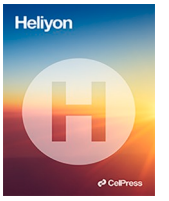
Heliyon 9 (2023) e17772

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Interpretable prediction of cardiopulmonary complications after non-small cell lung cancer surgery based on machine learning and SHapley additive exPlanations

Yihai Zhai [a], Xue Lin [b], Qiaolin Wei [c], Yuanjin Pu [a], Yonghui Pang [a,*]

[a] Guangxi Medical University Cancer Hospital, Department of Thoracic Surgery, Nanning, 530021, China
[b] The Second Affiliated Hospital of Guangxi Medical University, Department of Oncology, Nanning, 530000, China
[c] Guangxi Medical University Cancer Hospital, Department of Interventional Therapy, Nanning, 530021, China

ARTICLE INFO

Keywords:
Non–small cell lung cancer
Complications
Random forest
Interpretable model

ABSTRACT

Introduction: Lung cancer is a prevalent malignancy globally, with approximately 20% of patients developing cardiopulmonary complications after lobectomy. In order to prevent complications, an accurate and personalized method based on machine learning (ML) is required.

Methods: During the period of 2017–2021, a retrospective analysis was conducted on the medical records of patients who had undergone lobectomy for non-small cell lung cancer (NSCLC). We performed logical regression, decision tree (DT), random forest (RF), gradient boost DT, and eXtreme gradient boosting analyses to establish an ML model. The ten-fold cross-validation was used to evaluate the performance of multiple ML models based on various evaluation metrics, including accuracy, precision, recall, F1 score, and area under the receiver operating (AUC). Additionally, we also calculated the Kappa value of these model. Each model used grid search to optimize hyper-parameters and then used the interpretability method to provide explanations for the model's Decisions.

Results: The study included 718 eligible patients, among whom the incidence of postoperative cardiopulmonary complications was 20.89%. The RF model showed the best comprehensive performance among all models, and its ten-fold cross-validation accuracy, precision, recall, F1 score, and AUC were (OR and 95% confidence interval [CI]) 0.786 (0.738–0.834), 0.803 (0.735–0.872), 0.738 (0.678–0.797), 0.766 (0.714–0.818), 0.856 (0.815–0.898), respectively. The kappa value of the RF model was 0.696 (0.617–0.768). The SHAP method showed that gender, age, and intraoperative blood loss were closely associated with postoperative cardiopulmonary complications.

Conclusion: The application of ML methods for predicting postoperative cardiopulmonary complications based on clinical data of patients with NSCLC showed a good performance. The results indicate that ML combined with the SHAP individualized interpretation method has practical clinical value.

* Corresponding author.
E-mail address: pangyonghui003@163.com (Y. Pang).

https://doi.org/10.1016/j.heliyon.2023.e17772
Received 4 February 2023; Received in revised form 26 June 2023; Accepted 27 June 2023
Available online 3 July 2023

1. Introduction

Lung cancer exhibits the highest incidence and mortality rates among all malignant neoplasms at present [1]. Among them, non-small cell lung cancer (NSCLC) accounts for approximately 85% of the total lung cancers in clinically. Presently, the established surgical treatment for early-stage lung cancer is lobectomy and systematic lymph node dissection, the utilization of this treatment has been shown to greatly extend the lifespan of patients and enhance their quality of life [2,3]. However, pulmonary function damage due to surgery can occur, which includes pulmonary tissue traction, peripheral nerve, and tissue trauma during surgery, leading to postoperative cardiopulmonary complications. Some studies have shown that the incidence of pulmonary function damage is 8%–20% and includes symptoms such as pneumothorax, pneumonia, atelectasis, and pleural effusion [4–7]. This issue is worthy of further clinical studies. Moreover, the occurrence of postoperative cardiopulmonary complications increases the relevant hospitalization time by about 30%. Furthermore, it may also increase perioperative mortality in these patients by 7%–33% [8,9]. In a word, the early diagnosis of postoperative cardiopulmonary complications is of great significance for guiding the perioperative management of such patients.

As a means of predicting postoperative complications, previous studies used the POSSUM scoring system or the Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT) scoring system [10,11]. However, these systems cannot effectively predict lung cancer or related complications. Previous studies have attempted to establish a prediction model for postoperative cardiopulmonary complications [12,13]. Early prediction models for postoperative pulmonary complications have included established risk factors, including advanced age, chronic lung disease, American Society of Anesthesiologists (ASA) score, and smoking history [14]. More recent models have included the medical histories and laboratory test results of the patients. In a previous model, several factors, including asthma, nutrient lymphocyte ratio, male gender, and high body mass index (BMI), are considered important factors affecting postoperative complications [15].

However, these models do not work well in China [16], and these models are based on the method of binary logical regression (LR), which may lead to some research bias and is a major limitation [17–19]. Fortunately, machine learning (ML) technology has been increasingly prevalent in the medical industry owing to its ability to efficiently handle complex and time-consuming tasks, particularly in disease diagnosis, drug production, and medical data analysis. Presently, it has shown good ability in the diagnosis, management, and prediction of lung cancer. For instance, Lu and colleagues established a diagnosis model for high-risk smokers suffering from lung cancer using a large dataset and used a convolutional neural network and PLCOm2012 for further analysis [20]. The study findings revealed a diagnostic sensitivity of 0.749 for the convolutional neural network, which was slightly better than that of the risk prediction model, PLCOm2012, and could better predict lung cancer at the early stage. Wang et al. [21] used seven technologies, including LR, decision tree (DT), eXtreme gradient boosting (XGBoost), and support vector machines (SVM), to establish a prediction model for lung cancer viability in patients of different genders. In this case, the best one-year survival rate was shown by the XGBoost model (with 90.75% accuracy), followed by LR and SVM. Furthermore, the ML method also showed good performance in predicting various cancer complications. For example, Zeng et al. [22] used methods such as LR, DT C5.0, DT CART, SVM, and random forest (RF) under the supervision classification to predict the surgical complications after liver cancer resection and achieved an accuracy rate of 92.45%.

