Respiratory Research

RESEARCH

Open Access

The large language model diagnoses tuberculous pleural effusion in pleural effusion patients through clinical feature landscapes

Chaoling Wu[1†], Wanyi Liu[2†], Pengfei Mei[3†], Yunyun Liu[1], Jian Cai[4], Lu Liu[1], Juan Wang[3], Xuefeng Ling[1], Mingxue Wang[1], Yuanyuan Cheng[1], Manbi He[1], Qin He[1], Qi He[1], Xiaoliang Yuan[5*] and Jianlin Tong[1*]

Abstract

Background Tuberculous pleural effusion (TPE) is a challenging extrapulmonary manifestation of tuberculosis, with traditional diagnostic methods often involving invasive surgery and being time-consuming. While various machine learning and statistical models have been proposed for TPE diagnosis, these methods are typically limited by complexities in data processing and difficulties in feature integration. Therefore, this study aims to develop a diagnostic model for TPE using ChatGPT-4, a large language model (LLM), and compare its performance with traditional logistic regression and machine learning models. By highlighting the advantages of LLMs in handling complex clinical data, identifying interrelationships between features, and improving diagnostic accuracy, this study seeks to provide a more efficient and precise solution for the early diagnosis of TPE.

Methods We conducted a cross-sectional study, collecting clinical data from 109 TPE and 54 non-TPE patients for analysis, selecting 73 features from over 600 initial variables. The performance of the LLM was compared with logistic regression and machine learning models (k-Nearest Neighbors, Random Forest, Support Vector Machines) using metrics like area under the curve (AUC), F1 score, sensitivity, and specificity.

Results The LLM showed comparable performance to machine learning models, outperforming logistic regression in sensitivity, specificity, and overall diagnostic accuracy. Key features such as adenosine deaminase (ADA) levels and monocyte percentage were effectively integrated into the model. We also developed a Python package (https://pypi.org/project/tpeai/) for rapid TPE diagnosis based on clinical data.

Conclusions The LLM-based model offers a non-surgical, accurate, and cost-effective method for early TPE diagnosis. The Python package provides a user-friendly tool for clinicians, with potential for broader use. Further validation in larger datasets is needed to optimize the model for clinical application.

Keywords Tuberculous pleural effusion, Large language model, ChatGPT-4, Artificial intelligence, Diagnosis model

†Chaoling Wu, Wanyi Liu and Pengfei Mei contributed equally to this work.

*Correspondence:
Xiaoliang Yuan
yxlyyxs@126.com
Jianlin Tong
tjl8880@163.com
Full list of author information is available at the end of the article

Introduction

Tuberculous pleural effusion (TPE) is a frequently encountered form of extrapulmonary tuberculosis, and its nonspecific clinical and imaging features present significant diagnostic challenges. Early and accurate diagnosis of TPE is critical for timely treatment, especially in regions with a high burden of tuberculosis. However, traditional diagnostic methods, such as pleural biopsy and pleural effusion (PE) analysis, often demonstrate limited sensitivity. This limitation underscores the need for more advanced diagnostic tools. While numerous studies have explored machine learning models for TPE diagnosis, the potential of large language models (LLMs) such as ChatGPT-4 has not yet been thoroughly investigated. This study aims to create a diagnostic model for TPE using ChatGPT-4 and compare its performance with traditional TPE diagnosis models based on logistic regression and machine learning methods. We also explore the performance differences between these approaches.

In many countries, TPE is a leading cause of PE and one of the most prevalent types of extrapulmonary tuberculosis, posing a prominent public health issue in developing countries, including China [1, 2]. TPE is caused by Mycobacterium tuberculosis infection of the pleura, characterized by a substantial accumulation of chronic effusion and inflammatory cells in the pleural cavity [3]. The combination of elevated lymphocyte count, exudative PE, and increased adenosine deaminase (ADA) levels is crucial for TPE diagnosis. However, in early cases, neutrophils may predominant [4], ADA levels may be relatively low [5], and the optimal pleural fluid ADA threshold for TPE diagnosis varies across studies [1, 6]. The gold standard for diagnosing TPE is detecting Mycobacterium tuberculosis in PE or pleural biopsy specimens [1]. However, pleural fluid microbiological cultures have low positivity rates and are time-consuming, often requiring up to eight weeks. Additionally, obtaining pleural specimens via thoracoscopy or percutaneous pleural biopsy involves a surgical procedure, which poses substantial trauma and risks of complications, such as iatrogenic pneumothorax [7]. Therefore, diagnosing TPE remains challenging. This highlights the critical need for a less invasive, more accurate, and cost-effective method for early TPE diagnosis.

Recently, the use of artificial intelligence (AI) in healthcare has been gradually expanding. Machine learning, a subset of AI, creates algorithms that utilize large and complex datasets. This enables computers to exhibit intelligent behavior [8]. Machine learning algorithms (MLAs), such as k-Nearest Neighbors (KNN), Random Forests (RF), and Support Vector Machines (SVM), can build efficient, objective, and accurate disease diagnosis models. Machine learning has shown broad potential for clinical diagnosis [9]. Zhou et al. proposed a new algorithm, CFDE, for feature selection in the clinical feature analysis of TPE. This algorithm demonstrated significant advantages in global optimization and feature selection. When combined with the SVM model, it effectively identified key clinical indicators associated with TPE, supporting early diagnosis and treatment of TPE [10]. Ren et al. explored diagnostic biomarkers for TPE and incorporated patient clinical features into MLAs, including logistic regression, SVM, RF, and KNN. The results showed that RF achieved an area under the curve (AUC) value of 0.97, significantly higher than the AUC of pleural effusion ADA (0.89) [11]. Li et al. developed a new model called bGACO-SVM to classify TPE from non-TPE. The results showed that this model differed from classical MLAs [12]. Additionally, Li et al. combined a new algorithm, FS-MFO-SVM, with feature selection for diagnosing TPE. This approach demonstrated an average accuracy of 95%, an AUC of 0.9564, sensitivity of 93.35%, and specificity of 97.57% [13]. Despite these advancements, machine learning-based methods still face challenges in effectively integrating and analyzing complex, multi-dimensional clinical data, especially when dealing with high variability data.

