Respiratory Research

RESEARCH

Open Access

Machine learning-based Diagnostic model for determining the etiology of pleural effusion using Age, ADA and LDH

Qing-Yu Chen[1], Shu-Min Yin[1], Ming-Ming Shao[1,2], Feng-Shuang Yi[1,2*] and Huan-Zhong Shi[1*]

Abstract

Background Classification of the etiologies of pleural effusion is a critical challenge in clinical practice. Traditional diagnostic methods rely on a simple cut-off method based on the laboratory tests. However, machine learning (ML) offers a novel approach based on artificial intelligence to improving diagnostic accuracy and capture the non-linear relationships.

Method A retrospective study was conducted using data from patients diagnosed with pleural effusion. The dataset was divided into training and test set with a ratio of 7:3 with 6 machine learning algorithms implemented to diagnosis pleural effusion. Model performances were assessed by accuracy, precision, recall, F1 scores and area under the receiver operating characteristic curve (AUC). Feature importance and average prediction of age, Adenosine (ADA) and Lactate dehydrogenase (LDH) was analyzed. Decision tree was visualized.

Results A total of 742 patients were included (training cohort: 522, test cohort: 220), 397 (53.3%) diagnosed with malignant pleural effusion (MPE) and 253 (34.1%) with tuberculous pleural effusion (TPE) in the cohort. All of the 6 models performed well in the diagnosis of MPE, TPE and transudates. Extreme Gradient Boosting and Random Forest performed better in the diagnosis of the MPE, with F1 scores above 0.890, while K-Nearest Neighbors and Tabular Transformer performed better in the diagnosis of the TPE, with F1 scores above 0.870. ADA was identified as the most important feature. The ROC of machine learning model outperformed those of conventional diagnostic thresholds.

Conclusions This study demonstrates that ML models using age, ADA, and LDH can effectively classify the etiologies of pleural effusion, suggesting that ML-based approaches may enhance diagnostic decision-making.

Keywords Pleural effusion, Machine learning, Diagnostic model, Adenosine deaminase

Introduction

Pleural effusion is the accumulation of the fluid in the pleural cavity and often occurs in the clinical practice. The effective management requires identification of its underlying etiology [1]. The most common etiologies include congestive heart failure, pneumonia, and cancer [2]. However, existing diagnostic methods have limitations. Thoracentesis with fluid analysis is widely used to diagnose pleural effusion [3], but the diagnostic accuracy for malignant pleural effusion (MPE) varies widely [4]. In regions with a high tuberculosis burden, tuberculosis

*Correspondence:
Feng-Shuang Yi
yifengshuang@ccmu.edu.cn
Huan-Zhong Shi
shihuanzhong@sina.com
[1] Department of Respiratory and Critical Care Medicine, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing 100020, China
[2] Medical Research Center, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing 100020, China

pleural effusion (TPE) constitutes a larger proportion [5]. Light's criteria, though commonly employed, misclassify approximately 25% of transudates as exudates [6]. Furthermore, diagnosing parapneumonic effusion (PPE) is challenging, particularly in excluding other causes, as there are no definitive criteria for diagnosing uncomplicated parapneumonic effusion [7]. Thus, new tools to facilitate diagnosis are needed.

Though some invasive procedures, such as pleural needle biopsy and thoracoscopy, can provide definitive pathological diagnoses, but they carry a risk of complications, require time for pathological analysis and depend on the experience of the pathologists [8]. These challenges highlight the need for integrated diagnostic methods based on objective laboratory tests, which can support clinical decision-making and offer crucial diagnostic information for pleural effusion in a more efficient and less invasive way. The adenosine deaminase (ADA) in the pleural effusion is a biomarker for TPE, with a summary sensitivity and specificity of 0.92 and 0.90 respectively [9]. Lactate dehydrogenase (LDH) enhances the specificity in the detection of malignant and inflammatory exudates and is a key laboratory test in the Light's criteria [10]. Moreover, the diagnosis accuracy of pleural fluid ADA was affected by age [11, 12]. These laboratory-based biomarkers and demographic characteristic were considered as the potential features in developing more efficient diagnostic models for pleural effusion.

As the development of the algorithms and techniques, machine learning has been used in the diagnosis of various diseases [13], including pleural effusion. Machine learning approaches, unlike traditional methods based on predefined cut-off values, excel in capturing complex, non-linear relationships among variables [14]. Several studies have applied machine learning models with various features, such as demographic characteristics, clinical symptoms, blood and pleural fluid analyses, cytopathological slides, radiomic features, and even image-based data [9, 15–33]. The majority of these models have incorporated more than ten features. Although the results of these studies show promising AUC values, the inclusion of those features leads to an increasing number of tests, thereby raising the laboratory test expenses for patients. Therefore, we used machine learning model with fewer, yet clinically significant, features for the diagnosis of the pleural effusion.

In this study, we selected age, pleural fluid ADA, and pleural fluid LDH as the features and constructed diagnostic models for pleural effusion. We applied six machine learning techniques: multinomial linear regression (LR), support vector machine (SVM), Extreme Gradient Boosting (XGBoost), random forest (RF), K-Nearest Neighbors (KNN) and Tabular Transformer (TabTransformer), aiming to construct efficient models for improved diagnostic accuracy.

Method

Participants

This retrospective study included inpatients from Beijing Chao-Yang Hospital between January 2014 and May 2024. Patients with pleural effusion and underwent diagnostic thoracentesis were included in this study. Those with unclear or multiple etiologies were excluded. This study approved by the Beijing Chao-Yang Hospital affiliated to Capital Medical University (2018-ke-321, 2024-ke-502). Given the retrospective design of this study, the informed consent was not required for this study.

Features and Diagnostic criteria

The exclusion criteria were as follows: 1) undetermined etiologies of the pleural effusion, empyema, chylothorax. 2) Patients with incomplete clinical data. 3 features (Age, fluid ADA, fluid LDH) were collected from the patients' medical record system. If multiple results for ADA and LDH are available from the pleural fluid tests, the first result after the thoracentesis will be selected. Five main etiologies of pleural effusion were classified: Malignant pleural effusion (MPE), Tuberculous pleural effusion (TPE), Parapneumonic pleural effusion (PPE), transudative pleural effusion, other causes.

Malignant pleural effusion was defined as the pathologic findings of malignancy in the pleural effusion or the pleura. Tuberculous pleural effusion was defined as following criteria: 1) Acid-fast bacilli smear or culture were positive for Mycobacterium tuberculosis in sputum, pleural fluid, and bronchoalveolar lavage fluid; 2) Mycobacterium tuberculosis positive in bronchoalveolar brush samples, lung, or pleural biopsy; 3) caseous granuloma in pleura or lung; or 4) The ratio of lymphocytes to neutrophils in the pleural effusion exceeded 0.75, and the fluid ADA were above 40 IU/L, with effective antituberculosis treatment and other causes of pleural effusion excluded. Other etiologies of pleural effusion were classified as other causes, such as immune-related etiologies. Parapneumonic pleural effusion was diagnosed as the effusion was defined as exudative and associated with pneumonia, with other etiologies excluded.