As per our understanding, there has been limited application of ML in the study of cardiopulmonary complications after lobectomy. Therefore, the primary aim of this investigation was to investigate the potential of utilizing ML-based methods in predicting postoperative cardiopulmonary complications following lobectomy. To achieve this, we developed risk prediction models utilizing clinical and laboratory data obtained from electronic medical records. We compared the predictive performance of various supervised learning methods, including LR, DT, RF, gradient boost DT (GBDT), and XGBoost. Additionally, we used the SHapley Additive exPlanations (SHAP) method to identify the significance of the variables in prediction model. Ultimately, we aimed to provide a tool for the early diagnosis of patients with cardiopulmonary complications after lobectomy, which could help clinicians make better-informed decisions and improve patient outcomes.

2. Methods

2.1. Study population

We retrospectively reviewed medical records of patients who underwent thoracoscopic lobectomy at a thoracic oncology department of a cancer research center from January 2017 to December 2021. Patient eligibility criteria were as follows: (1) Patients who had undergone their first operation. (2) NSCLC was diagnosed by pathological examination. (3) Age of the patients $\geq$18 years. (4) The physical fitness score of Eastern Cooperative Oncology Group was in the range of 0–2. Criteria for exclusion were as follows: (1) Conversion to thoracotomy; (2) presence of other cancers, (3) unplanned discharge.

2.2. Information collection indicators

Patient information was obtained from Hospital Information System, which included (1) Demographic characteristics such as gender, age, and BMI. (2) Medical history: hypertension, diabetes, coronary heart disease, or smoking. (3) Tumor information: maximum tumor size, tumor site, and stage. (4) Auxiliary examination results: maximum size of the left atrium, pulmonary function (FEV1%, FEV1/FVC); (5) Operative data: preoperative ASA score, anesthesia mode, duration of time, intraoperative blood loss, and histological type. (6) Laboratory examination: albumin, hemoglobin (HB), fasting blood glucose, platelet (PLT), alanine

aminotransferase (ALT), aspartate transaminase (AST), and neutrophils (N%). This is a retrospective study; therefore, the Ethics Review Committee agreed that an informed consent form is not needed.

2.3. Operation method

Surgeons performed a two-hole thoracoscopic lobectomy on the patients. The first hole was the observation hole, and the surgeons made a 1.0–2.0 cm incision at the 7th or 8th intercostal position of the midaxillary line. The second incision, known as the main operation hole, was made at the anterior axillary line's 4th or 5th intercostal position, with an incision size of 3.0–4.0 cm.

2.4. Data preprocessing

The binary independent variable was coded by assigning a value of 1 to the positive event and 0 to the negative event. We set one-hot variables for unordered multi-category variables. The continuous variables were standardized by mean normalization to prevent errors caused by large data differences. The main results were defined as cardiopulmonary complications occurring within two weeks after lobectomy, including pneumothorax, pleural effusion, atelectasis, pulmonary infection, acute respiratory failure, acute heart failure, arrhythmia, and bronchopleural fistula. The diagnostic standard for pulmonary complications was ARISCAT study-related definitions [23]. Arrhythmias include atrial fibrillation and ventricular fibrillation. Data with less than 20% missing content were completed by multiple imputations.

2.5. Sample equalization processing

Postoperative cardiopulmonary complications accounted for 20.89% of the total data of the study. Supervised learning algorithms that are optimized for overall classification accuracy may fail to learn important features from the minority classes, which can lead to imbalanced classification performance. Therefore, to ensure the efficiency of ML, the SMOTE Tomek Link algorithm with oversampling and undersampling was used [24]. This method can remove all noise points or boundary points in the sample while balancing the sample size.

2.6. Learning model

LR, DT, RF, GBDT, and XGBoost were used for model development.

Table 1
General information of patients.

Variable	No event	Event	P value	Variable	No event	Event	P value
Age (y)	58.0 (52.0–64.0)	62.0 (56.0–68.0)	<0.001	ASA			0.041
BMI	23.2 (20.8–25.0)	22.6 (20.7–24.5)	0.266	I/II	548 (76.32%)	139 (19.36)	
Duration of time (min)	177.0 (145.0–216.8)	186.5 (146.8–226.5)	0.188	III/IV	20 (2.79%)	11 (1.53%)	
Blood loss (ml)	80.0 (30.0–150.0)	100.0 (50.0–200.0)	0.001	Inhalation anesthesia			0.678
GLU (mmol/L)	4.6 (4.3–5.1)	4.5 (4.1–5.0)	0.080	Yes	430 (59.89%)	116 (16.16%)	
ALB (g/L)	39.0 (37.1–41.0)	38.5 (36.1–40.6)	0.062	No	138 (19.22%)	34 (4.74%)	
N (%)	57.7 (51.2–63.3)	58.7 (51.3–65.6)	0.350	Tumor site			0.375
WBC (10^12/L)	6.3 (5.4–7.5)	6.4 (5.2–7.9)	0.793	left upper lobe	107 (14.90%)	27 (3.76%)	
AST (U/L)	24.0 (20.0–28.0)	24.0 (20.0–30.0)	0.087	left lower lobe	86 (11.98%)	20 (2.79%)	
ALT (U/L)	17.0 (13.0–24.0)	17.0 (13.0–25.0)	0.555	right upper lobe	153 (21.30%)	43 (5.99%)	
HGB (g/L)	132.0 (121.0–142.0)	131.0 (120.8–142.3)	0.900	right middle lobe	45 (6.27%)	5 (0.70%)	
PLT (10^9/L)	247.0 (208.0–289.0)	260.5 (216.8–307.3)	0.023	right lower lobe	108 (15.04%)	30 (4.18%)	
Gender			0.205	bipulmonary lobe	69 (9.61%)	25 (3.48%)	
Male	285 (39.69%)	84 (11.70%)		Tumor size (cm)			0.027
Female	283 (39.42%)	66 (9.195)		<3	441 (61.42%)	103 (14.35%)	
Hypertension			0.092	3–5	96 (13.37%)	37 (5.15%)	
Yes	136 (18.94%)	46 (6.41%)		>5	31 (4.32%)	10 (1.39%)	
No	432 (60.17%)	104 (14.48%)		Tumor stage			0.357
Diabetes			0.835	I	381 (53.06%)	104 (14.48%)	
Yes	64 (8.91%)	16 (2.23%)		II	93 (12.95%)	31 (4.32%)	
No	504 (70.19%)	134 (18.66%)		IIIa	94 (13.09%)	15 (2.09%)	
Coronary disease			0.308	Histological typing			0.032
yes	20 (2.79%)	8 (1.11%)		Adenocarcinoma	498 (69.36%)	121 (16.85%)	
no	548 (76.32%)	142 (19.78%)		SCC	45 (6.27%)	21 (2.92%)	
FEV_1 (<80%)			0.039	Other	25 (3.48%)	8 (1.11%)	
normal	483 (67.27%)	117 (16.30%)		Smoking			0.027
abnormal	85 (11.84%)	33 (4.60%)		Yes	173 (24.09%)	60 (8.36%)	
FEV_1/FVC (<80%)			0.059	No	395 (55.01%)	90 (12.53%)	
normal	552 (76.88%)	131 (18.25%)					

Note: ASA:American Society of Anesthesiologists.