LLMs are AI systems based on deep learning [14, 15]. By learning from vast amounts of data, they can analyze complex clinical information and provide medical diagnostic suggestions [16–20]. Significant progress has been made in applying LLMs to disease diagnosis and treatment. Studies have shown that LLMs, such as ChatGPT, can assist clinicians quickly access and summarize large volumes of medical literature. This enables them to stay updated on recent studies about rare diseases and facilitates more precise diagnosis [21]. Tassallah et al. evaluated the performance of three LLMs—ChatGPT 3.5, ChatGPT-4, and Google Bard—in diagnosing conditions such as chylous tuberculosis and primary adrenal cortical insufficiency. The results showed that these models outperformed the average diagnostic accuracy of physicians [22]. Zheng et al. pointed out that ChatGPT excelled in assisting the diagnosis of diseases like primary pulmonary arterial hypertension and Parkinson's disease with an early onset. It demonstrated the ability to quickly analyze medical literature and patient data while formulating personalized treatment plans [23]. Hu et al. evaluated ChatGPT-4's ability to diagnose rare eye diseases in different scenarios. The results showed that ChatGPT-4 helped primary care ophthalmologists diagnose rare eye conditions more quickly and accurately [24]. Additionally, Carlo et al. assessed the performance of various AI LLMs (ChatGPT 3.5, ChatGPT-4, Bing Chat, Google Bard, and Claude) in answering medical questions about diseases such as thymoma and Good's syndrome. The results showed that ChatGPT-4 and Bard outperformed

others in terms of information accuracy, responsiveness, and clinical applicability [25]. These studies demonstrate that LLMs offer superior efficiency compared to traditional methods and may also provide advantages in diagnostic accuracy. However, while these LLM models have shown promise in various clinical scenarios, their application to TPE diagnosis remains unexplored.

This study aims to bridge this gap by developing a diagnostic model for TPE using the LLM. We compare its performance with traditional diagnostic approaches, including logistic regression and various MLAs, to evaluate its ability to diagnose TPE. The results show that LLMs, particularly ChatGPT-4, excel at integrating clinical data and identifying potential relationships between complex features, offering new insights and support for the early diagnosis of TPE. Furthermore, we developed and published a ChatGPT-4-based diagnostic LLM software package for distinguishing between TPE and non-TPE, making it accessible for clinical use. Future refinement of this tool could significantly enhance diagnostic accuracy and efficiency, ultimately facilitating earlier diagnosis and more personalized treatment of TPE.

Materials and methods

Patients and study design

This study included 38,885 hospitalized patients from January 2011 to June 2024 at the Affiliated Hospital of Jiujiang University. A cross-sectional study was conducted. Patients were eligible for enrollment if they met the following criteria: (1) a diagnosis of pleural effusion (PE) confirmed by ultrasound, chest computed tomography (CT), or X-ray; (2) a diagnosis of PE confirmed by pleural biopsy. The exclusion criteria were: (1) patients who had undergone anti-tuberculosis treatment prior to admission; (2) pregnant women; (3) patients with incomplete clinical data (more than 20% missing); (4) patients with an unknown cause of PE. All patients included in the study were newly diagnosed with PE and had not received any prior treatment. We collected relevant demographic, laboratory, and clinical information from the hospital's clinical electronic records system. In total, 163 patients were included in the final analysis. Among them, 109 had TPE, and 54 had non-TPE. Initially, over 600 clinical features were screened for the 163 patients. Variables with more than 20% missing data were excluded, leaving 73 variables for analysis. Differences between variables were visualized using the ggplot2 package. The patient selection process and study flowchart are shown in Fig. 1.

Diagnostic criteria for TPE

The diagnosis of TPE is based on one of the following criteria: (a) the detection of Mycobacterium tuberculosis in pleural fluid or pleural tissue culture; (b) histological examination showing granulomatous inflammation in pleural biopsy, with Mycobacterium tuberculosis isolated from another site; or (c) histological examination showing granulomatous inflammation in pleural biopsy, with clinical response to anti-tuberculosis therapy [26].

Data collection and variable selection

We selected candidate variables based on key literature on TPE diagnosis models and our clinical experience. These variables were chosen for their clinical availability and non-surgical nature. The potential diagnostic variables included the following clinical characteristics: age, sex, routine PE parameters (color, turbidity, specific gravity, and Leifant test), biochemical parameters of PE (total protein, glucose, lactate dehydrogenase (LDH), adenosine deaminase (ADA), and albumin), serum biochemical parameters (C-reactive protein (CRP), erythrocyte sedimentation rate (ESR)), complete blood count [(white blood cells (WBC), lymphocytes, neutrophils], and tumor markers [carcinoembryonic antigen (CEA), non-small cell lung cancer-related antigen, neuron-specific enolase (NSE)], among others. Samples were sourced from peripheral blood, PE, or pleural tissue collected during hospitalization. We collected clinical data from eligible patients using a structured data sheet customized for this study. These clinical data were obtained from the patients' discharge records. Two experienced

(See figure on next page.)

