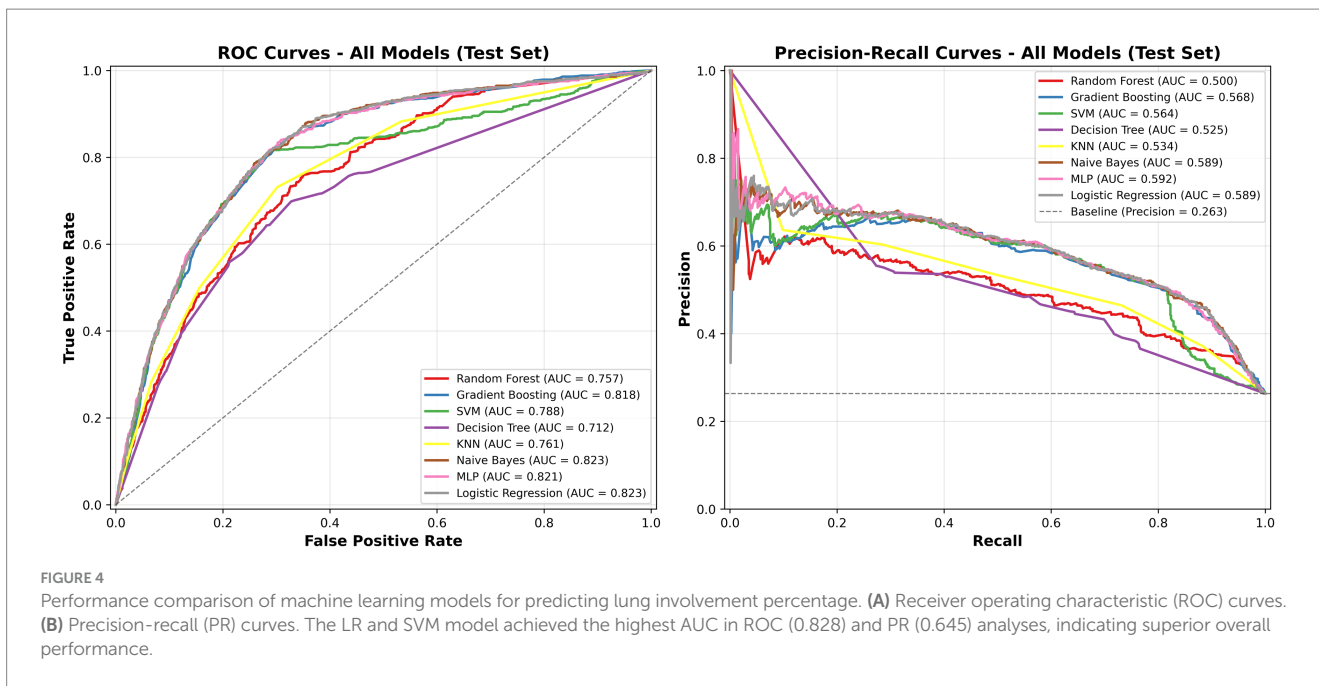


**FIGURE 3** Threshold optimization based on F1 score, precision, and recall. F1 score, precision, and recall were plotted across different thresholds. The three metrics intersected around 0.4%, which was selected as the threshold for classification.

**TABLE 2** Comparison of model performance before and after incorporating daily case counts.

Model	Condition	ACC	F1	ROC AUC	DeLong <i>p</i> -value	PR AUC	Z-test <i>p</i> -value
LR	Before	0.7894	0.5783	0.8228	Reference	0.5903	Reference
LR	After	0.7873	0.5865	0.8229	0.7640	0.5894	0.5020
SVM	Before	0.7860	0.5465	0.7868	Reference	0.5723	Reference
SVM	After	0.7876	0.5971	0.8208	0.0000	0.5903	0.0520
RF	Before	0.7863	0.5545	0.8145	Reference	0.5725	Reference
RF	After	0.7502	0.5061	0.7572	0.0000	0.4994	0.0000
GB	Before	0.7851	0.5449	0.8199	Reference	0.5769	Reference
GB	After	0.7811	0.5613	0.8180	0.3780	0.5684	0.2520
DT	Before	0.7832	0.5555	0.8127	Reference	0.5725	Reference
DT	After	0.7260	0.5127	0.7099	0.0000	0.5230	0.0000
KNN	Before	0.7486	0.4704	0.7057	Reference	0.4851	Reference
KNN	After	0.7570	0.5207	0.7593	0.0000	0.5212	0.0440
NB	Before	0.7870	0.5817	0.8230	Reference	0.5907	Reference
NB	After	0.7879	0.5917	0.8228	0.5900	0.5894	0.0400
MLP	Before	0.7860	0.5967	0.8228	Reference	0.5895	Reference
MLP	After	0.7792	0.5587	0.8147	0.0000	0.5714	0.0000