2.7. Statistical methods

The study data were added to Excel, and the statistical software used Python 3.10. Frequency was used to express the counting data, while the mean ± standard deviation was used to express the normally distributed measurement data. Non-normally distributed continuous variables were reported as median and interquartile range (IQR). Statistical comparisons between groups were performed using appropriate statistical tests such as *t*-test, chi-square test, or non-parametric tests based on the distribution of data, with a significance level of $\alpha = 0.05$.

The model results were evaluated by accuracy (percentage of correct prediction results in the total sample), precision (percentage of all predicted positive samples in the actual value of positive samples), recall (percentage of predicted positive samples in all actual values of positive samples), F1 score (harmonic mean of precsion and recall), and area under the receiver operating characteristic curve (AUC). All indicators ranged from 0 to 1, with higher values indicating superior classification performance. Kappa values were used to assess the performance of classification models. The data were divided into test sets and validation sets at a ratio of 7:3. At the same time, to improve the generalization ability of the model, the test set was calculated by the form of ten fold cross-validation, and the hyper-parameters of the model were adjusted by the Gridsearch method. Finally, the mean of the evaluation metrics obtained from the validation set were reported as the performance indicators of the model. The 95% confidence interval (CI) was also calculated to

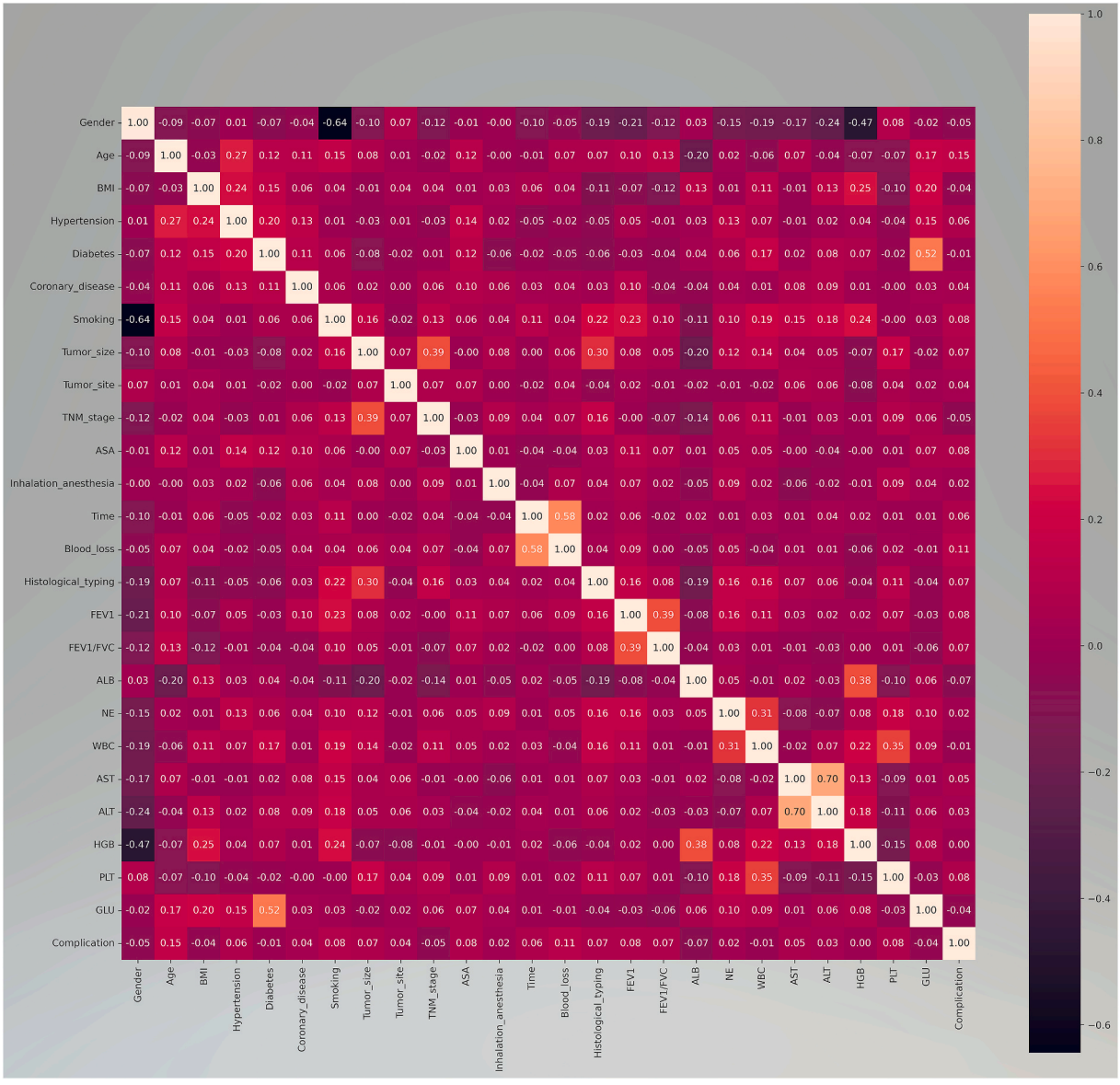


Fig. 1. The diagram of thermodynamic correlations between features.

estimate the variability of the results. The importance of variables was evaluated using the SHAP method. Each sample in the prediction process was assigned a predictive value by the model, and for each feature in the sample, a SHAP value was calculated [25].