Study design

The sample size was calculated using the following formula: $\mathbf{N} = \frac{\mathbf{Z}^2 \times \mathbf{P} \times (1-\mathbf{P})}{\mathbf{d}^2}$,

N = required sample size, Z = Z-value, set to 1.96 for a 95% confidence interval, P = Expected model accuracy, d = Margin of error.

$$\mathbf{N} = \frac{1.96^2 \times 0.85 \times (1 - 0.85)}{0.05^2} = 196$$

To achieve the expected total accuracy of 0.85 with a margin of error of 0.05 at a 95% confidence level, a minimum of 196 samples in the train set was required.

The patient datasets from Beijing Chao-Yang Hospital were divided into training and test sets with randomization both in a 7:3 ratio, resulting in 522 cases in the training set. As the patients with missing data for age, pleural fluid ADA, and pleural fluid LDH were excluded from the dataset, no imputations were applied. The datasets were centered to a mean of 0 and scaled to a standard deviation of 1 for each feature. Six machine learning methods were used to construct diagnostic models: LR, SVM, XGBoost, RF, KNN and Tab Transformer. Bayesian optimization was employed to tune the hyperparameters of the models. The details of the hyperparameters in each model were listed (Supplementary Table 1).

As a comparation to the machine learning models, traditional diagnostic methods based on the cut-off values were applied to assess the performance. For MPE, the cancer ratio, defined as the ratio of blood LDH to pleural fluid ADA, was employed, with a threshold value set at greater than 20 [34].Similarly, for TPE, a pleural fluid ADA level greater than 40 U/L was used as the diagnostic criterion [35].

Primary outcome and performance metrics

The primary output of this study is the classification of the etiological types of pleural effusion, which include MPE, TPE, PPE, transudative effusion, and other causes. The primary endpoint of the study was the diagnostic performance of the machine learning models in classifying pleural effusion etiology. The performance of the models was evaluated based on the accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC). True positives (TP), true negatives (TN), false positive (FP) and false negatives (FN) were obtained from the confusion matrix. The parameters were calculated by the following formula: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$, F1 Score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. The AUC is calculated based on the true positive rate (TPR) and false positive rate (FPR) across different thresholds. Feature importance was assessed to determine the contribution of the selected features. The feature importance in XGBoost was assessed by gain, while in the RF model, it was evaluated by mean decrease in Gini. Plot of the first decision tree form the RF model is presented to illustrate the splitting logic and feature importance at the individual tree level. The average prediction of each feature on different etiologies were assessed and visualized in a partial dependence plot. Bootstrap resampling on the test set were used to provide a reliable assessment of model performance by calculating the mean accuracy and AUC along with their 95% confidence intervals.

Statistical analysis

Qualitative data (gender and disease classification) were summarized as frequencies and percentages. Chi-square tests were used to assess significant differences between groups for categorical variables. Quantitative data included Age, ADA, LDH, Total protein, Glucose, Chloride, Total cell counts, and mononuclear cell percentage levels. Normality testing was performed using the Shapiro–Wilk test.

For normally distributed data, results are presented as mean ± standard deviation (SD), and comparisons between groups were made using independent t-tests for two groups or one-way analysis of variance (ANOVA) for more than two groups. For non-normally distributed data, values are expressed as median (interquartile range, IQR), and non-parametric tests, such as the Mann–Whitney U test for two groups or Kruskal–Wallis H test for more than two groups, were used. When the Kruskal–Wallis test indicated significant differences, Dunn's test was used for pairwise comparisons to assess specific group differences. Pearson correlation coefficient was calculated to assess the strength and direction of the relationship. P-values < 0.05 were considered statistically significant.

All the statistical analyses were performed using R (version 4.2.3) or Python (version 3.11). More detailed information about the necessary packages and their versions can be found in the supplementary file (Supplementary Files 1–4).

Results

Baseline information of the cohort

1172 patients in Beijing Chao-Yang Hospital underwent diagnostic thoracentesis during the specific time, 430 patients were excluded (Fig. 1). The basic clinical information as well as the cytological and biochemical tests of pleural effusion in total, in the training set and the test set are shown (Table 1). In total, the etiologies of the pleural effusion were as the followings: malignant pleural effusion (53.5%), tuberculous pleural effusion (34.1%), parapneumonic pleural effusion (4.2%), Transudative pleural effusion (4.4%), others (3.8%). The clinical characteristics classified by the etiologies are listed (Table 2).

The relationship among Age, ADA and LDH

Three factors (Age, ADA, LDH) were used as the feature in the machine leaning. The distribution of the features