Fig. 1 Patient selection and study flowchart. The flowchart illustrates the steps involved in developing a diagnostic model for tuberculous pleural effusion (TPE) using various modeling approaches. Data was collected from 38,885 patients admitted to the Department of Respiratory Medicine at Ganjiang University Affiliated Hospital between January 2011 and June 2024. A total of 245 patients with pleural effusion were included after excluding 38,640 patients without pleural effusion. Among these, 163 patients underwent thoracoscopic biopsy and were included in the study. Patients were further divided into two groups: the TPE group (n = 109), diagnosed with TPE, and the non-TPE group (n = 54), diagnosed with other types of pleural effusion (PE). The data was divided into training (n = 115) and test sets (n = 48) for model development and evaluation. Lasso regression was applied for variable selection. Three model types were developed: H_2O AutoML, including XGBoost, GBM, GLM, XRT, ensemble stacking, and deep learning algorithms; traditional logistic regression models using forward, backward, and bidirectional stepwise regression; and ChatGPT-based large language models (LLMs), including ChatGPT-4o and ChatGPT-4. The diagnostic performance of these models was compared with eight previously published TPE diagnosis models to assess their accuracy and effectiveness. The final goal was to evaluate and compare the diagnostic performance of machine learning, traditional logistic regression, and LLMs in diagnosing TPE

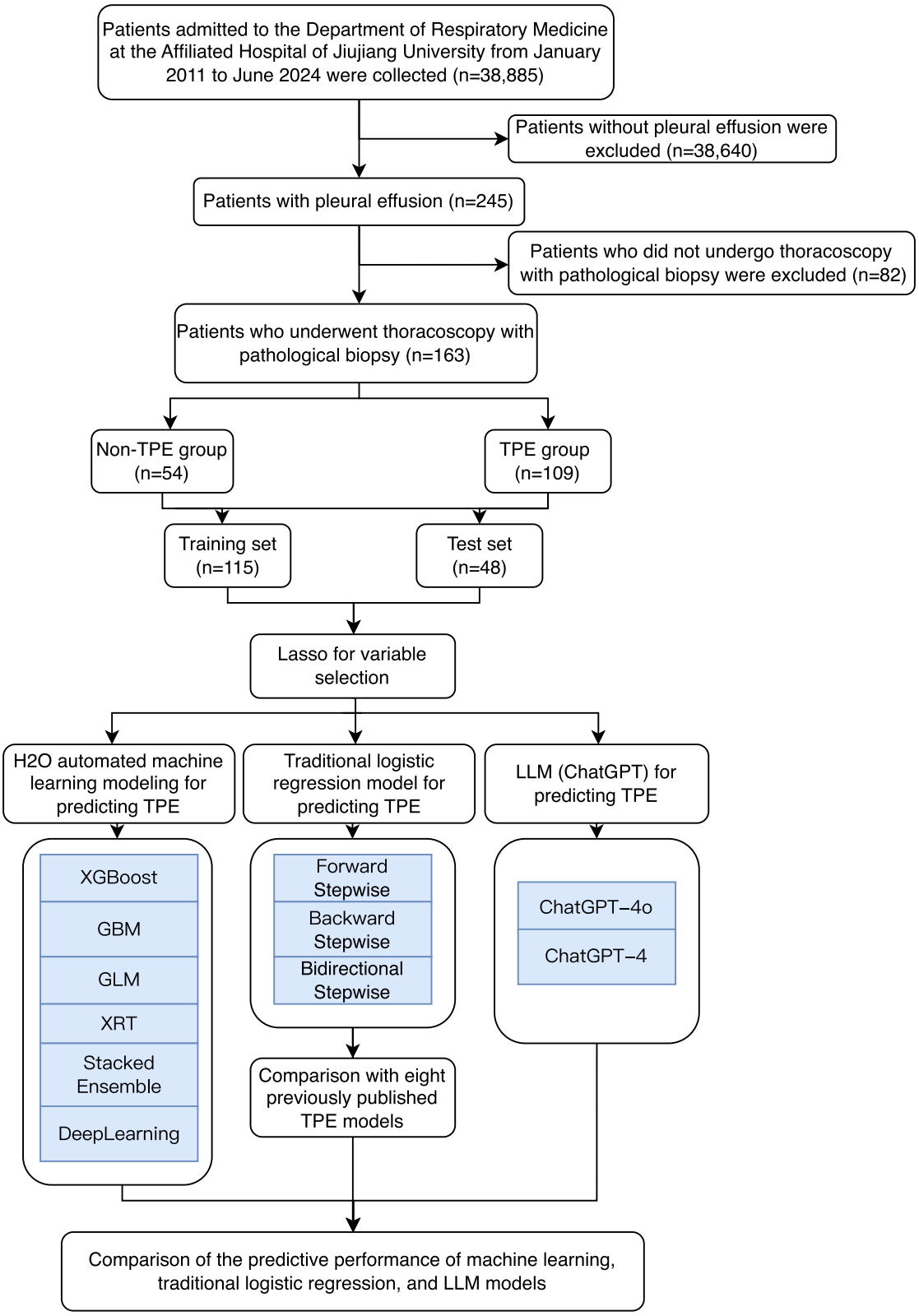

Fig. 1 (See legend on previous page.)

pulmonologists reviewed, refined, and cross-checked the clinical data. All data were collected by research staff who were blinded to the final outcome measurements.

Data preprocessing and feature selection

The cohort data used in this study contained missing values. Deleting all incomplete data could reduce the sample size, compromise data quality, and affect diagnostic results. Therefore, we excluded data with more than 20% missing values. For data with missing values $\leq 20\%$, we applied different imputation methods depending on the data type. We used the "norm" method for continuous data, "logreg" for binary classification data, and "polyreg" for multiclass data. These imputation methods were implemented using the "mice" package in R. All continuous variables were converted into binary variables, with the optimal classification threshold determined by receiver operating characteristic (ROC) curve analysis. We used the coords function from the pROC package to select the optimal threshold, which is determined based on the trade-off between sensitivity and specificity. This method finds the balance point between sensitivity (maximizing the identification of positive samples) and specificity (minimizing false negatives), thereby optimizing the classification performance (detailed in Code Sect. 1 of Supplementary Materials).