3. Results

3.1. Patient characteristics

In accordance with the study's eligibility criteria, the data from 718 patients were included. Postoperative cardiopulmonary complications occurred in 150 patients (20.89%). Among them, 92 patients were of pneumothorax (61.33%), 18 of atelectasis (12.00%), 22 of pulmonary infection (14.67%), 5 patients were acute respiratory failure (3.33%), 1 of acute heart failure (0.67%), and 39 of arrhythmia (26.00%). Table 1 shows the basic information of these patients. The average hospital stay in the normal group was 8.79 ± 3.99 days, whereas that in the complication group was 11.86 ± 7.36 days ($t = -6.844$, $P < 0.001$). Fig. 1 presents a thermodynamic correlation diagram of all features, whereas Fig. 2 depicts a violin plot showing the multi-feature after normalized.

3.2. Prediction effect of multiple ML models

We optimized the hyper-parameters of all models by the grid search method. Table 2 displays the effectiveness of final ten-fold cross-validation model and its 95% CI. Each model has its own advantages in learning cardiopulmonary complications after lobectomy. Additional information regarding the ML methods used in this study, including hyper-parameter optimization and the results of such optimization, can be found in the appendix. The comprehensive performance indicators of each learning model were better than those of the LR method. Regarding the performance of the RF model, except for the recall rate, other performances were optimal. The AUC of RF is shown in Fig. 3. Simultaneously, we calculated the kappa value and 95% CI of the RF model by sampling 1000 random samples. Its performance on the test set were 0.696 (0.617–0.768).

3.3. Assessment of important variables

To explain the ML model, we selected two best-preforming models to explain these characteristics. The importance of each feature of the SHAP method is shown in Figs. 4 and 5, where the top 20 important variables are listed. The left halves of Figs. 4 and 5 represent the mean value of the important variables categorized according to their importance and whether they have protective or harmful

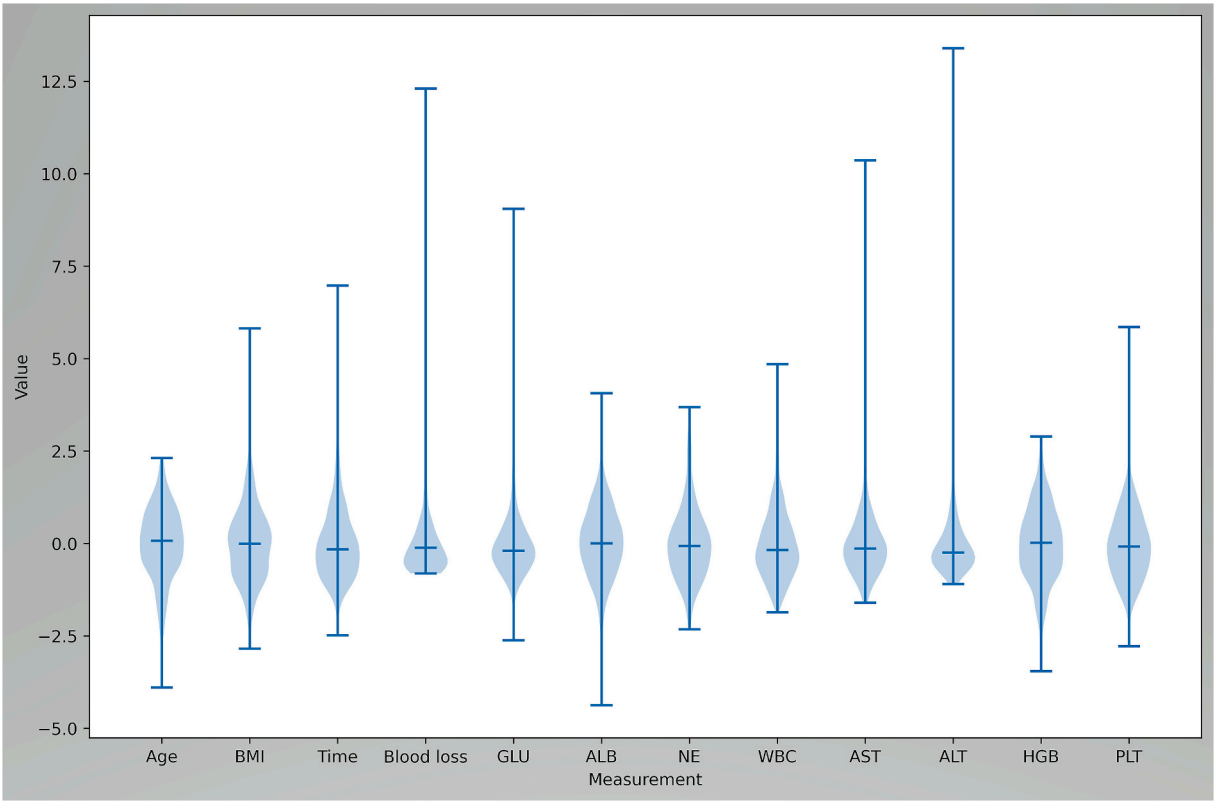


Fig. 2. Normalized multi-feature violin plot.

Table 2
The prediction of the model after ten-fold cross-validation.

	Accuracy (95% CI)	Precision (95% CI)	Recall rate (95% CI)	F1 score(95% CI)	AUC (95% CI)
LR	0.679 (0.609–0.749)	0.670 (0.588–0.751)	0.675 (0.609–0.741)	0.669 (0.603–0.734)	0.726 (0.666–0.786)
DT	0.696 (0.638–0.755)	0.663 (0.601–0.726)	0.756 (0.673–0.840)	0.702 (0.643–0.762)	0.720 (0.659–0.782)
RF	0.786 (0.738–0.834)	0.803 (0.735–0.872)	0.738 (0.678–0.797)	0.766 (0.714–0.818)	0.856 (0.815–0.898)
GBDT	0.750 (0.705–0.795)	0.749 (0.691–0.808)	0.731 (0.651–0.811)	0.734 (0.682–0.785)	0.833 (0.796–0.869)
XGBoost	0.777 (0.749–0.805)	0.778 (0.735–0.821)	0.756 (0.686–0.827)	0.761 (0.725–0.797)	0.825 (0.785–0.864)