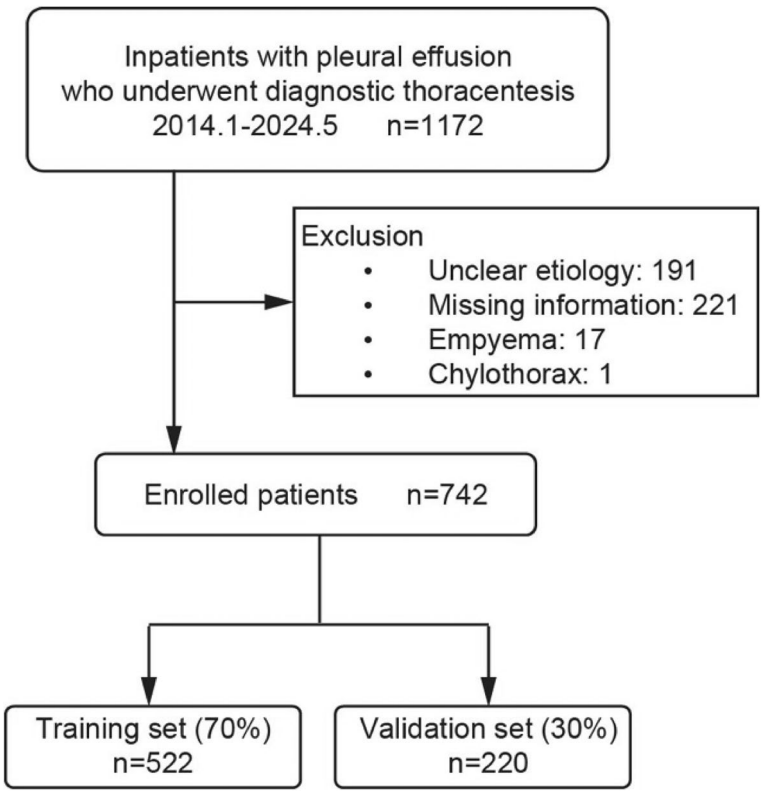

Fig. 1 Workflow chart of the patient enrollment

Table 1 Clinical characteristic of the cohorts

	Beijing Chao-Yang Cohort	Training Set	Validation Set	*p*
n	742	522	220	
Diagnosis (%)				0.998
Malignant pleural effusion	397 (53.5)	278 (53.3)	119 (54.1)	
Tuberculous pleural effusion	253 (34.1)	178 (34.1)	75 (34.1)	
Parapneumonic pleural effusion	31 (4.2)	22 (4.2)	9 (4.1)	
Transudative pleural effusion	33 (4.4)	24 (4.6)	9 (4.1)	
Others	28 (3.8)	20 (3.8)	8 (3.6)	
Age (median [IQR]), years	64 [51, 74]	64 [50, 73]	65 [52, 75]	0.368
Sex (%)				0.806
Female	280 (37.7)	195 (37.4)	85 (38.6)	
Male	462 (62.3)	327 (62.6)	135 (61.4)	
Pleural effusion				
ADA (median [IQR]), U/L	15 [9, 36]	15.00 [9, 36]	14 [9, 36]	0.911
LDH (median [IQR]), U/L	302 [186, 488]	285 [184, 479]	330 [191, 515]	0.104
Total protein (median [IQR]), g/L	47.4 [41.5, 51.8]	47.2 [41.5, 51.9]	47.5 [41.5, 51.3]	0.931
Glucose (median [IQR]), mmol/L	5.94 [4.63, 7.29]	6.00 [4.63, 7.29]	5.74 [4.64, 7.32]	0.578
Chloride (median [IQR]), mmol/L	106.2 [103.4, 108.8]	106.2 [103.2, 109.0]	106.5 [103.7, 108.6]	0.988
Total cell counts (median [IQR]), cells/μl	6972 [3105, 22154]	6903 [3037,22406]	7087 [3313, 20708]	0.931
Mononuclear cell percentage (median [IQR]), %	91[78, 96]	90 [78, 96]	91 [78, 96]	0.839

Data are presented as the median [IQR] for Non-normally distributed continuous variables, number (%) for categorical variables. For non-normally distributed continuous variables, Mann–Whitney U test was used for comparison; for categorical variables, Chi-square test was used for comparison. *ADA* Adenosine Deaminase, *LDH* lactate dehydrogenase

Table 2 Clinical characteristic according to the cohorts

	Malignant pleural effusion	Tuberculous pleural effusion	Parapneumonic pleural effusion	Transudative pleural effusion	Others
n	397	253	31	33	28
Age (median [IQR]), years	67 [58, 74]	53 [29, 69]	67 [54, 80]	73 [66, 80]	61 [50, 70]
Sex (%)					
Female	166 (41.8)	86 (34.0)	8 (25.8)	7 (21.2)	13 (46.4)
Male	231 (58.2)	167 (66.0)	23 (74.2)	26 (78.8)	15 (53.6)
Pleural effusion					
ADA (median [IQR]), U/L	11 [7, 15]	45 [32, 59]	12 [8, 18]	5 [4, 9]	17 [12, 28]
LDH (median [IQR]), U/L	319 [204, 498]	324 [223, 512]	171 [122, 232]	94 [78, 120]	178 [124, 464]
Total protein (median [IQR]), g/L	46.5 [41.4, 50.5]	50.4 [46.0, 54.1]	45.2 [36.4, 51.5]	27.1 [20.6, 33.9]	48.0 [31.4, 51.2]
Glucose (median [IQR]), mmol/L	6.03 [4.60, 7.35]	5.51 [4.47, 6.55]	7.43 [6.18, 8.55]	7.11 [6.48, 8.60]	6.06 [5.48, 7.50]
Chloride (median [IQR]), mmol/L	106.4 [103.3, 108.9]	105.6 [103.2, 107.6]	108.8 [104.1, 110.1]	108.7 [105.4, 111.5]	107.5 [103.8, 109.0]
Total cell counts (median [IQR]), cells/μl	10,358 [2935, 44948]	6297 [3741, 10371]	4677 [2579, 18146]	1522 [318, 4198]	6824 [3333, 22789]
Mononuclear cell percentage (median [IQR]), %	87 [73, 93]	95 [88, 98]	87 [54, 94]	90 [83, 96]	89 [77, 96]

Data are presented as the median [IQR] for Non-normally distributed continuous variables, number (%) for categorical variables. *ADA* Adenosine Deaminase, *LDH* lactate dehydrogenase

across the different etiologies were compared (Fig. 2A), each feature was significant among groups. Pairwise comparisons were made between diagnostic groups for these features (Fig. 2B), most of the comparison were of significance. To assess if there were linear relationships between each pair of features, we drew the scatter plots and fit lines (Fig. 2C). ADA levels were negative associated with age. However, no linear relationship was observed between LDH and age. Moreover, there was a strong positive linear relationship between LDH and ADA. As these factors were important demographic and laboratory factors in the clinical decision-making and there were complex relationships among them, we constructed the diagnostic model based on these three features.

Model performance evaluation

The accuracy of LR, SVM, XGBoost, RF, KNN and Tab Transformer in train and test sets were presented (Table 3). The accuracies of XGBoost and RF models were high in both train set and test set, above 0.820. The accuracies between train set and test set show no overfitting in the models.

To evaluate the model performance in different etiologies, the ROC curves were plotted and AUC values were calculated for each etiology (Fig. 3, Table 4). All six models demonstrated high AUC values for the classification of MPE, TPE and transudates, which were above 0.890. The performance for PPE classification was generally around 0.700. To further evaluate the performance of these six models in the diagnosis of MPE and TPE, we calculated their precision, recall and F1 score (Table 5). All of this machine learning models have high recall above 0.950 in the diagnosis of MPE. XGBoost and RF performed better in the diagnosis of the MPE, while KNN and Tab Transformer performed better in the diagnosis of the TPE.

To obtain more robust estimates of model performance, we applied Bootstrap resampling to the test set and evaluated accuracy and AUC by calculating their averages and corresponding 95% confidence intervals (Table 6).

For comparison with the traditional cut-off method, we used commonly accepted diagnostic criteria from the literature. The AUC for MPE using the cut-off method with a cancer ratio greater than 20 was 0.670 and the AUC for TPE using a cut-off value of pleural fluid ADA greater than 40U/L was 0.800 (Fig. 4). Both values were lower than the AUCs obtained by the machine learning models for specific diagnosis. Also, we calculated the precision, recall and F1 score of traditional cut-off method (Table 5). All of the six models performed better than the traditional cut-off methods in the classification of both MPE and TPE.