We then clustered the data using partial least squares discriminant analysis (PLS-DA). Variables with variable importance projection (VIP) values greater than 0.5 were extracted, resulting in 73 selected variables. We removed variables with area under the curve (AUC) values less than 0.6, leaving 17 variables. Next, we removed highly correlated variables with pairwise correlations greater than 0.5. We assessed multicollinearity using the variance inflation factor (VIF). Variables with a VIF value less than 5 were retained, reducing the number of variables to 16. Finally, we used least absolute shrinkage and selection operator (LASSO) regression, conducted with the "glmnet" package in R, to select 12 variables for the final analysis. Using the "caret" package, we randomly split the patients into a training and test set in a 7:3 ratio.

Establishment and evaluation of the TPE machine learning diagnosis model

In this study, we used H2O AutoML to integrate a series of classical and advanced machine learning algorithms to effectively diagnose TPE. The algorithms employed included "XGBoost", "GBM", "GLM", "XRT", "DeepLearning", and "StackedEnsemble" (detailed in Supplementary Materials). We comprehensively evaluated these algorithms to select the optimal model for disease diagnosis.

Hyperparameter tuning and cross-validation

To optimize model performance and prevent overfitting, we used fivefold cross-validation. This method splits the training dataset into five mutually exclusive subsets. Each subset serves as the validation set while the remaining four subsets are used for training. Additionally, the AutoML process automatically performs hyperparameter tuning to explore the best model configuration.

Model selection and evaluation

We set the number of automatically generated models to 1000, and successfully generated 453 models. These models underwent hyperparameter tuning and fivefold cross-validation. We implemented an early stopping mechanism, using AUC as the performance evaluation metric. The training process was automatically halted if the AUC improvement was less than 0.001 for three consecutive training cycles, preventing overfitting. We also filtered out models with an AUC of 1, as this could indicate overfitting. The final model was selected based on the largest average AUC value from both the training and validation sets. This ensured optimal diagnostic performance and generalizability.

Model performance evaluation and diagnostic interpretation

We evaluated model performance using ROC curves, F1 score, and SHAP (R package) analysis. First, we used the trained model to make diagnosis on the test set (testdata) and extracted the diagnostic probabilities for the positive class (class 1).We then used the pROC package to generate the ROC curve and calculate the model's AUC with its 95% confidence interval. The F1 score on the test set was calculated using the confusionMatrix function. We also visualized the confusion matrix and saved it. Finally, we performed SHAP value analysis using the shapviz package to interpret the impact of each variable on the model's diagnosis. This analysis helped explain the model's diagnosis of TPE likelihood in patients.

Establishment of the traditional logistic regression model and comparison with previous models

Logistic regression model construction and variable selection

We constructed a logistic regression model to fit the data for effective TPE diagnosis. First, we built a full-variable logistic regression model using variables selected by LASSO in the training dataset. We applied three different variable selection strategies: forward selection, backward elimination, and stepwise regression. Forward selection adds variables progressively based on the minimum AIC value. Backward elimination removes non-significant variables step by step. Stepwise regression combines

both strategies. During the variable selection process, we chose the logistic regression model from forward selection with the highest AUC for subsequent analysis.

Model evaluation and performance visualization

ROC curve and AUC calculation

We evaluated model performance using the ROC curve and calculated the AUC to quantify classification performance. The pROC package generated ROC curves for both the training and test sets. We recorded the AUC values and their 95% confidence intervals (CI). To assess the model's reliability, we applied a bootstrap method with 1000 resamples. This produced multiple ROC curves to evaluate the model's stability.

Decision curve analysis (DCA)

We performed DCA to assess the clinical applicability of the model at different thresholds. DCA evaluates the net benefit at various decision thresholds, helping us determine the model's practical significance in specific clinical scenarios.

Variable importance and forest plot visualization

We used a forest plot to visually display each variable's contribution to the model's diagnosis. The forestplot package created the plot, displaying the importance of variables through their respective odds ratios (OR). The visual results also included the confidence intervals and significance levels for the variables.

Nomogram and individualized diagnosis

We used a nomogram to show how the model can be used for individualized diagnosis. The nomogram converts each variable into a scoring system to diagnose TPE in an individual. This approach enhances the interpretability and practical application of the model.

Comparison with published TPE diagnosis model performance

We collected the variables from eight previously published TPE diagnosis models and applied them to our training dataset for modeling and diagnosis on the test dataset. We compared the AUC values of the different models to evaluate and determine the classification performance of each model.

Establishment and evaluation of the large language model (LLM) (ChatGPT-4) diagnosis model for TPE

To further explore the application of LLM (ChatGPT-4) in diagnosing TPE, we employed the following methods:

Variable importance scoring

First, we used the variable set selected by LASSO regression and assigned an importance score to each variable using ChatGPT-4. The scores ranged from 1 to 10. This process was repeated 10 times, and we calculated the mean score for each variable across all 10 iterations. We ranked the variables in descending order based on their mean importance scores. Variables with an average importance score greater than 5 were selected. A total of 8 key variables were identified: a. Biochemical parameters of PE: ADA, total protein, albumin; b. Blood cell analysis parameters: Lymphocyte count, neutrophil percentage, monocyte percentage, neutrophil count; c. Patient age.

Model training and diagnosis

We input the 8 key variables into ChatGPT-4 to train the model, ensuring it accurately learned and understood the data features. After training, we used the model to diagnose the TPE.

Model evaluation

We evaluated the model's performance using the ROC curve and the F1 score. By calculating the AUC and F1 score, we quantified the model's classification performance. Finally, we compared the diagnosis results of the ChatGPT-4 with those of previous best-performing machine learning models (such as Support Vector Machines (SVM), Random Forests (RF)) and the traditional logistic regression model. This comparison helped us assess its superiority or limitations.