Performance metrics of eight machine learning models—including LR, SVM, RF, GB, DT, KNN, NB, and MLP—before and after inclusion of daily case counts as an additional variable. Model performance was evaluated using ACC, F1 score, ROC AUC, and PR AUC. Statistical differences were assessed using the DeLong test (for ROC AUC) and Z-test (for PR AUC). 1. “Before” indicates model performance using age and gender as predictors. 2. “After” indicates model performance after incorporating daily case counts. 3. “Reference” indicates the baseline value for comparison in statistical tests.

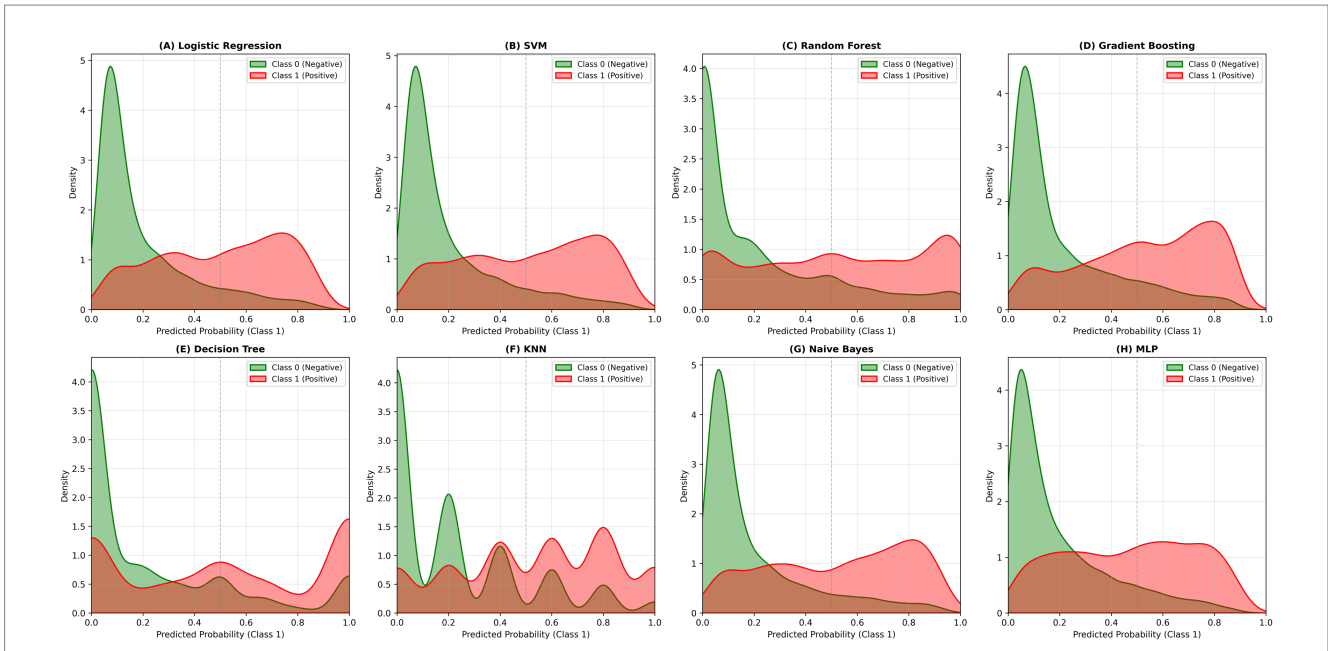


younger age groups (0–40 years), greater variability in predicted positivity rates was observed across models, with lower predicted probabilities in some classifiers. In older age groups ( $\geq 50$  years), predicted and observed positivity rates were generally concordant across models.