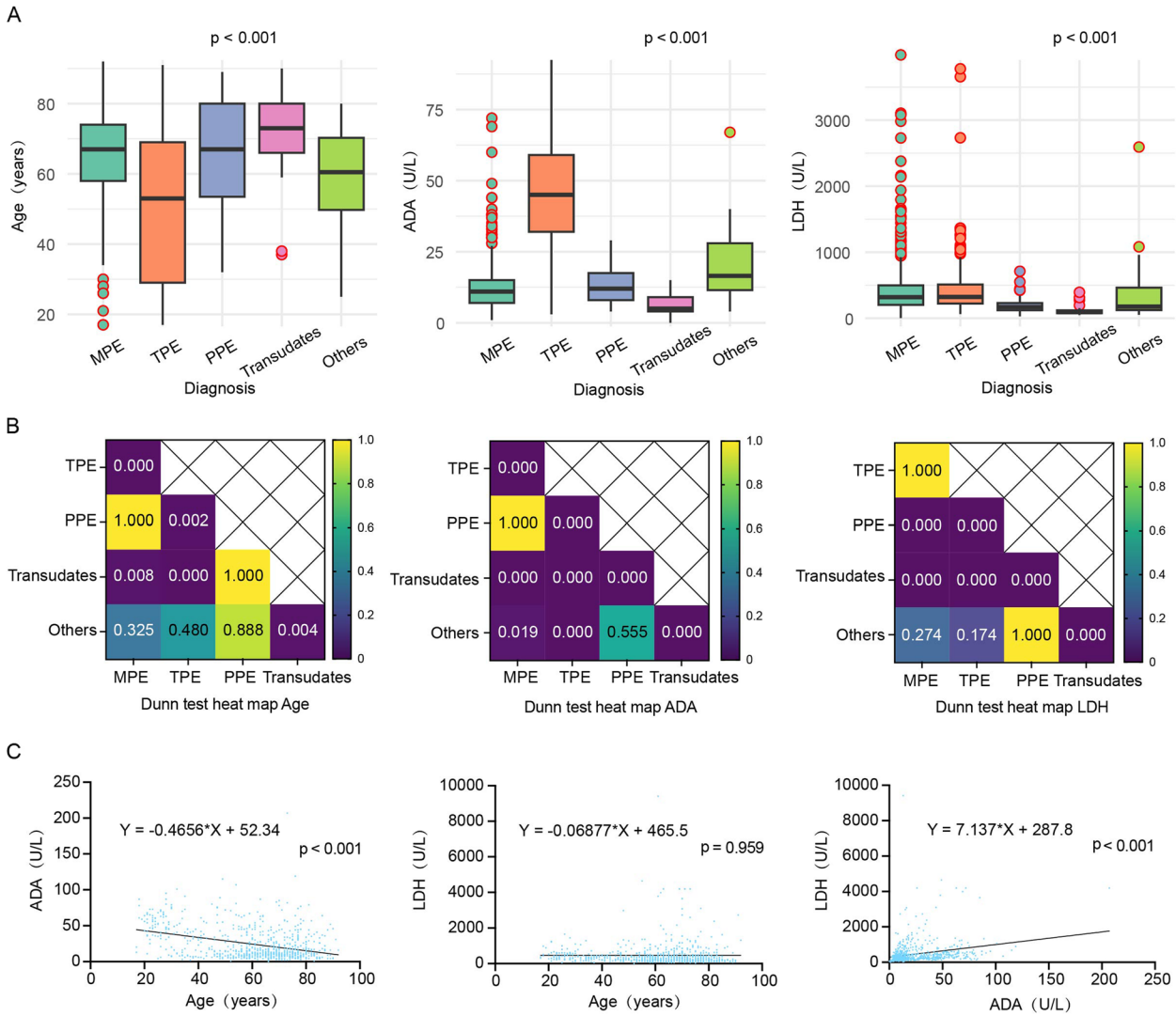

Fig. 2 Diagnostic group comparisons and relationships between variables. **A**, Box plots of Age, ADA, and LDH from left to right, showing the distribution of each variable across diagnostic groups, with outliers indicated. *P*-values for the differences between groups are calculated using Kruskal–Wallis tests. **B**, Dunn test heatmap displaying pairwise comparisons between diagnostic groups for Age, ADA, and LDH. The heatmap shows the significance of pairwise differences, with darker colors representing stronger statistical significance. MPE, malignant pleural effusion, TPE, tuberculous pleural effusion, PPE, parapneumonic pleural effusion. **C**, Scatter plots from left to right illustrating the relationships between Age and ADA, Age and LDH, and ADA and LDH. The plots include fitted curves and Pearson correlation coefficients to highlight the strength and direction of the associations between variables

Table 3 Model accuracy in train and test

	LR	SVM	XGBoost	RF	KNN	Tab Transformer
Accuracy in train set	0.785	0.795	0.835	0.828	0.803	0.753
Accuracy in test set	0.818	0.809	0.827	0.832	0.832	0.836

LR multinomial linear regression, SVM support vector machine, *XGBoost* Extreme Gradient Boosting, *RF*, random forest, *KNN* K-Nearest Neighbors, *Tab Transformer* Tabular Transformer

Impact of features on model prediction

To assess the feature importance in pleural effusion diagnosis, we ranked the feature contributions based on the gain in XGBoost model and the mean decrease in Gini in RF model (Fig. 5). In both models, ADA exhibited the largest importance, followed by LDH and age.

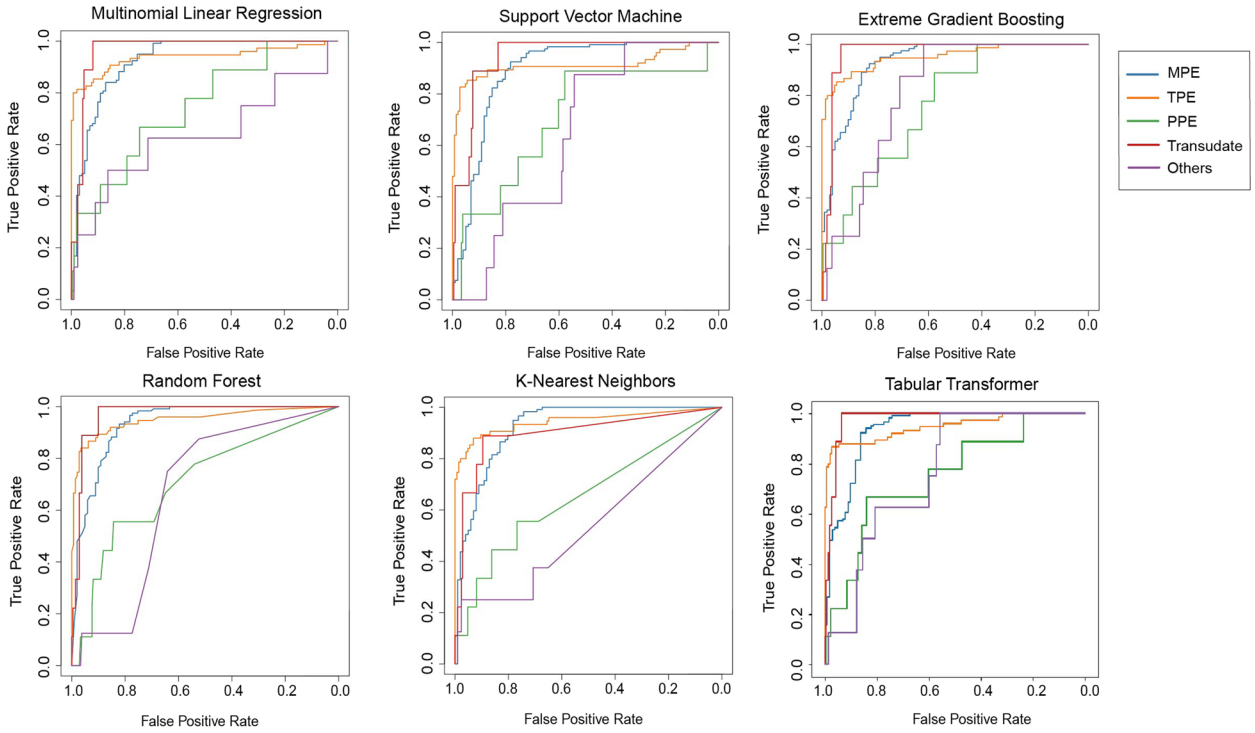

Fig. 3 Receiver operating characteristic curves of five models in the test set. MPE, malignant pleural effusion, TPE, tuberculous pleural effusion, PPE, parapneumonic pleural effusion