Development of the ChatGPT-4 diagnostic model python package for TPE

We developed a Python package named tepai (https://pypi.org/project/tpeai/) to quickly differentiate between TPE and non-TPE using the diagnostic power of ChatGPT-4. By inputting a set of key variables related to the patient's biochemistry and blood cell analysis, the model provides an intelligent diagnosis. The required variables include: pleural fluid biochemistry (ADA, total protein, albumin), blood cell analysis (lymphocyte count, neutrophil percentage, monocyte percentage, neutrophil count), and patient age. The model uses these inputs to generate a diagnosis through ChatGPT-4, assisting clinicians quickly identify the type of PE and make informed diagnostic and treatment decisions.

Statistical methods

Statistical analyses and software development for this study were performed using R 4.2.3 and Python 3.10.

We first tested continuous variables for normality. Data that followed a normal distribution are presented as mean ± standard deviation (SD). We used the independent samples t-test for pairwise comparisons and ANOVA for multiple group comparisons. For non-normally distributed data, the median and interquartile range [P25, P75] are presented. Group comparisons were made using the Mann–Whitney U test for two groups and the Kruskal–Wallis test for multiple groups. Categorical data are expressed as frequencies and percentages (%), with the Chi-square (χ^2) test used to compare rates between groups. We set a significance level of $\alpha = 0.05$ (two-tailed). A P-value of less than 0.05 was deemed statistically significant.

Results

Clinical characteristics of TPE

We analyzed 73 clinical variables from 163 patients (including 109 TPE patients and 54 non-TPE patients) who underwent thoracoscopic biopsy (Fig. 1). A baseline table of clinical characteristics for the TPE and non-TPE groups was generated using the tableone package (Table 1). We classified non-TPE and TPE samples using clinical characteristics data based on routine blood tests and biochemical markers, with PLS-DA (Fig. 2A). The results showed that the first two principal components (PC1 and PC2) explained 6.83% and 3.44% of the variance, respectively, and the two groups exhibited a clear separation on the score plot. This indicates that clinical features have strong discriminative power in distinguishing non-TPE group from TPE group. To further explore the expression patterns of these clinical features across different samples, we used a heatmap to display the expression levels of various biochemical and hematological variables (Fig. 2B). The results indicated significant differences in the expression levels of these variables between non-TPE and TPE groups. The cut-off values and area under the curves (AUCs) of the clinical characteristics are shown in Supplementary Table 1. We illustrated the distribution of key biomarkers in the non-TPE and TPE groups (Fig. 2C). The findings revealed significant intergroup differences ($P < 0.05$) in the levels of adenosine deaminase (ADA) in pleural effusion (PE), total protein distribution, and the percentage of monocytes in the blood count. These biomarkers may hold potential diagnostic value for TPE. Furthermore, we assessed the diagnostic performance of these biomarkers using ROC curve analysis (Fig. 2D). The AUC for ADA in PE was 0.8488 (95% CI: 0.7696–0.928), the AUC for the percentage of monocytes in the blood count was 0.7645 (95% CI: 0.6903–0.8387), and the AUC for total protein in PE was 0.7202 (95% CI: 0.6332–0.8072). These results indicate that ADA levels in PE, total protein distribution, and monocyte percentage in the blood count have high diagnostic accuracy and can effectively differentiate between TPE and non-TPE groups.

Machine learning modeling effectively diagnose TPE

We further refined the variable selection by excluding variables with AUC values less than 0.6 and those with pairwise correlations greater than 0.5. We also assessed multicollinearity using the variance inflation factor (VIF) and retained variables with VIF values less than 5 (Supplementary Table 2, Supplementary Fig. 1). This process resulted in 16 selected variables (detailed in the Methods section). Subsequently, we performed diagnostic analysis for TPE using a machine learning model. We visualized the coefficient paths of different biochemical and hematological variables using the LASSO regression model (Fig. 3A). As the regularization parameter λ increased, the coefficients of the variables gradually shrank toward zero, indicating that the influence of some variables on the model's diagnosis was reduced during regularization. We then examined the model's performance at different λ values using a validation curve (Fig. 3B). The Mean Squared Error initially decreased and then increased with λ, suggesting that an optimal λ value corresponding to the best model complexity that could maintain high accuracy while avoiding overfitting. We selected 12 variables for further modeling. Using the H2O automated machine learning platform, we created 453 models with six algorithms: "XGBoost", "GBM", "GLM", "XRT", "DeepLearning", "StackedEnsemble". We ranked the top 93 models based on their average AUC values from both the training and test sets (Fig. 3C, Supplementary Tables 3–6). The results indicate that most models performed excellently on these two metrics. Additionally, we used the Gain method to reflect the importance of each variable in the optimal XGBoost model (Fig. 3D). The variable importance ranking shows that biochemical markers in PE, such as albumin and ADA, possess the strongest diagnostic ability in distinguishing between the TPE and non-TPE groups. To further explain the diagnostic mechanism of the model, we used SHAP to analyze the contribution of each feature to the sample outcomes (Fig. 3E). The analysis showed that features like PFB ADA and PFB albumin significantly impacted the model's diagnosis. In contrast, features such as age and lipid profile triglycerides contributed less. The results showed that "albumin = 0" in PE had the strongest negative impact on the model output, with a SHAP value of − 1.34, while "ADA = 1" in PE had a significant positive impact (SHAP value of + 1.29). Additionally, the SHAP force plot demonstrated the decomposition of multiple features' contributions to the diagnostic result for a specific sample (Fig. 3F). In this sample, the positive contributions of