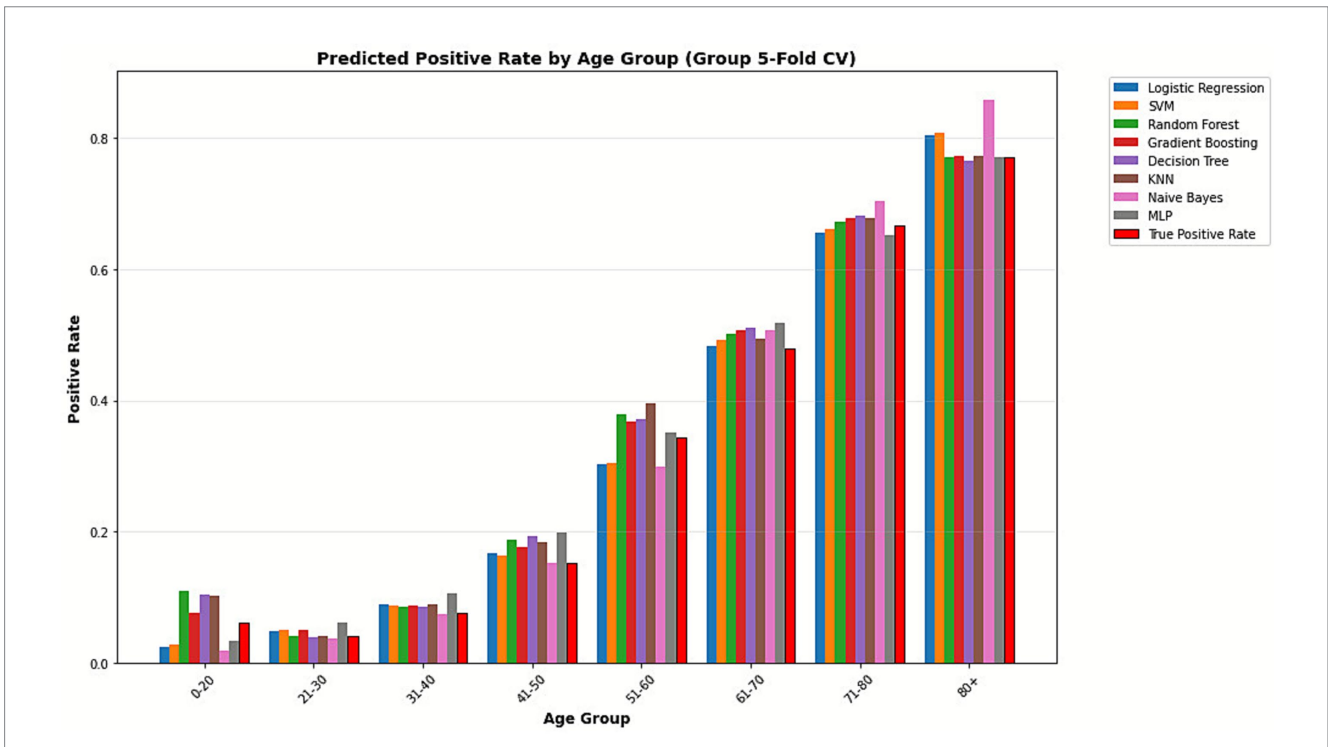
Using gradient boosting, the relative importance of three factors for predicting lung involvement percentage was quantified as follows: age (0.8829), daily case counts (0.1125), and sex (0.0046) (see Figure 8).

### SHAP analysis of gradient boosting model

SHAP analysis was used to examine both global and local feature contributions in the gradient boosting model (Figures 9, 10). At the global level, age showed the largest SHAP contributions, followed by daily case counts, whereas sex exhibited comparatively smaller contributions. Positive SHAP values were associated with higher age and higher daily case counts in the model-predicted lung



**FIGURE 6** Predicted probability density distributions of lung involvement percentage by different machine learning models. Density plots show the predicted probability distributions for non-involvement (class 0, green) and lung involvement (class 1, red) across eight classifiers: Logistic Regression, SVM, Random Forest, Gradient Boosting, Decision Tree, KNN, Naive Bayes, and MLP. Models such as gradient boosting and MLP demonstrate better class separation, indicating higher discriminative performance.