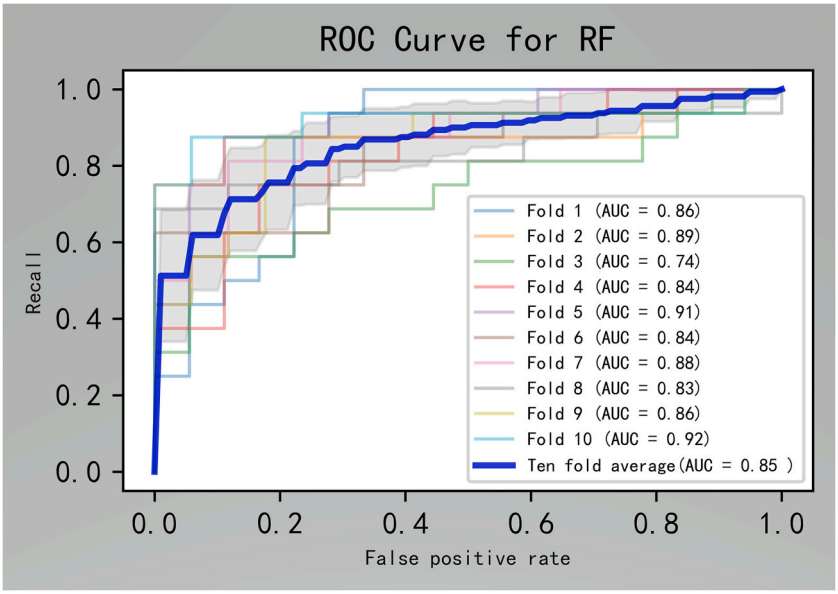


Fig. 3. Receiver operating characteristic curve of the random forest model.

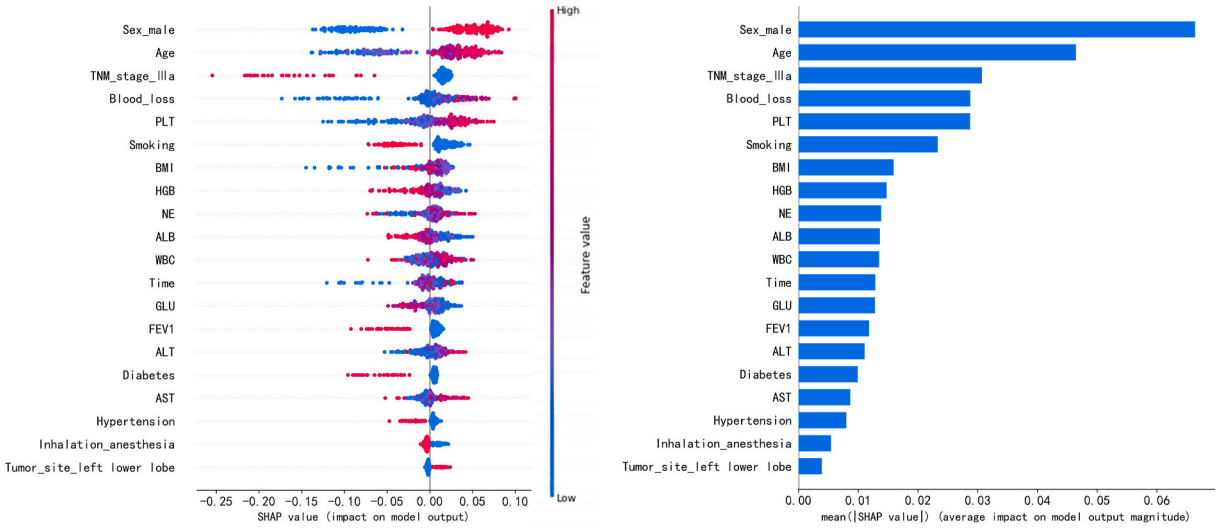


Fig. 4. The importance of the SHapley Additive exPlanation method based on the random forest model.

effects. Each point in the right halves of Figs. 4 and 5 represents the actual attributes of the patients as follows: (1) the color gradient from blue to red represents an increase in the variable value from low to high, (2) the virtualized variable 1 represents the variable has this attribute, and (3) the virtualized variable 0 represents the variable does not have this attribute. The magnitude and direction of the feature's effect on the prediction were quantified using SHAP values. We found that factors including gender, age, tumor stage,

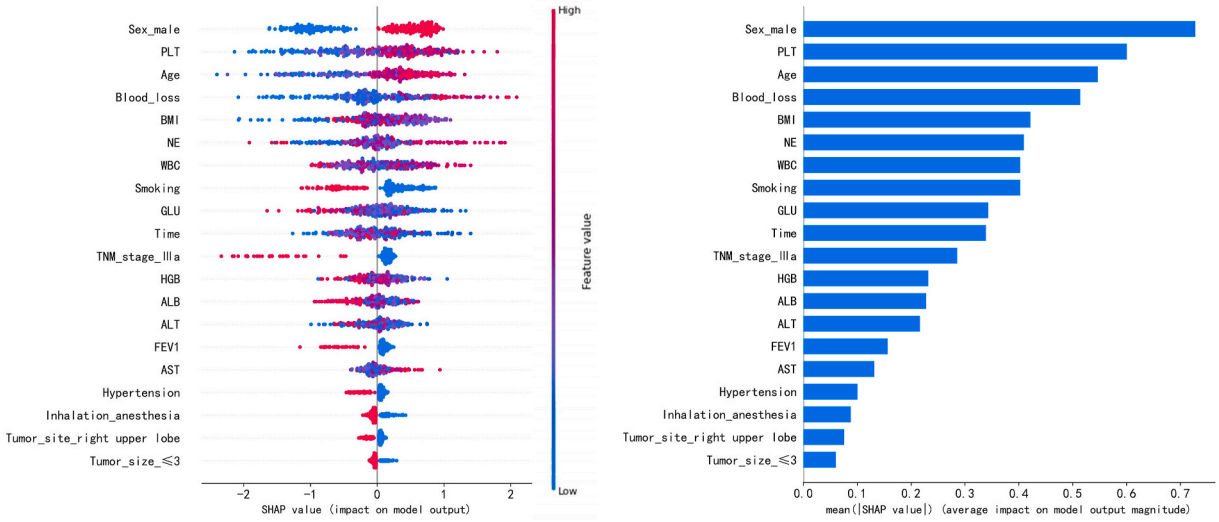


Fig. 5. The importance of the SHapley Additive exPlanation method based on the XGBoost model.

intraoperative blood loss, and platelet count were most important in the RF model. Furthermore, factors such as male gender, older age, earlier tumor stage, more blood loss during operation, and higher platelet count aggravated the occurrence of postoperative cardiopulmonary complications. The most important factors in the SHAP chart based on XGBoost were gender, platelet count, age, intraoperative blood loss, and BMI. According to the SHAP chart of the two models, gender, age, and intraoperative blood loss were the top affecting predictors of postoperative cardiopulmonary complications.