Table 1 Baseline clinical characteristics of TPE and non-TPE patients

	Level	Overall	Non-TPE	TPE	p value
n		163	54	109	
GROUP (%)	0	54 (33.13)	54 (100.00)	0 (0.00)	< 0.0001
	1	109 (66.87)	0 (0.00)	109 (100.00)	
Gender (%)	0	39 (23.93)	15 (27.78)	24 (22.02)	0.5378
	1	124 (76.07)	39 (72.22)	85 (77.98)	
COHORT (mean (SD))		2018.859 (3.222)	2018.944 (3.123)	2018.817 (3.283)	0.8123
Age (%)	0	28 (17.18)	1 (1.85)	27 (24.77)	0.0006
	1	135 (82.82)	53 (98.15)	82 (75.23)	
CRP (%)	0	36 (22.09)	19 (35.19)	17 (15.60)	0.0084
	1	127 (77.91)	35 (64.81)	92 (84.40)	
D-dimer (%)	0	81 (49.69)	32 (59.26)	49 (44.95)	0.1205
	1	82 (50.31)	22 (40.74)	60 (55.05)	
Electrolytes_Chloride (%)	0	118 (72.39)	33 (61.11)	85 (77.98)	0.0374
	1	45 (27.61)	21 (38.89)	24 (22.02)	
Electrolytes_Osmolality (%)	0	81 (49.69)	20 (37.04)	61 (55.96)	0.035
	1	82 (50.31)	34 (62.96)	48 (44.04)	
Electrolytes_Phosphorus (%)	0	92 (56.44)	34 (62.96)	58 (53.21)	0.3106
	1	71 (43.56)	20 (37.04)	51 (46.79)	
Electrolytes_Calcium (%)	0	99 (60.74)	30 (55.56)	69 (63.30)	0.4337
	1	64 (39.26)	24 (44.44)	40 (36.70)	
Electrolytes_Natrium (%)	0	43 (26.38)	7 (12.96)	36 (33.03)	0.0109
	1	120 (73.62)	47 (87.04)	73 (66.97)	
Electrolytes_Potassium (%)	0	52 (31.90)	22 (40.74)	30 (27.52)	0.1271
	1	111 (68.10)	32 (59.26)	79 (72.48)	
Electrolytes_Magnesium (%)	0	32 (19.63)	16 (29.63)	16 (14.68)	0.0401
	1	131 (80.37)	38 (70.37)	93 (85.32)	
Electrolytes_Anion gap (%)	0	74 (45.40)	20 (37.04)	54 (49.54)	0.1796
	1	89 (54.60)	34 (62.96)	55 (50.46)	
CEA (%)	0	75 (46.01)	16 (29.63)	59 (54.13)	0.0053
	1	88 (53.99)	38 (70.37)	50 (45.87)	
NSE (%)	0	35 (21.47)	18 (33.33)	17 (15.60)	0.0167
	1	128 (78.53)	36 (66.67)	92 (84.40)	
Liver function_GGT (%)	0	119 (73.01)	34 (62.96)	85 (77.98)	0.065
	1	44 (26.99)	20 (37.04)	24 (22.02)	
Liver function_Carbon dioxide (%)	0	62 (38.04)	24 (44.44)	38 (34.86)	0.3103
	1	101 (61.96)	30 (55.56)	71 (65.14)	
Liver function_Total bile acids (%)	0	78 (47.85)	16 (29.63)	62 (56.88)	0.0019
	1	85 (52.15)	38 (70.37)	47 (43.12)	
Liver function_Total bilirubin (%)	0	43 (26.38)	17 (31.48)	26 (23.85)	0.3946
	1	120 (73.62)	37 (68.52)	83 (76.15)	
Liver function_Total protein (%)	0	145 (88.96)	53 (98.15)	92 (84.40)	0.0178
	1	18 (11.04)	1 (1.85)	17 (15.60)	
Liver function_Globulin (%)	0	98 (60.12)	37 (68.52)	61 (55.96)	0.1704
	1	65 (39.88)	17 (31.48)	48 (44.04)	
Liver function_Albumin/Globulin ratio (%)	0	82 (50.31)	30 (55.56)	52 (47.71)	0.4372
	1	81 (49.69)	24 (44.44)	57 (52.29)	
Liver function_Albumin (%)	0	75 (46.01)	31 (57.41)	44 (40.37)	0.0591
	1	88 (53.99)	23 (42.59)	65 (59.63)	
Liver function_Direct bilirubin (%)	0	138 (84.66)	39 (72.22)	99 (90.83)	0.0041

Table 1 (continued)