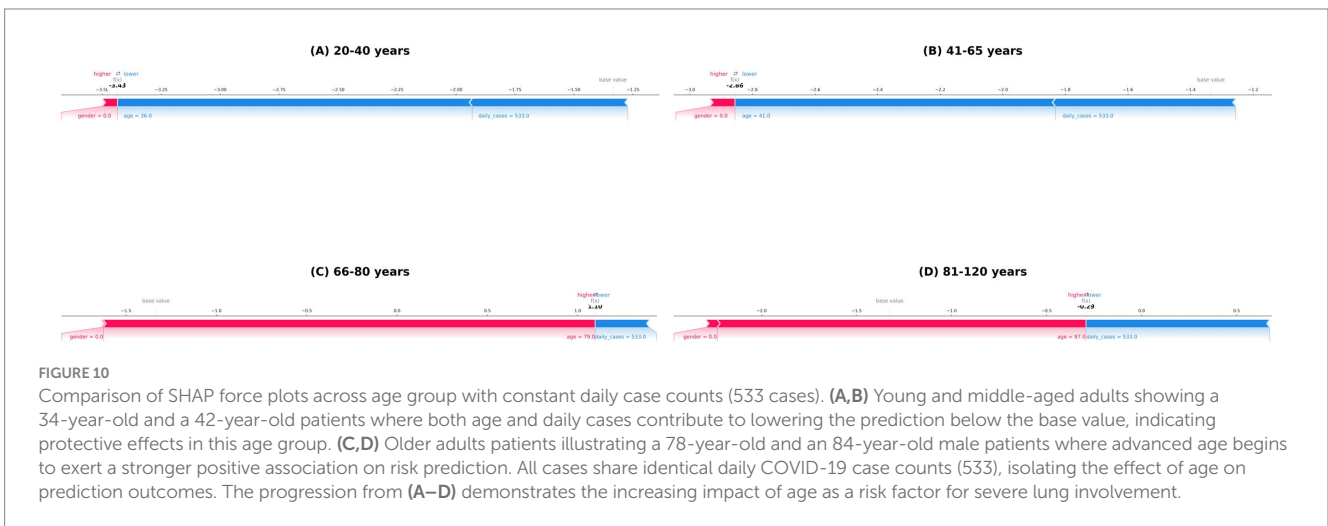
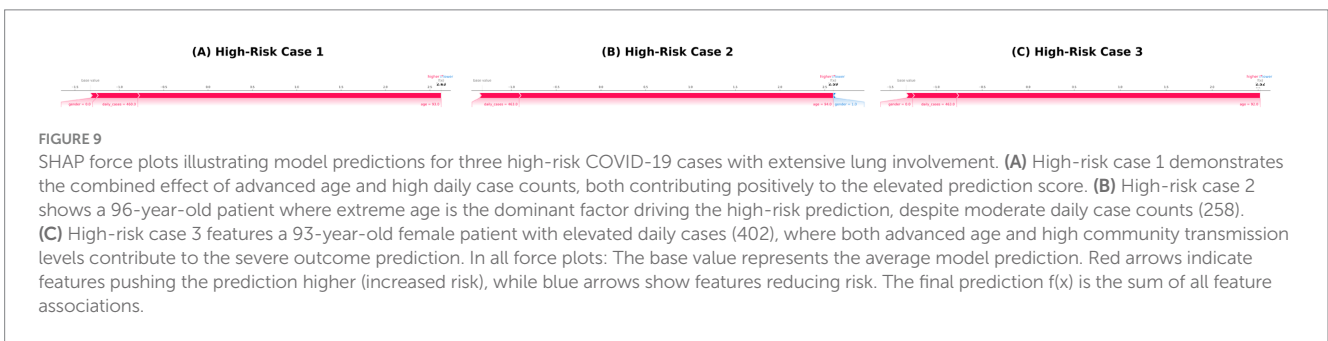
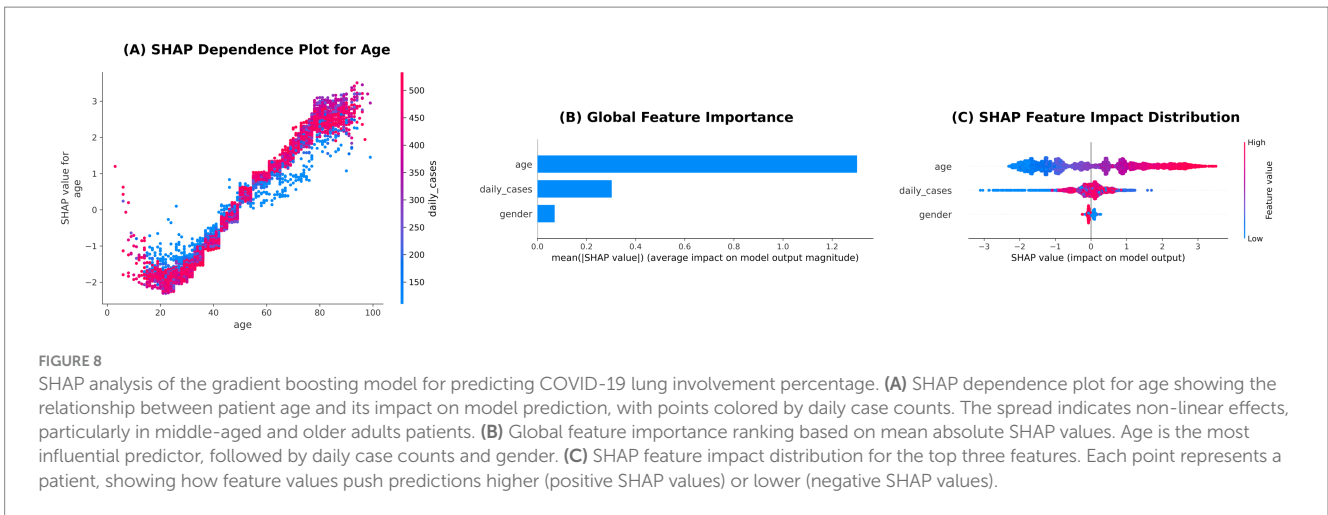