Table 4 Area under the receiver operating characteristic curve of single etiologies in each machine learning method in the test set

	LR	SVM	XGBoost	RF	KNN	Tab Transformer
MPE	0.925	0.891	0.931	0.934	0.925	0.935
TPE	0.939	0.913	0.953	0.950	0.946	0.947
PPE	0.745	0.706	0.765	0.706	0.652	0.752
Transudative	0.966	0.946	0.967	0.969	0.899	0.975
Others	0.636	0.644	0.813	0.666	0.538	0.766

MPE Malignant pleural effusion, *TPE* Tuberculous pleural effusion, *PPE* Parapneumonic pleural effusion, *Transudative* Transudative pleural effusion, *Others* Other causes. *LR* multinomial linear regression, *SVM* support vector machine, *XGBoost* Extreme Gradient Boosting, *RF* random forest, *KNN* K-Nearest Neighbors, *Tab Transformer* Tabular Transformer

Table 5 Precision, recall and F1 score in the diagnosis of MPE and TPE

	MPE			TPE		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Multinomial linear regression	0.780	0.983	0.870	0.910	0.813	0.859
Support vector machine	0.763	0.975	0.856	0.912	0.827	0.867
Extreme Gradient Boosting	0.832	0.958	0.891	0.877	0.853	0.865
Random forest	0.844	0.958	0.898	0.877	0.853	0.865
K-Nearest Neighbors	0.783	1.000	0.878	0.939	0.827	0.879
Tabular Transformer	0.779	0.975	0.866	0.952	0.800	0.870
Traditional method	0.482	0.803	0.602	0.613	0.958	0.748

MPE Malignant pleural effusion, *TPE* Tuberculosis pleural effusion

Table 6 Bootstrap Evaluation of Model Performance

	LR	SVM	XGBoost	RF	KNN	Tab Transformer
Accuracy	0.818 [0.768,0.868]	0.810 [0.759,0.859]	0.822 [0.768,0.868]	0.837 [0.786,0.886]	0.833 [0.782,0.882]	0.815 [0.786,0.841]
AUC	0.846 [0.786,0.911]	0.821 [0.768,0.870]	0.880 [0.834,0.923]	0.845 [0.796,0.894]	0.792 [0.726,0.857]	0.839 [0.775,0.902]

AUC Area under the receiver operating characteristic curve, *LR* multinomial linear regression, *SVM* support vector machine, *XGBoost* Extreme Gradient Boosting, *RF* random forest, *KNN* K-Nearest Neighbors, *Tab Transformer* Tabular Transformer

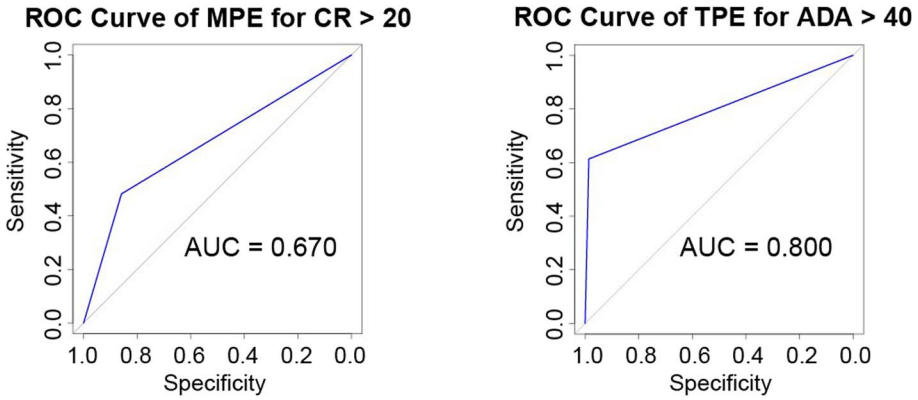

Fig. 4 Receiver operating characteristic curves of traditional methods. CR, cancer ratio; ADA, Adenosine deaminase

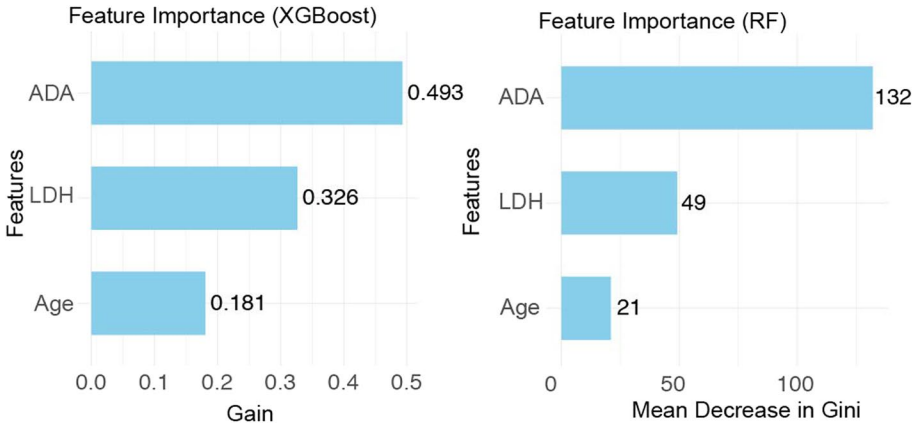

Fig. 5 Feature importance of extreme gradient boost measured and random forest. XGBoost, extreme gradient boost; RF, random forest; ADA, Adenosine deaminase; LDH, lactate dehydrogenase

To understand the process of decision-making in the RF model, we visualized the first tree (Fig. 6). The first split in the tree was based on ADA levels, the subsequent splits were based on LDH levels and the final split was made by age. This tree structure reflects the systematic process of the random forest model in handing multiple clinical variables in an interpretable way.

To assess the specific effects of the features in the XGBoost model, we drew the Partial Dependence Plots (Fig. 7), which indicated distinct patterns for ADA, Age and LDH in relationship to the etiological prediction. The average prediction of ADA elevated in the TPE cases, which indicates a strong association. The average prediction of age trend to elevated in MPE and reduced in TPE, which is consist with the typical patient demographics observed in these two etiologies. The curves of LDH show a marked increase in the MPE and sharp declines in PPE, transudates and other