	Level	Overall	Non-TPE	TPE	p value
	1	25 (15.34)	15 (27.78)	10 (9.17)	
Liver function_Alkaline phosphatase (%)	0	67 (41.10)	14 (25.93)	53 (48.62)	0.0092
	1	96 (58.90)	40 (74.07)	56 (51.38)	
Liver function_ALT (%)	0	135 (82.82)	48 (88.89)	87 (79.82)	0.2207
	1	28 (17.18)	6 (11.11)	22 (20.18)	
Liver function_AST/ALT ratio (%)	0	51 (31.29)	12 (22.22)	39 (35.78)	0.1147
	1	112 (68.71)	42 (77.78)	70 (64.22)	
Liver function_AST (%)	0	45 (27.61)	20 (37.04)	25 (22.94)	0.0874
	1	118 (72.39)	34 (62.96)	84 (77.06)	
Liver function_Indirect bilirubin (%)	0	81 (49.69)	23 (42.59)	58 (53.21)	0.2671
	1	82 (50.31)	31 (57.41)	51 (46.79)	
Liver function_BUN (%)	0	30 (18.40)	12 (22.22)	18 (16.51)	0.5026
	1	133 (81.60)	42 (77.78)	91 (83.49)	
Liver function_Uric acid (%)	0	18 (11.04)	4 (7.41)	14 (12.84)	0.4372
	1	145 (88.96)	50 (92.59)	95 (87.16)	
Liver function_Creatinine (%)	0	29 (17.79)	14 (25.93)	15 (13.76)	0.0903
	1	134 (82.21)	40 (74.07)	94 (86.24)	
Liver function_LDH (%)	0	37 (22.70)	17 (31.48)	20 (18.35)	0.0919
	1	126 (77.30)	37 (68.52)	89 (81.65)	
Pleural fluid biochemistry_Total protein (%)	0	61 (37.42)	35 (64.81)	26 (23.85)	< 0.0001
	1	102 (62.58)	19 (35.19)	83 (76.15)	
Pleural fluid biochemistry_Albumin (%)	0	45 (27.61)	26 (48.15)	19 (17.43)	0.0001
	1	118 (72.39)	28 (51.85)	90 (82.57)	
Pleural fluid biochemistry_ADA (%)	0	55 (33.74)	42 (77.78)	13 (11.93)	< 0.0001
	1	108 (66.26)	12 (22.22)	96 (88.07)	
Pleural fluid biochemistry_Glucose (%)	0	100 (61.35)	25 (46.30)	75 (68.81)	0.0091
	1	63 (38.65)	29 (53.70)	34 (31.19)	
Coagulation profile_Prothrombin time (%)	0	27 (16.56)	14 (25.93)	13 (11.93)	0.0414
	1	136 (83.44)	40 (74.07)	96 (88.07)	
Coagulation profile_Thrombin time (%)	0	71 (43.56)	30 (55.56)	41 (37.61)	0.0448
	1	92 (56.44)	24 (44.44)	68 (62.39)	
Coagulation profile_INR (%)	0	48 (29.45)	23 (42.59)	25 (22.94)	0.016
	1	115 (70.55)	31 (57.41)	84 (77.06)	
Coagulation profile_APTT (%)	0	82 (50.31)	32 (59.26)	50 (45.87)	0.1491
	1	81 (49.69)	22 (40.74)	59 (54.13)	
Coagulation profile_Fibrinogen (%)	0	92 (56.44)	27 (50.00)	65 (59.63)	0.3175
	1	71 (43.56)	27 (50.00)	44 (40.37)	
ESR (%)	0	43 (26.38)	21 (38.89)	22 (20.18)	0.0182
	1	120 (73.62)	33 (61.11)	87 (79.82)	
Blood cell analysis_Neutrophil percentage (%)	0	129 (79.14)	34 (62.96)	95 (87.16)	0.0007
	1	34 (20.86)	20 (37.04)	14 (12.84)	
Blood cell analysis_Neutrophil count (%)	0	101 (61.96)	20 (37.04)	81 (74.31)	< 0.0001
	1	62 (38.04)	34 (62.96)	28 (25.69)	
Blood cell analysis_Monocyte percentage (%)	0	50 (30.67)	31 (57.41)	19 (17.43)	< 0.0001
	1	113 (69.33)	23 (42.59)	90 (82.57)	
Blood cell analysis_Monocyte count (%)	0	136 (83.44)	40 (74.07)	96 (88.07)	0.0414
	1	27 (16.56)	14 (25.93)	13 (11.93)	
Blood cell analysis_Basophil percentage (%)	0	144 (88.34)	49 (90.74)	95 (87.16)	0.6803
	1	19 (11.66)	5 (9.26)	14 (12.84)	

Table 1 (continued)

	Level	Overall	Non-TPE	TPE	p value
Blood cell analysis_Basophil count (%)	0	82 (50.31)	24 (44.44)	58 (53.21)	0.375
	1	81 (49.69)	30 (55.56)	51 (46.79)	
Blood cell analysis_Eosinophil percentage (%)	0	134 (82.21)	37 (68.52)	97 (88.99)	0.0027
	1	29 (17.79)	17 (31.48)	12 (11.01)	
Blood cell analysis_Eosinophil count (%)	0	107 (65.64)	24 (44.44)	83 (76.15)	0.0001
	1	56 (34.36)	30 (55.56)	26 (23.85)	
Blood cell analysis_Large platelet ratio (%)	0	140 (85.89)	42 (77.78)	98 (89.91)	0.0636
	1	23 (14.11)	12 (22.22)	11 (10.09)	
Blood cell analysis_Red cell distribution width-SD (%)	0	64 (39.26)	13 (24.07)	51 (46.79)	0.0087
	1	99 (60.74)	41 (75.93)	58 (53.21)	
Blood cell analysis_Immature granulocyte percentage (%)	0	23 (14.11)	9 (16.67)	14 (12.84)	0.6739
	1	140 (85.89)	45 (83.33)	95 (87.16)	
Blood cell analysis_Immature granulocyte count (%)	0	145 (88.96)	44 (81.48)	101 (92.66)	0.0604
	1	18 (11.04)	10 (18.52)	8 (7.34)	
Blood cell analysis_Lymphocyte percentage (%)	0	51 (31.29)	24 (44.44)	27 (24.77)	0.0178
	1	112 (68.71)	30 (55.56)	82 (75.23)	
Blood cell analysis_Lymphocyte count (%)	0	107 (65.64)	26 (48.15)	81 (74.31)	0.0017
	1	56 (34.36)	28 (51.85)	28 (25.69)	
Blood cell analysis_Red cell distribution width-CV (%)	0	94 (57.67)	21 (38.89)	73 (66.97)	0.0012
	1	69 (42.33)	33 (61.11)	36 (33.03)	
Blood cell analysis_Hematocrit (%)	0	62 (38.04)	26 (48.15)	36 (33.03)	0.0891
	1	101 (61.96)	28 (51.85)	73 (66.97)	
Blood cell analysis_WBC (%)	0	84 (51.53)	24 (44.44)	60 (55.05)	0.2678
	1	79 (48.47)	30 (55.56)	49 (44.95)	
Blood cell analysis_Plateletcrit (%)	0	142 (87.12)	44 (81.48)	98 (89.91)	0.2065
	1	21 (12.88)	10 (18.52)	11 (10.09)	
Blood cell analysis_Platelet count (%)	0	146 (89.57)	47 (87.04)	99 (90.83)	0.6365
	1	17 (10.43)	7 (12.96)	10 (9.17)	
Blood cell analysis_Hemoglobin (%)	0	129 (79.14)	39 (72.22)	90 (82.57)	0.185
	1	34 (20.86)	15 (27.78)	19 (17.43)	
Lipid profile_LDL-C (%)	0	39 (23.93)	5 (9.26)	34 (31.19)	0.0038
	1	124 (76.07)	49 (90.74)	75 (68.81)	
Lipid profile_Total cholesterol (%)	0	84 (51.53)	21 (38.89)	63 (57.80)	0.0351
	1	79 (48.47)	33 (61.11)	46 (42.20)	
Lipid profile_Triglycerides (%)	0	133 (81.60)	34 (62.96)	99 (90.83)	< 0.0001
	1	30 (18.40)	20 (37.04)	10 (9.17)	
Lipid profile_HDL-C (%)	0	84 (51.53)	22 (40.74)	62 (56.88)	0.076
	1	79 (48.47)	32 (59.26)	47 (43.12)	
Random blood glucose (%)	0	91 (55.83)	27 (50.00)	64 (58.72)	0.375
	1	72 (44.17)	27 (50.00)	45 (41.28)	
Non-small cell lung cancer-associated antigen (%)	0	74 (45.40)	19 (35.19)	55 (50.46)	0.0937
	1	89 (54.60)	35 (64.81)	54 (49.54)	
Blood cell analysis_MCV (mean (SD))		90.702 (8.244)	91.048 (9.478)	90.530 (7.601)	0.7067
Blood cell analysis_MCH (mean (SD))		29.691 (2.952)	29.514 (2.925)	29.780 (2.974)	0.5896
Blood cell analysis_MCHC (mean (SD))		327.977 (12.777)	326.465 (13.344)	328.727 (12.482)	0.2889
Blood cell analysis_MPV (mean (SD))		9.956 (1.673)	9.919 (2.119)	9.974 (1.412)	0.8419
Blood cell analysis_PDW (mean (SD))		14.099 (3.036)	14.231 (3.341)	14.034 (2.887)	0.6976