**FIGURE 7** Predicted positive rate by age group across machine learning models. Bar plots show the predicted positive rates (lung involvement) across different age groups. True positive rate (red bars) increases progressively with age.

involvement percentage, while sex showed predominantly small SHAP values.

At the individual level, SHAP force plots indicated that predictions of higher lung involvement percentage were mainly associated with

advanced age and elevated daily case counts, with minimal contributions from sex. Age-stratified SHAP plots demonstrated an increasing magnitude of SHAP contributions with advancing age, with the oldest age group ( $\geq 81$  years) showing the highest age-related contributions.



## Discussion

### Overview

In this study, we analyzed chest CT scans of COVID-19 patients from Suzhou, China, over a five-week period following the cessation of the dynamic zero-COVID policy, using artificial intelligence-based pneumonia analysis software. We examined associations between lung involvement and age, sex, daily case counts, and applied machine learning models with SHAP explanations. While numerous studies

have investigated the effects of age and sex on COVID-19 outcomes, few have explored the impact of outbreaks, and there is a lack of research on the relationship between daily case counts and lung involvement percentage.

### CT findings

Among our Omicron cases, CT findings included ground-glass opacities, consolidation, poorly defined margins with bronchial

aeration signs, with or without pleural effusion. Previous studies have reported highly variable incidence rates of these features (12). Lesions were commonly distributed bilaterally, peripherally, and in the posterior lung regions. These findings are nonspecific and overlap with other infections, limiting the diagnostic specificity of chest CT for COVID-19 (13, 14). Some literature also indicates that early CT imaging has limited value in predicting Omicron disease severity and outcomes (13). Recent studies have further suggested that overall lung involvement in Omicron infections is relatively mild (15, 16).

In our cohort, using data from patients' first CT scans, lung lesions similarly exhibited bilateral, peripheral, and posterior distribution. Evidence suggests that COVID-19 is fundamentally an endothelial disease (17–19). From this perspective, the formation of intravascular microthrombi causing obstruction in pulmonary capillaries can manifest as ischemic necrosis in the lung interstitium (20), which aligns with radiological findings and may explain the formation of pulmonary cavities, pneumothorax, and the bilateral, peripheral, posterior distribution of lesions due to gravitational effects on microthrombi. Clinically, this mechanism may also underlie manifestations such as skin discoloration, renal impairment, stroke, and myocardial injury. Some studies indicate that chest CT can reflect not only pulmonary damage but also pathological changes associated with myocardial injury (21). Nevertheless, these interpretations remain indirect, and the extent to which quantitative lung involvement on CT represents underlying microvascular pathology warrants further investigation.

## Significant impact of age on lung involvement percentage

Our analyses consistently indicate that age is the strongest factor associated with lung involvement. GEE Logistic regression showed that older age was strongly associated with higher lung involvement percentage (OR 1.0813, 95% CI 1.0703–1.0925,  $p < 0.0001$ ). Gradient boosting analysis confirmed this result, with age exhibiting the highest feature importance (0.8829). This finding is consistent with extensive prior literature, indicating that older patients are more likely to develop severe pulmonary damage following SARS-CoV-2 infection (22). Zhou et al. (23) also identified age as a major risk factor for severe COVID-19, with older individuals exhibiting higher risks of serious complications and mortality. Our results further corroborate these observations, highlighting that even during outbreaks without strict containment measures, age remains a critical association of lung involvement percentage.