3.4. Interaction between variables

In retrospective studies, many researchers pay more attention to the recall rate [26], which refers to the proportion of true positive samples that are correctly predicted as positive. This rate can indicate the recognition ability of a classifier for positive classes. Among

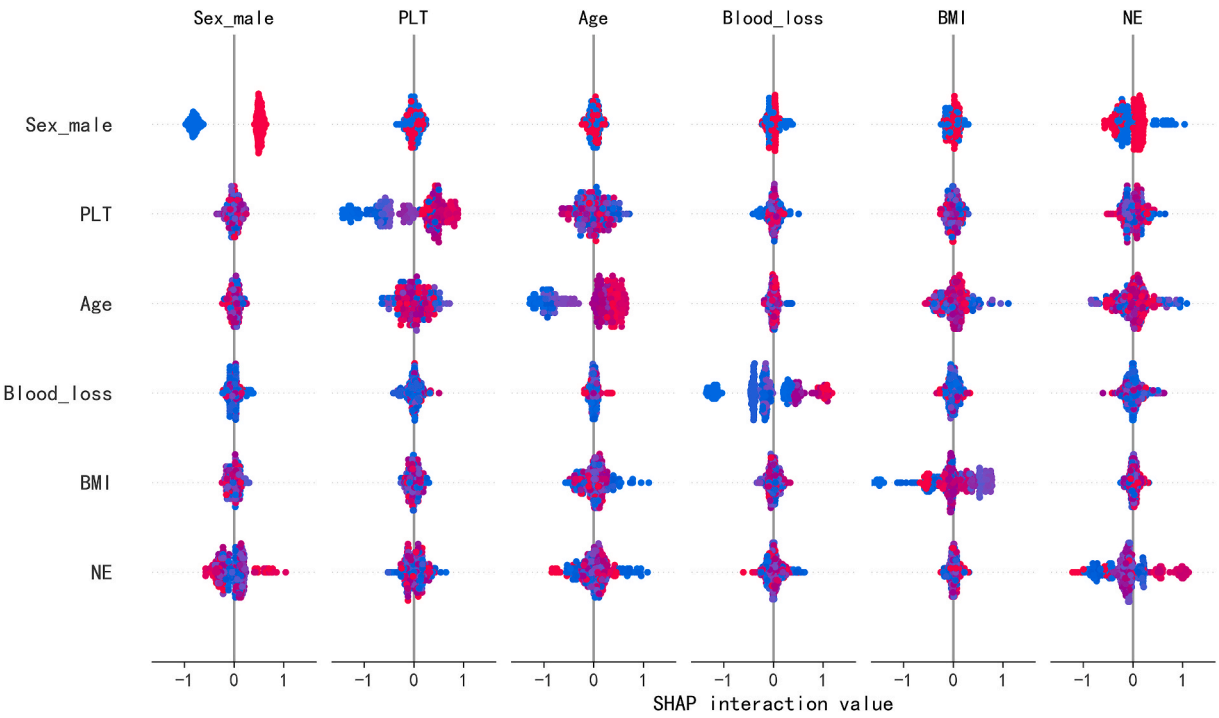


Fig. 6. Variable interactions based on the XGBoost model.

our five models, XGBoost, with a recall rate of 0.756, was the best model. XGBoost is one of the boosting algorithms. Its advantage lies in training; new learners rely on previous models and combine them according to a certain deterministic strategy to reduce deviations. We have drawn the interaction diagram of the top 6variables based on the XGBoost model (Fig. 6). The horizontal and vertical axes contain the top 6 variables in the SHAP chart. The greater the interaction between two variables, including age, PLT, and neutrophil counts, the greater the width displayed in the chart. However, the interaction between intraoperative blood loss and other variables was less, and most of the SHAP values were concentrated in the negative area, indicating that intraoperative blood loss might be an independent risk factor.

4. Discussion

In this study, we develop an ML-based model for predicting cardiopulmonary complications, with the goal of enabling early prevention and treatment. The model was designed to be interpretable, allowing for easier integration into clinical practice. The incidence of cardiopulmonary complications was 20.89%. The most common cardiopulmonary complications were pleural effusion/pneumothorax, followed by arrhythmia. Previous studies showed pulmonary infection as the most common complication [13,14]. However, our patients consumed prophylactic antibiotics before or during surgeries and actively dealt with possible infections; thus, the number of patients with pulmonary infections was less in the present study.

Among the five ML-based classification models, each comprehensive prediction index of the DT, RF, GBDT, and XGBoost models was higher than that of the LR model. Regarding accuracy, RF showed an accuracy value of 0.786, followed by 0.777 of XGBoost, 0.750 of GBDT, 0.696 of DT and 0.679 of LR. These results indicated that ML was practically valuable in clinical practice and could improve the performance of risk assessment tools. In the training process of RF, multiple DTs were generated, i.e., the classification results of several weak classifiers were selected by voting to form a strong classifier. The forest composed of multiple DTs had higher prediction accuracy than a single DT, and it also eliminated the problem of over-fitting of a single DT [27], thus showing good predictability for patients with cardiopulmonary complications. XGBoost created a tree based on the classification rules of category variables. The model could well handle the combination of multi-class and continuous variables and effectively improve the recall rate, and the comprehensive performance was not weak. Thus, consistent with other studies, ML has great advantages in prediction [28,29].

LR is a form of generalized linear regression model that can provide clear explanations for the analysis results [30]; however, its prediction accuracy was low in our study. We speculated that linear and non-linear relationships are present between the baseline characteristics of patients and postoperative cardiopulmonary complications. However, other models were not affected by the non-linear relationship and showed better comprehensive performance.

We calculated the kappa value of the RF model in the test set, which was 0.696. Compared with other metrics, Kappa values are a reliable indicator for evaluating the performance of classification models because they are not affected by imbalanced data and offer a more precise evaluation. According to common standards, when the kappa value is > 0.6, the classifier is considered to have good performance. Therefore, we concluded that the RF model performed well in the test set, exhibiting high accuracy. This result also indicates that our model training has certain advantages and can be used for predictive tasks in practical scenarios.