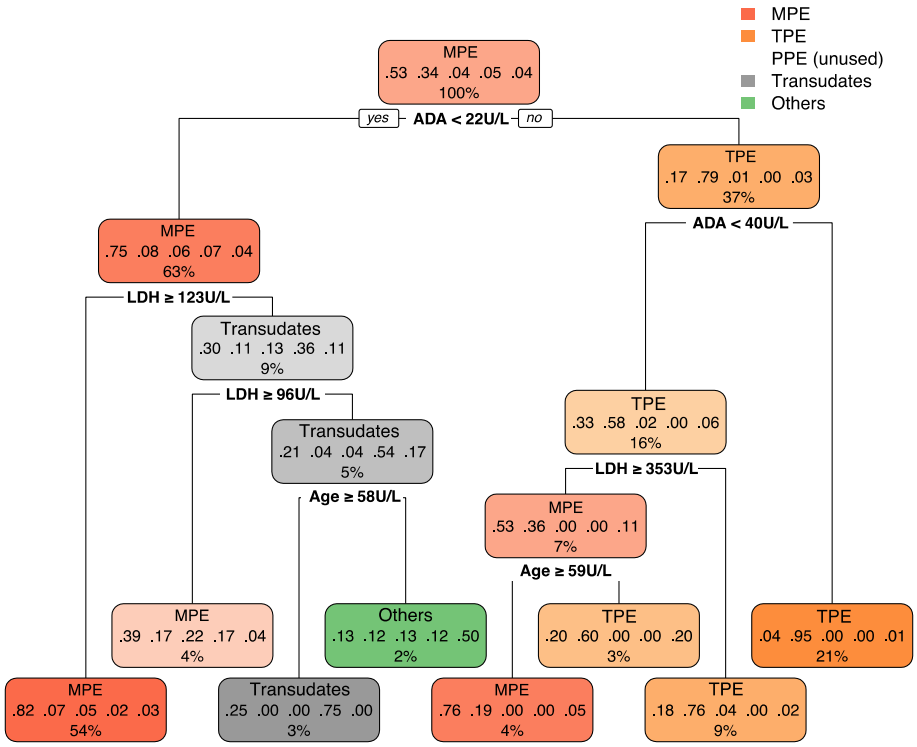

Fig. 6 Decision tree model for the pleural effusion etiology classification. The value list at each node shows the distribution of samples across different classes, with percentages indicating the proportion of cases for each class within that node. Each class represents an etiological category in the pleural effusion diagnosis, and the tree splits the data by sequentially choosing the most informative features at each node to make predictions. MPE, malignant pleural effusion, TPE, tuberculous pleural effusion, PPE, parapneumonic pleural effusion

causes, indicating that LDH serve as a distinguishing factor contributing to the MPE.

Discussion

As the development of the artificial intelligence, machine learning has taken a leading place in setting up the algorithms by the improvement through experience. Numerous studies have applied the machine learning as a tool to the early diagnosis of diseases and it showed a promising value in the identification of diseases [13]. Different aspects of medical data have been collected to calculate the machine learning models, including demographics, symptoms, medical history, laboratory tests, radiologic reports and images. Though machine leaning is good at processing high-dimensional data [36], it still faces challenges. When dealing with complex features and large datasets, large amounts of computational power were needed, especially when handling with medical images [37, 38]. As adding redundant or irrelevant features led to the overfitting and unnecessary computational cost, selection of the informative features is important [39]. In this article, we selected laboratory test and demographic characteristic as the tabular data and chose machine learning methods capable of handling this information for multi-class classification tasks.

LR is efficient for modeling linear relationships [40], but multicollinearity and outliers can reduce its performance and lead to biased results [41]. SVM constructs a hyperplane for classification and handles non-linear relationships well using kernel functions [42], but it has long training times and multiple parameters [43]. XGBoost combines decision trees for classification and regression, and it is known for its high robustness and ability to model non-linear relationships. But it is prone to overfitting [44]. RF aggregates decision trees through voting, offering strong resilience to outliers and high-dimensional data, but its interpretability is limited [45]. KNN classifies based on proximity to training samples, suitable for small datasets but requires large storage for large-scale data [46]. Tab Transformer captures relationships between categorical features using multi-head self-attention and non-linear transformations [47], which has not yet been applied in the diagnosis of pleural effusion.

In our studies, all of these six models act well in the diagnosis of MPE and TPE due to the sufficient sample size and the specific selection of the features to reduce the model complexity.

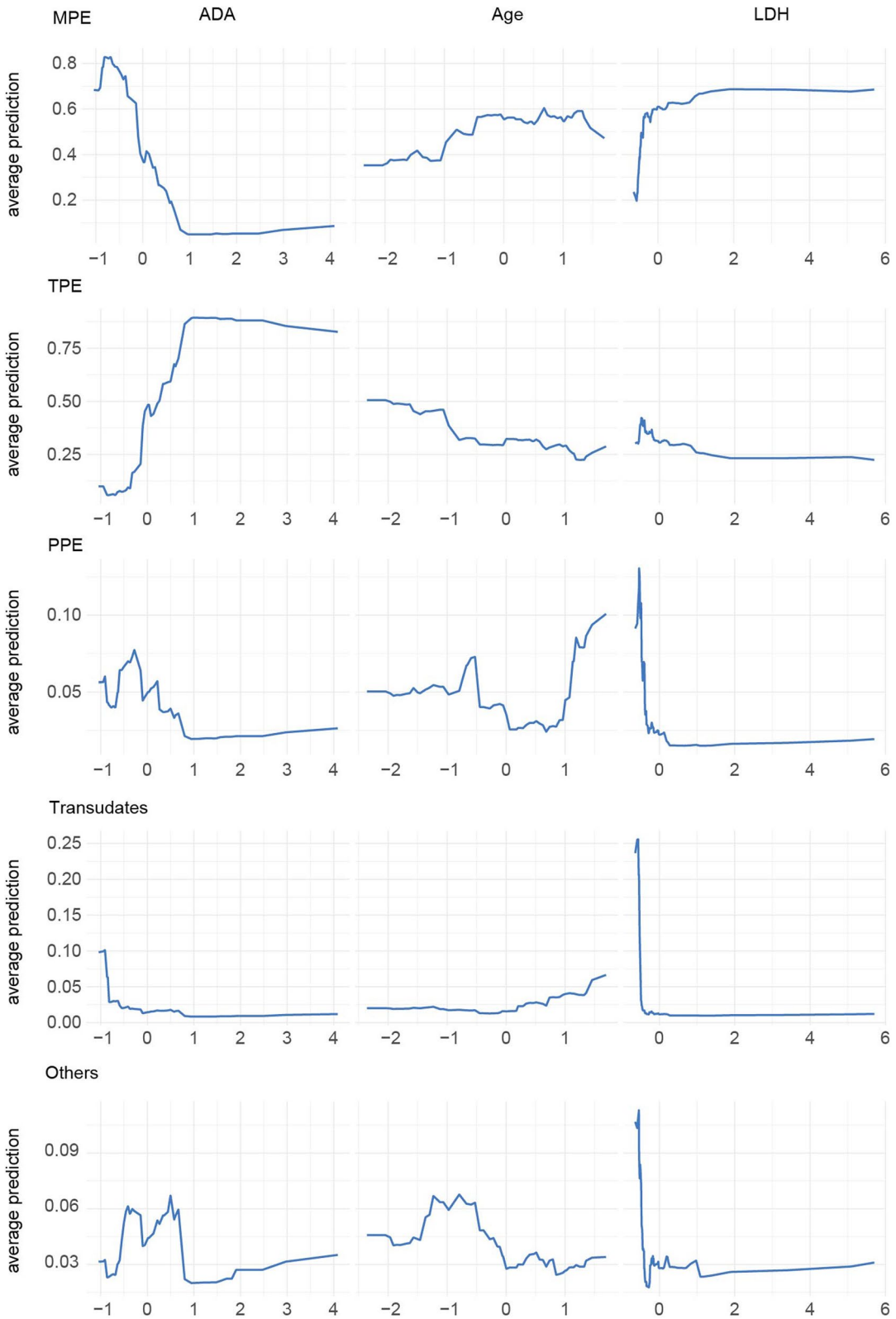

Fig. 7 Partial Dependence Plots of ADA, Age and LDH on pleural effusion etiologies prediction based on Extreme Gradient Boosting model. Each line depicts the single-variable effects on the prediction outcome. MPE, malignant pleural effusion, TPE, tuberculous pleural effusion, PPE, parapneumonic pleural effusion. ADA, Adenosine deaminase; LDH, lactate dehydrogenase