The mechanisms underlying the impact of age on lung involvement percentage may include immunosenescence, accumulation of comorbidities, and decreased host response to viral infection. Previous studies have shown that older adults patients are more prone to cytokine storms, which exacerbate pulmonary inflammatory responses (24). The findings of our study support these mechanisms and emphasize that, particularly during uncontrolled epidemic surges, age is a key factor determining the extent of lung involvement percentage.

## Relationship between daily case counts and lung involvement percentage

Daily case counts emerged as the second most important factor. GEE Logistic regression indicated a modest but statistically significant association with lung involvement percentage (OR 1.0033, 95% CI 1.0018–1.0047,  $p < 0.0001$ ). Gradient boosting showed that daily case counts had a feature importance of 0.1125. These findings suggest that periods of higher epidemic intensity are associated with increased lung involvement on CT. An increase in daily case counts reflects widespread viral transmission within the community. At the population level, intensified transmission has been associated with higher viral burden and increased clinical heterogeneity among infected individuals. Previous studies have reported that the extent of lung involvement in Omicron infections varies with viral load as reflected by Ct values (25), whereas lung involvement was generally minimal under strict lockdown conditions. As the intensity of viral transmission increases, the composition of infected individuals may shift toward older or more vulnerable populations, who are more susceptible to lung involvement. This epidemiological mechanism may partly explain the observed association between daily case counts and lung involvement percentage.

Another consideration is that surges in daily cases are often accompanied by increased strain on healthcare resources, potentially delaying optimal treatment and worsening patient outcomes. Wang et al. (26) demonstrated that higher hospitalization rates during outbreak periods were significantly associated with increased proportions of severe cases, likely due to healthcare systems operating under high capacity constraints. Other studies also indicate that mortality significantly rises for patients admitted during “surge” periods (27, 28). However, in our study, the Suzhou Hospital of Integrated Traditional Chinese and Western Medicine operated with sufficient preparedness and high efficiency; we did not observe systematic delays or exclusion of patients from imaging, including those with mild disease. Therefore, the observed association between daily case counts and lung involvement is unlikely to be solely attributable to imaging delays related to healthcare system overload.

Our results indicate that although the feature importance of daily case counts is lower than that of age, it remained a meaningful factor in predicting lung involvement.

## Minimal association of sex on lung involvement percentage

Sex showed a minimal association with lung involvement percentage in our analyses. GEE Logistic regression indicated a slight protective association for females patients (OR 0.8098, 95% CI 0.6983–0.9391,  $p = 0.0053$ ), and gradient boosting analysis yielded a feature importance of 0.0046. Indicating a minimal association on lung involvement in our study. This finding differs from some previous reports, which suggest that male patients tend to exhibit more severe symptoms and have higher rates of critical illness and mortality following SARS-CoV-2 infection (3). In our study, however, the influence of sex on lung involvement percentage was limited. This discrepancy may be related to cohort characteristics or study design.

Under uncontrolled epidemic conditions, large-scale case surges may attenuate or obscure sex-related differences in pulmonary involvement across the overall patient population.

## SHAP-based interpretations

SHAP analysis provided interpretable insights into the gradient boosting model predictions, reaffirming that age was the most influential predictor of extensive lung involvement percentage, followed by daily case counts. This indicates that, according to the model, patients admitted during periods of high daily case counts were predicted to have greater lung involvement. Importantly, SHAP values reflect predictive influence within the model and do not establish formal statistical associations or causal relationships. The observed patterns likely correspond to increased viral exposure during epidemic surges rather than constraints on healthcare resources.

Furthermore, the negative SHAP values associated with female sex indicate that, within the model predictions, being female slightly reduced the predicted probability of extensive lung involvement compared with male sex. Importantly, SHAP values represent the influence of features on model outputs and do not constitute formal statistical associations or causal effects. This approach facilitates transparent interpretation of machine learning outputs in a clinically meaningful manner.

## Limitations and future directions

Although this study analyzed the primary associations of lung involvement percentage in COVID-19 using multiple machine learning models and quantified feature importance via gradient boosting, several limitations remain. Daily case counts does not directly correspond to the total number of newly diagnosed COVID-19 cases; therefore, it serves only as a proxy indicator of epidemic intensity. We only considered three independent variables—age, sex, and daily case counts—without including other potentially influential factors such as comorbidities, vaccination status, length of hospitalization, and treatment regimens. Future studies incorporating a broader range of variables could improve the predictive accuracy of the models.