Previous ML models could not reveal the mechanism of their internal systems, even though they performed well. It can be difficult to determine which patient features are most relevant for accurate prediction when using ML algorithms, given their black box nature and lack of interpretability. We overcame this limitation using the SHAP method and provided the importance maps of RF and XGBoost models, which clearly and intuitively showed the importance of variables. The SHAP diagram helped doctors understand which factors contributed to having a higher or lower risk of cardiopulmonary complications. In clinical practice, the SHAP importance characteristic map based on the RF and XGBoost models is more convenient to use for the following reasons. First, it can be used in preoperative communication. More predictive factors can be combined with patients' personalized information for surgical communication and can intuitively explain which characteristics put such patients at risk. Second, doctors can use this model to estimate risks and provide information to patients for decision-making. Doctors can help patients via personalized prevention or intensive treatment strategies. For example, the time of postoperative electrocardiogram monitoring for a high-risk population can be prolonged and the monitoring of a low-riskpopulation can be reduced.

We further determined the predictors of postoperative cardiopulmonary complications caused by patients' baseline and surgical characteristics. We found that gender, age, and intraoperative blood loss were important factors in the ranking of cardiopulmonary complications. Other factors included BMI, smoking, hypertension, diabetes, time duration, tumor location, blood indicators, and lung function, which were listed as the top 20 important variables. These results are consistent with previous study results [31,32]. Due to data defects, we only distinguished whether patients smoke or not without further calculating the smoking index, which led to smoking becoming a protective factor in our study. Notably, we found that many blood factors, including albumin, PLT count, HB levels, blood sugar levels, and liver functions, played a role in predicting postoperative cardiopulmonary complications. The combined effects of these factors require further exploration. A previous study has shown a relationship between multiple blood routine tests and lung cancer diagnosis [33,34]. Blood examination may be associated with some complex events. Thus, more conventional blood indicators can be combined to predict postoperative cardiopulmonary complications in the future.

4.1. Limitations

This study is subject to certain limitations that warrant acknowledgement. First, the study was performed at a single-center, and the sample size is not large. Although cross-validation was used to avoid over-fitting the model, the ML model could not reach the optimal state. Second, this is a retrospective study, and the survey content lacks relevant information, such as cardiopulmonary complications

related to variables including smoking index and smoking cessation duration [35], which may help improve prediction efficiency. Lastly, we selected only five single ML algorithms to predict cardiopulmonary complications after lobectomy for patients with NSCLC. Future studies should use more and newer or integrated ML algorithms to achieve accurate prediction efficiency.

5. Conclusion

To conclude, we have established an ML model that accurately predicts the likelihood of cardiopulmonary complication occurrence after thoracoscopic lobectomy in patients with NSCLC. The findings indicated that gender, age, and intraoperative blood loss were important predictors of cardiopulmonary complications after NSCLC. Furthermore, RF showed the best prediction performance among all models. Relevant risk factors were determined by combining the SHAP ML interpretability method, which can be used as a tool to assist clinicians in formulating targeted interventions and making better medical decisions. However, our findings should be externally validated and ML ensemble models should be considered in the future. Nonetheless, our results provide valuable insights into predicting cardiopulmonary complications after thoracoscopic lobectomy in patients with NSCLC.

Author contribution statement

Yihai Zhai: Conceived and designed the experiments, Performed the experiments, Wrote the paper.
Xue Lin: Performed the experiments, Wrote the paper.
Qiaolin Wei: Conceived and designed the experiments, Analyzed and interpreted the data.
Yuanjin Pu: Conceived and designed the experiments, Analyzed and interpreted the data.
Yonghui Pang: Contributed reagents, materials, analysis tools or data, Wrote the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Bullet Edits Limited for the linguistic editing and proofreading of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e17772.

References

[1] C. Xia, X. Dong, H. Li, et al., Cancer statistics in China and United States, 2022: profiles, trends, and determinants, Chin. Med. J. 135 (5) (2022) 584–590, https://doi.org/10.1097/CM9.0000000000002108. Published 2022 Feb 9.
[2] E. Lim, T. Batchelor, M. Shackcloth, et al., Study protocol for VIdeo assisted thoracoscopic lobectomy versus conventional Open LobEcTomy for lung cancer, a UK multicentre randomised controlled trial with an internal pilot (the VIOLET study), BMJ Open 9 (10) (2019), e029507.
[3] R. Oda, K. Okuda, S. Osaga, et al., Long-term outcomes of video-assisted thoracoscopic surgery lobectomy vs. thoracotomy lobectomy for stage IA non-small cell lung cancer, Surg. Today 49 (5) (2019) 369–377, https://doi.org/10.1007/s00595-018-1746-4.
[4] Y. Tsutani, N. Tsubokawa, M. Ito, et al., Postoperative complications and prognosis after lobar resection versus sublobar resection in elderly patients with clinical Stage I non-small-cell lung cancer, Eur. J. Cardio. Thorac. Surg. 53 (2) (2018) 366–371, https://doi.org/10.1093/ejcts/ezx296.
[5] S. Okada, M. Shimomura, S. Ishihara, et al., Clinical significance of postoperative pulmonary complications in elderly patients with lung cancer, Interact. Cardiovasc. Thorac. Surg. 35 (2) (2022).
[6] R. Li, M. Xue, Z. Ma, et al., Construction and validation of a nomogram for predicting prolonged air leak after minimally invasive pulmonary resection, World J. Surg. Oncol. 20 (1) (2022) 249, https://doi.org/10.1186/s12957-022-02716-w. Published 2022 Aug 3.
[7] H. Ishibashi, R. Wakejima, A. Asakawa, et al., Postoperative atrial fibrillation in lung cancer lobectomy-analysis of risk factors and prognosis, World J. Surg. 44 (11) (2020) 3952–3959, https://doi.org/10.1007/s00268-020-05694-w.
[8] W. Baar, A. Semmelmann, J. Knoerlein, et al., Risk factors for postoperative pulmonary complications leading to increased in-hospital mortality in patients undergoing thoracotomy for primary lung cancer resection: a multicentre retrospective cohort study of the German thorax registry, J. Clin. Med. 11 (19) (2022) 5774, https://doi.org/10.3390/jcm11195774.
[9] K. Kaufmann, S. Heinrich, Minimizing postoperative pulmonary complications in thoracic surgery patients, Curr. Opin. Anaesthesiol. 34 (1) (2021) 13–19, https://doi.org/10.1097/ACO.0000000000000945.
[10] G.P. Copeland, D. Jones, M. Walters, POSSUM: a scoring system for surgical audit, Br. J. Surg. 78 (3) (1991) 355–360, https://doi.org/10.1002/bjs.1800780327.
[11] J. Nithiuthai, A. Siriussawakul, R. Junkai, et al., Do ARISCAT scores help to predict the incidence of postoperative pulmonary complications in elderly patients after upper abdominal surgery? An observational study at a single university hospital, Perioperat. Med. 10 (1) (2021) 43.
[12] K. Brat, P. Homolka, Z. Merta, et al., Prediction of postoperative complications: ventilatory efficiency and rest end-tidal carbon dioxide [published online ahead of print, 2022 jan 21], Ann. Thorac. Surg. S0003–4975 (22) (2022) 60–61, https://doi.org/10.1016/j.athoracsur.2021.11.073.
[13] G. Chouinard, P. Roy, M.C. Blais, et al., Exercise testing and postoperative complications after minimally invasive lung resection: a cohort study, Front. Physiol. 13 (2022), 951460.