Pleural fluid ADA, pleural fluid LDH were common test in the diagnosis of the pleural effusion. Though 40U/L was a commonly used cut-off in the TPE diagnosis, the optimum cut-off remains controversial [35]. Our results showed that there was a negative correlation between pleural fluid ADA and age, which is consist with other study [48]. Age/ADA was reported as a promising diagnostic index for differentiating between TPE and MPE [12]. Thus, we included age as a feature in our study to assist the diagnosis. It is reported that pleural fluid LDH exhibits a weak positive correlation with pleural fluid ADA in the TPE, whereas in non-TPE cases the correlation is strong and positive [48]. However, in our study, there is a linear relationship between pleural fluid LDH and pleural ADA in the analysis without grouping by etiologies. Meanwhile, pleural fluid LDH is included in the Light's criteria in differentiating exudates from transudates. Pleural fluid LDH/ADA ratios could differentiate between TPE and PPE [49, 50]. Pleural effusion caused by the autoimmune disease have an elevated pleural fluid ADA and pleural fluid LDH [51]. Given all those studies and our results, we chose pleural fluid ADA, pleural fluid LDH and age as features.

In the diagnosis of MPE, machine learning models have both higher precision and recall compared to the traditional method based on cancer ratio. Given the pathological analysis is time-consuming, the enhanced performance achieved on the laboratory biomarkers and demographic characteristic is significant for reducing the diagnostic time. In the diagnosis of TPE, machine learning models have higher precision but lower recall than traditional method based simply on pleural ADA, which means the machine learning have stricter criteria. The clinical manifestations of tuberculosis may be added as considerable features for the diagnosis of TPE to elevated the recall of the models.

The visualization of the first tree in the random forest gave us a good model explanation. The splits of ADA and LDH reveals similar diagnosis patterns in the clinical decision-making. Patients with high pleural fluid ADA ($\geq$ 22U/L) and low pleural fluid LDH (< 353U/L) are indicative of TPE. Whereas patients with low pleural fluid ADA (< 22U/L) and high pleural effusion LDH ($\geq$ 123U/L) are more likely to have MPE. The patients with median levels of pleural fluid ADA and LDH were hard to classify, so age is a critical differentiator as a final split, with older patients ($\geq$ 59 years) have higher likelihood of MPE and younger patients (< 59 years) have higher likelihood of TPE. This strategy is similar in the clinical practice but the tree model gave us a specific split points with clear criteria and logical relationship.

The average prediction of the features in XGBoost indicated the potential contribution of the features to the predictions. The fluctuations for age suggest a complex relationship between age and the predicted outcome, indicating that age's impact may vary depending on the values of other features.

Many studies on pleural effusion diagnosis have shed light on the potential of the machine learning in promoting the diagnostic accuracy and clinical decision-making. To differentiate MPE, tumor biomarker [18], demographic characteristic, symptom, volume of the pleural effusion, site of the pleural effusion, blood routine test, pleural fluid routine and biochemical analyses [32], radiomic features [33], and radiomic features [15]. Machine learning has also been employed to investigate the pathological subtypes of the malignant pleural effusion in lung cancer [20], breast cancer [33] and malignant pleural mesothelioma [26]. To differentiate TPE, pleural fluid ADA as well as other features [19, 27, 28, 31]. have been utilized, with pleural fluid ADA identified as the most important feature in the model, which is consistent in our results. Machine learning has also been applied to multi-class classification for etiological diagnosis [4, 29], so as our study.

The patients included in our study had a higher proportion of TPE diagnosis (34.1%), compared to 9.2% [4] and 15.1% [29] in other studies, probably due to China's status as a country with high-burden tuberculosis. This finding highlights the potential feasibility of using only three features for diagnosis in resource-limited countries with a high tuberculosis burden, where the cost of laboratory tests should be carefully considered for clinical application.

Our study has following limitations: 1) The data were sourced from a single center which is a public hospital in a large city, which may result in differences in disease composition compared to primary care hospitals. 2) The number of cases of PPE, exudative effusion, and other types of pleural effusions was limited, which could impact the model's ability to accurately predict these uncommon categories. Further studies were needed to provide a more representative and diverse dataset to refine the predictive models.

Conclusions

By simply collecting the clinical parameters (age, pleural fluid ADA and pleural fluid LDH), machine leaning demonstrates strong performance in the etiological diagnosis of the pleural effusion, particularly for MPE, TPE, and transudative pleural effusion. This approach has the potential to serve as a valuable tool in assisting clinicians with identifying the underlying causes of pleural effusion.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-025-03253-2.

Additional file 1: Supplementary table 1. The hyperparameters of the models after Bayesian optimization are provided.

Additional file 2: Supplementary file 1. The necessary R packages, along with their dependencies and respective versions, are specified.

Additional file 3: Supplementary file 2. This file includes all the necessary R scripts for the analysis.

Additional file 4: Supplementary file 3. The necessary Python packages, along with their dependencies and respective versions, are listed.

Additional file 5: Supplementary file 4. This file contains all the necessary Python scripts for the analysis.

Authors' contributions

Q-Y. C, M-M. S designed the study and analyzed the data. F-S. Y and Q-Y. C drafted the manuscript. Q-Y. C and S-M. Y collected the data. F-S. Y and H-Z. S conceived the idea, supervised the research, and revised the manuscript. All authors read the manuscript and approved the final version for submission.

Funding

This work was supported by grants from Natural Science Foundation of Beijing Municipality (No. 7232066), National Natural Science Foundation of China (No. 82200111), the Beijing Scholars Program (No. 048) and Beijing Hospitals Authority Youth Program (QML20230303).

Data availability

The processed results are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Beijing Chao-Yang Hospital of Capital Medical University (2018-ke-321, 2024-ke-502).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 March 2025 Accepted: 21 April 2025
Published online: 02 May 2025