Furthermore, while machine learning models such as random forests and gradient boosting perform well in capturing nonlinear relationships, they may fail to account for more complex interactions. Future research could explore more sophisticated approaches, such as deep learning models, to further enhance predictive performance.

## Conclusion

This study analyzed the associations of lung involvement percentage in COVID-19 under non-lockdown conditions and found that age was the most important predictor of lung involvement, followed by daily case counts, whereas sex had a minimal effect. Lung involvement percentage increased with both patient age and daily case numbers. Our findings provide novel insights into the clinical management of COVID-19 patients in uncontrolled epidemic settings and offer data-driven support for responding to similar outbreaks in the future.

## Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving humans were approved by Ethics Committee of Suzhou Hospital of Integrated Traditional Chinese and Western Medicine. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

## Author contributions

YM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. LY: Data curation, Resources, Writing – original draft. JP: Data curation, Resources, Writing – original draft. DS: Data curation, Resources, Writing – original draft. QW: Data curation, Resources, Writing – original draft. BS: Data curation, Resources, Writing – original draft. YS: Data curation, Resources, Writing – original draft. XZ: Data curation, Resources, Writing – original draft. FC: Data curation, Resources, Writing – original draft. JS: Data curation, Resources, Writing – original draft. QY: Writing – original draft. SQ: Writing – original draft. RR: Writing – original draft. XL: Writing – original draft. JX: Writing – original draft. JZ: Writing – original draft. DZ: Writing – original draft. QZ: Writing – review & editing. YZ: Writing – review & editing. BZ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2025.1728282/full#supplementary-material>