[14] T. Nakada, Y. Noda, D. Kato, et al., Risk factors and cancer recurrence associated with postoperative complications after thoracoscopic lobectomy for clinical stage I non-small cell lung cancer, Thorac Cancer 10 (10) (2019) 1945–1952, https://doi.org/10.1111/1759-7714.13173.
[15] P. Agostini, H. Cieslik, S. Rathinam, et al., Postoperative pulmonary complications following thoracic surgery: are there any modifiable risk factors? Thorax 65 (9) (2010) 815–818, https://doi.org/10.1136/thx.2009.123083.
[16] G. Huang, L. Liu, L. Wang, Z. Wang, Z. Wang, S. Li, External validation of five predictive models for postoperative cardiopulmonary morbidity in a Chinese population receiving lung resection, PeerJ 10 (2022), e12936.
[17] N. Motono, M. Ishikawa, S. Iwai, et al., Analysis of risk factors for postoperative complications in non-small cell lung cancer: comparison with the Japanese National Clinical Database risk calculator, BMC Surg. 22 (1) (2022) 180, https://doi.org/10.1186/s12893-022-01628-6.
[18] C.T. Bevilacqua Filho, A.P. Schmidt, E.A. Felix, et al., Risk factors for postoperative pulmonary complications and prolonged hospital stay in pulmonary resection patients: a retrospective study, Braz J Anesthesiol 71 (4) (2021) 333–338, https://doi.org/10.1016/j.bjane.2021.02.003.
[19] S. Gupta, R.J. Fernandes, J.S. Rao, et al., Perioperative risk factors for pulmonary complications after non-cardiac surgery, J. Anaesthesiol. Clin. Pharmacol. 36 (1) (2020) 88–93, https://doi.org/10.4103/joacp.JOACP_54_19.
[20] M.T. Lu, V.K. Raghu, T. Mayrhofer, et al., Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model, Ann. Intern. Med. 173 (9) (2020) 704–713, https://doi.org/10.7326/M20-1868.
[21] Y. Wang, S. Liu, Z. Wang, et al., A machine learning-based investigation of gender-specific prognosis of lung cancers, Medicina (Kaunas). 57 (2) (2021) 99, https://doi.org/10.3390/medicina57020099.
[22] S. Zeng, L. Li, Y. Hu, et al., Machine learning approaches for the prediction of postoperative complication risk in liver resection patients, BMC Med. Inf. Decis. Making 21 (1) (2021) 371, https://doi.org/10.1186/s12911-021-01731-3.
[23] J. Canet, L. Gallart, C. Gomar, et al., Prediction of postoperative pulmonary complications in a population-based surgical cohort, Anesthesiology 113 (6) (2010) 1338–1350, https://doi.org/10.1097/ALN.0b013e3181fc6e0a.
[24] Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance[J], Pattern Recogn. Lett. 93 (2017) 3–12, https://doi.org/10.1016/j.patrec.2016.10.006.
[25] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, 31st Annual Conference on Neural Information Processing Systems (NIPS) 30 (2017) 4765–4774.
[26] C.S. Lee, C. Parise, J. Burleson, et al., Assessing the recall rate for screening mammography: comparing the medicare hospital compare dataset with the national mammography database, AJR Am. J. Roentgenol. 211 (1) (2018) 127–132, https://doi.org/10.2214/AJR.17.19229.
[27] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[28] M. Hanko, M. Grendár, P. Snopko, et al., Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy, World Neurosurg 148 (2021) e450–e458, https://doi.org/10.1016/j.wneu.2021.01.002.
[29] X. Zhao, C. Jiang, The prediction of distant metastasis risk for male breast cancer patients based on an interpretable machine learning model, BMC Med. Inf. Decis. Making 23 (1) (2023) 74, https://doi.org/10.1186/s12911-023-02166-8.
[30] C.S. Lee, C. Conway, The role of generalized linear models in handling cost and count data, Eur. J. Cardiovasc. Nurs. 21 (4) (2022) 392–398, https://doi.org/10.1093/eurjcn/zvac002.
[31] M. Benker, N. Citak, T. Neuer, et al., Impact of preoperative comorbidities on postoperative complication rate and outcome in surgically resected non-small cell lung cancer patients, Gen Thorac Cardiovasc Surg 70 (3) (2022) 248–256, https://doi.org/10.1007/s11748-021-01710-5.
[32] S. Li, K. Zhou, Y. Lai, et al., Estimated intraoperative blood loss correlates with postoperative cardiopulmonary complications and length of stay in patients undergoing video-assisted thoracoscopic lung cancer lobectomy: a retrospective cohort study, BMC Surg. 18 (1) (2018), https://doi.org/10.1186/s12893-018-0360-0.
[33] J. Wu, X. Zan, L. Gao, et al., A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study, JMIR Med Inform 7 (3) (2019), e13476, https://doi.org/10.2196/13476.
[34] R. Zu, L. Wu, R. Zhou, et al., A new classifier constructed with platelet features for malignant and benign pulmonary nodules based on prospective real-world data, J. Cancer 13 (8) (2022) 2515–2527, https://doi.org/10.7150/jca.67428. Published 2022 May 9.
[35] M.K. Gould, B.Z. Huang, M.C. Tammemagi, et al., Machine learning for early lung cancer identification using routine clinical and laboratory data, Am. J. Respir. Crit. Care Med. 204 (4) (2021) 445–453, https://doi.org/10.1164/rccm.202007-2791OC.