References

1. Jany B, Welte T. Pleural Effusion in Adults-Etiology, Diagnosis, and Treatment. Dtsch Arztebl Int. 2019;116(21):377–86.
2. Light RW. Clinical practice. Pleural effusion N Engl J Med. 2002;346(25):1971–7.
3. McGrath EE, Anderson PB. Diagnosis of pleural effusion: a systematic approach. Am J Crit Care. 2011;20(2):119–27 quiz 128.
4. Kim NY, et al. Differential Diagnosis of Pleural Effusion Using Machine Learning. Ann Am Thorac Soc. 2024;21(2):211–7.
5. Yu R, et al. Clinical diagnostic algorithm in defining tuberculous unilateral pleural effusion in high tuberculosis burden areas short of diagnostic tools. J Thorac Dis. 2022;14(4):866–76.
6. Porcel JM. Identifying transudates misclassified by Light's criteria. Curr Opin Pulm Med. 2013;19(4):362–7.
7. Park HJ, Choi CM. Can parapneumonic effusion be diagnosed only with pleural fluid analysis? J Thorac Dis. 2020;12(6):3422–5.
8. Metintas M, et al. Image-Assisted Pleural Needle Biopsy or Medical Thoracoscopy: Which Method for Which Patient? A Randomized Controlled Trial Chest. 2024;166(2):405–12.
9. Aggarwal AN, et al. Comparative accuracy of pleural fluid unstimulated interferon-gamma and adenosine deaminase for diagnosing pleural tuberculosis: A systematic review and meta-analysis. PLoS One. 2021;16(6):e0253525.
10. Chubb SP, Williams RA. Biochemical Analysis of Pleural Fluid and Ascites. Clin Biochem Rev. 2018;39(2):39–50.
11. Korczynski P, et al. Impact of age on the diagnostic yield of four different biomarkers of tuberculous pleural effusion. Tuberculosis (Edinb). 2019;114:24–9.
12. Zhou J, et al. Age : pleural fluid ADA ratio and other indicators for differentiating between tubercular and malignant pleural effusions. Medicine (Baltimore). 2022;101(26):e29788.
13. Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022;10(3):541.
14. Rajula HSR, et al. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina (Kaunas). 2020;56(9):455.
15. Ozcelik N, et al. Deep learning for diagnosis of malign pleural effusion on computed tomography images. Clinics (Sao Paulo). 2023;78:100210.
16. Wei TT, et al. Development and validation of a machine learning model for differential diagnosis of malignant pleural effusion using routine laboratory data. Ther Adv Respir Dis. 2023;17:17534666231208632.
17. Chen Z, et al. Machine learning applied to near-infrared spectra for clinical pleural effusion classification. Sci Rep. 2021;11(1):9411.
18. Zhang Y, et al. Diagnosis of malignant pleural effusion with combinations of multiple tumor markers: A comparison study of five machine learning models. Int J Biol Markers. 2023;38(2):139–46.
19. Ren Z, Hu Y, Xu L. Identifying tuberculous pleural effusion using artificial intelligence machine learning algorithms. Respir Res. 2019;20(1):220.
20. Perumal J, et al. Machine Learning Assisted Real-Time Label-Free SERS Diagnoses of Malignant Pleural Effusion due to Lung Cancer. Biosensors (Basel). 2022;12(11):940.
21. Wang J, et al. The Diagnosis of Malignant Pleural Effusion Using Tumor-Marker Combinations: A Cost-Effectiveness Analysis Based on a Stacking Model. Diagnostics (Basel). 2023;13(19):3136.
22. Widodo CE, Adi K, Gernowo R. A support vector machine approach for identification of pleural effusion. Heliyon. 2024;10(1):e22778.
23. Sexauer R, et al. Automated Detection, Segmentation, and Classification of Pleural Effusion From Computed Tomography Scans Using Machine Learning. Invest Radiol. 2022;57(8):552–9.
24. Liu Y, et al. Diagnostic and comparative performance for the prediction of tuberculous pleural effusion using machine learning algorithms. Int J Med Inform. 2024;182:105320.
25. Khemasuwan D, et al. Machine Learning Model Predictors of Intrapleural Tissue Plasminogen Activator and DNase Failure in Pleural Infection: A Multicenter Study. Ann Am Thorac Soc. 2025;22(2):187–92.
26. Li Y, et al. Differentiating malignant pleural mesothelioma and metastatic pleural disease based on a machine learning model with primary CT signs: A multicentre study. Heliyon. 2022;8(11):e11383.
27. Garcia-Zamalloa A, et al. Diagnostic accuracy of adenosine deaminase for pleural tuberculosis in a low prevalence setting: A machine learning approach within a 7-year prospective multi-center study. PLoS One. 2021;16(11):e0259203.
28. Li C, et al. Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. Comput Methods Programs Biomed. 2018;153:211–25.
29. Lee JH, et al. Classification of pleural effusions using deep learning visual models: contrastive-loss. Sci Rep. 2022;12(1):5532.
30. Liu J, Gallego B, Barbieri S. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. Sci Rep. 2022;12(1):1762.
31. Wu C, et al. The large language model diagnoses tuberculous pleural effusion in pleural effusion patients through clinical feature landscapes. Respir Res. 2025;26(1):52.
32. Li Y, et al. Driverless artificial intelligence framework for the identification of malignant pleural effusion. Transl Oncol. 2021;14(1):100896.

33. Cai F, et al. An Integrated Clinical and Computerized Tomography-Based Radiomic Feature Model to Separate Benign from Malignant Pleural Effusion. Respiration. 2024;103(7):406–16.
34. Verma A, Abisheganaden J, Light RW. Identifying Malignant Pleural Effusion by A Cancer Ratio (Serum LDH: Pleural Fluid ADA Ratio). Lung. 2016;194(1):147–53.
35. Aggarwal AN, et al. Adenosine deaminase for diagnosis of tuberculous pleural effusion: A systematic review and meta-analysis. PLoS One. 2019;14(3):e0213728.
36. Caballé-Cervigón N, et al. Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review. Applied Sciences. 2020;10(15):5135.
37. Litjens G, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
38. Rajpurkar P, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv, 2017. abs/1711.05225. https://arxiv.org/abs/1711.05225.
39. Ying X. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 2019:1168. https://doi.org/10.1088/1742-6596/1168/2/022022.
40. Schober P, Vetter TR. Logistic Regression in Medical Research. Anesth Analg. 2021;132(2):365–6.
41. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011;18(10):1099–104.
42. Xue H, Chen S, Yang Q. Structural regularized support vector machine: a framework for structural large margin classifier. IEEE Trans Neural Netw. 2011;22(4):573–87.
43. Chen Z, Li J, Wei L. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artif Intell Med. 2007;41(2):161–75.
44. Salehi F, et al. Machine Learning Prediction of Treatment Response to Biological Disease-Modifying Antirheumatic Drugs in Rheumatoid Arthritis. J Clin Med. 2024;13(13):3890.
45. Shamraeva MA, et al. The Application of a Random Forest Classifier to ToF-SIMS Imaging Data. J Am Soc Mass Spectrom. 2024;35(12):2801–14.
46. Hu LY, et al. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus. 2016;5(1):1304.
47. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. 2020. https://arxiv.org/abs/2012.06678.
48. Tay TR, Tee A. Factors affecting pleural fluid adenosine deaminase level and the implication on the diagnosis of tuberculous pleural effusion: a retrospective cohort study. BMC Infect Dis. 2013;13:546.
49. Wang J, et al. The pleural fluid lactate dehydrogenase/adenosine deaminase ratio differentiates between tuberculous and parapneumonic pleural effusions. BMC Pulm Med. 2017;17(1):168.
50. Nyanti LE, Rahim MAA, Huan NC. Diagnostic Accuracy of Lactate Dehydrogenase/Adenosine Deaminase Ratio in Differentiating Tuberculous and Parapneumonic Effusions: A Systematic Review. Tuberc Respir Dis (Seoul). 2024;87(1):91–9.
51. Lin L, et al. A retrospective study on the combined biomarkers and ratios in serum and pleural fluid to distinguish the multiple types of pleural effusion. BMC Pulm Med. 2021;21(1):95.

Publisher's Note