## References

- Docherty, AB, Harrison, EM, Green, CA, Hardwick, HE, Pius, R, Norman, L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO clinical characterisation protocol: prospective observational cohort study. *BMJ*. (2020) 369:m1985. doi: 10.1136/bmj.m1985
- Statsenko, Y, Al Zahmi, F, Habuza, T, Gorkom, KN, and Zaki, N. Prediction of COVID-19 severity using laboratory findings on admission: informative values, thresholds, ML model performance. *BMJ Open*. (2021) 11:e044500. doi: 10.1136/bmjopen-2020-044500
- Peckham, H, de Grujter, NM, Raine, C, Radziszewska, A, Ciurtin, C, Wedderburn, LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nat Commun*. (2020) 11:6317. doi: 10.1038/s41467-020-19741-6
- Klein, SL. Sex influences immune responses to viruses, and efficacy of prophylaxis and treatments for viral diseases. *BioEssays*. (2012) 34:1050–9. doi: 10.1002/bies.201200099
- Williamson, EJ, Walker, AJ, Bhaskaran, K, Bacon, S, Bates, C, Morton, CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. (2020) 584:430–6. doi: 10.1038/s41586-020-2521-4
- Pontecorvi, G, Bellenghi, M, Ortona, E, and Carè, A. microRNAs as new possible actors in gender disparities of Covid-19 pandemic. *Acta Physiol (Oxf)*. (2020) 230:e13538. doi: 10.1111/apha.13538
- Pagano, MT, Peruzzo, D, Ruggieri, A, Ortona, E, and Gagliardi, MC. Vitamin D and sex differences in COVID-19. *Front Endocrinol*. (2020) 11:567824. doi: 10.3389/fendo.2020.567824
- Wray, S, and Arrowsmith, S. The physiological mechanisms of the sex-based difference in outcomes of COVID19 infection. *Front Physiol*. (2021) 12:627260. doi: 10.3389/fphys.2021.627260
- Jin, JM, Bai, P, He, W, Wu, F, Liu, XF, Han, DM, et al. Gender differences in patients with COVID-19: focus on severity and mortality. *Front Public Health*. (2020) 8:152. doi: 10.3389/fpubh.2020.00152
- Chidambaram, V, Tun, NL, Haque, WZ, Majella, MG, Sivakumar, RK, Kumar, A, et al. Factors associated with disease severity and mortality among patients with COVID-19: a systematic review and meta-analysis. *PLoS One*. (2020) 15:e0241541. doi: 10.1371/journal.pone.0241541
- Booth, A, Reed, AB, Ponzo, S, Yassaee, A, Aral, M, Plans, D, et al. Population risk factors for severe disease and mortality in COVID-19: a global systematic review and meta-analysis. *PLoS One*. (2021) 16:e0247461. doi: 10.1371/journal.pone.0247461
- Fatima, N, Khokhar, SA, and Farooq Ur Rehman, RM. Correlation between oxygen saturation of patient and severity index of Covid 19 pneumonia on CT. *J Pak Med Assoc*. (2023) 73:60–3. doi: 10.47391/JPMA.5586
- Alsharif, W, and Qurashi, A. Effectiveness of COVID-19 diagnosis and management tools: a review. *Radiography*. (2021) 27:682–7. doi: 10.1016/j.radi.2020.09.010
- Raptis, CA, Hammer, MM, Short, RG, Shah, A, Bhalla, S, Bierhals, AJ, et al. Chest CT and coronavirus disease (COVID-19): a critical review of the literature to date. *AJR Am J Roentgenol*. (2020) 215:839–42. doi: 10.2214/AJR.20.23202
- Tsakok, MT, Watson, RA, Saujani, SJ, Kong, M, Xie, C, Peschl, H, et al. Reduction in chest CT severity and improved hospital outcomes in SARS-CoV-2 omicron compared with Delta variant infection. *Radiology*. (2023) 306:261–9. doi: 10.1148/radiol.220553
- Han, X, Chen, J, Chen, L, Jia, X, Fan, Y, Zheng, Y, et al. Comparative analysis of clinical and CT findings in patients with SARS-CoV-2 original strain, Delta and omicron variants. *Biomedicine*. (2023) 11:903. doi: 10.3390/biomed11030901
- Libby, P, and Lüscher, T. COVID-19 is, in the end, an endothelial disease. *Eur Heart J*. (2020) 41:3038–44. doi: 10.1093/eurheartj/ehaa623
- Bonaventura, A, Vecchiè, A, Dagna, L, Martinod, K, Dixon, DL, van Tassel, BW, et al. Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. *Nat Rev Immunol*. (2021) 21:319–29. doi: 10.1038/s41577-021-00536-9
- Portier, I, Campbell, RA, and Denorme, F. Mechanisms of immunothrombosis in COVID-19. *Curr Opin Hematol*. (2021) 28:445–53. doi: 10.1097/MOH.0000000000000666
- Conway, EM, Mackman, N, Warren, RQ, Wolberg, AS, Mosnier, LO, Campbell, RA, et al. Understanding COVID-19-associated coagulopathy. *Nat Rev Immunol*. (2022) 22:639–49. doi: 10.1038/s41577-022-00762-9
- Zhong, Y, Sun, Z, Xu, P, Bai, Y, Zhang, Z, and Wang, G. The value of non-contrast chest CT in the prediction of myocardial injury in patients with the COVID-19 omicron variant. *Sci Rep*. (2023) 13:10321. doi: 10.1038/s41598-023-37335-2
- Petrilli, CM, Jones, SA, Yang, J, Rajagopalan, H, O'Donnell, L, Chernyak, Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in new York City: prospective cohort study. *BMJ*. (2020) 369:m1966. doi: 10.1136/bmj.m1966
- Zhou, F, Yu, T, Du, R, Fan, G, Liu, Y, Liu, Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet (London, England)*. (2020) 395:1054–62. doi: 10.1016/S0140-6736(20)30566-3
- Mehta, P, McAuley, DF, Brown, M, Sanchez, E, Tattersall, RS, and Manson, JJ. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet (London, England)*. (2020) 395:1033–4. doi: 10.1016/S0140-6736(20)30628-0
- Ying, WF, Chen, Q, Jiang, ZK, Hao, DG, Zhang, Y, and Han, Q. Chest computed tomography findings of the omicron variants of SARS-CoV-2 with different cycle threshold values. *World J Clin Cases*. (2023) 11:756–63. doi: 10.12998/wjcc.v11.i4.756
- Wang, D, Hu, B, Hu, C, Zhu, F, Liu, X, Zhang, J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA*. (2020) 323:1061–9. doi: 10.1001/jama.2020.1585
- Keene, AB, Admon, AJ, Brenner, SK, Gupta, S, Lazarous, D, Leaf, DE, et al. Association of Surge Conditions with mortality among critically ill patients with COVID-19. *J Intensive Care Med*. (2022) 37:500–9. doi: 10.1177/08850666211067509
- Kadri, SS, Sun, J, Lawandi, A, Strich, JR, Busch, LM, Keller, M, et al. Association between caseload surge and COVID-19 survival in 558 U.S. hospitals, march to august 2020. *Ann Intern Med*. (2021) 174:1240–51. doi: 10.7326/M21-1213