total protein and ADA significantly increased the diagnostic value, while albumin and neutrophil count had the greatest negative impact. As a result, the model diagnosed the patient as non-TPE type ($f(x) = 0.0963, < 0.5$). The confusion matrix indicated that the model exhibited high sensitivity and specificity on both the training set (sensitivity = 0.909, accuracy = 0.938) and the test set (sensitivity = 0.977, accuracy = 0.957) (Fig. 3G, H). The F1 score and Kappa statistics further demonstrated the model's excellent diagnostic consistency on both the training set (F1 score = 0.944, Kappa statistics = 0.908) and the test set (F1 score = 0.87, Kappa statistics = 0.829).

Superior diagnostic performance of traditional logistic model for TPE

We established a traditional multivariate logistic regression model to diagnose TPE. We selected variables sequentially using forward stepwise logistic regression, resulting in 12 variables. We then performed a comprehensive evaluation of its diagnostic performance. First, we presented the regression analysis results of each biochemical and hematological variable in the logistic regression model using a forest plot (Fig. 4A). The results revealed that ADA in pleural fluid biochemistry (PFB), albumin in PFB, and alkaline phosphatase had odds ratios (OR) of 24.63 (95% CI: 5.22–169.75, $P < 0.001$), 10.75 (95% CI: 1.47–121.88, $P = 0.03$), and 3.62 (95% CI: 0.73–21.61, $P = 0.13$), respectively. These variables played a significant role in diagnosing TPE. We displayed the logistic regression model's scoring system through a nomogram (Fig. 4B). Each variable's score was color-coded and mapped to the probability of TPE occurrence, with higher scores corresponding to a higher likelihood of TPE. This nomogram provides a convenient diagnosis tool for clinicians. ROC curves in the training and test sets demonstrated the diagnostic performance of the logistic regression model (Fig. 4C, D). The AUC in the training set was 0.96 (95% CI: 0.93–0.99), and in the test set, the AUC was 0.95 (95% CI: 0.89–1.00), indicating the model's high diagnostic accuracy and robustness. Moreover, we assessed the clinical utility of the model at various treatment thresholds using DCA (Fig. 4E, F). The results showed that, across a wide range of treatment thresholds, the logistic regression model offered a greater net benefit than the "treat all" and "treat none" strategies. This supports its potential value in clinical practice. In the training set, after 1000 bootstrap resamplings, the ROC curve further validated the model's robustness, with an AUC consistently around 0.96 (Fig. 4G). Additionally, comparisons between forward selection, backward selection, and forward–backward selection methods showed minimal variation in AUC. The AUC remained in the range of 0.95–0.96, confirming the model's consistent diagnostic performance (Fig. 4H). Finally, we compared the performance of eight published models in the training and test sets (Fig. 4I, J). Our logistic regression model outperformed other models, such as the Wu model (training AUC = 0.87, test AUC = 0.78) and the Li model (training AUC = 0.74, test AUC = 0.81). Our model achieved significantly higher AUC values (training AUC = 0.96, test AUC = 0.95), suggesting superior performance in differentiating between TPE and non-TPE patients.

Effective diagnosis of TPE using large language model (LLM)

We innovatively employed LLMs such as ChatGPT-4 and ChatGPT-4o to assess their performance in diagnosing TPE. First, ChatGPT-4 rated the importance of different biochemical and hematological indicators (Fig. 5A, Supplementary Table 7), revealing that biochemical markers

(See figure on next page.)

Fig. 2 Clinical characteristic landscape of TPE. **A** PLS-DA score plot showing the separation between non-TPE and TPE samples based on blood routine and biochemical markers. The percentage of variance explained by each principal component (PC1 and PC2) is indicated in the figure. The distribution of the TPE group and non-TPE group was distinguishable along PC1 (16.76%) and PC2 (16.37%). **B** Heatmap displaying the expression levels of various biochemical and hematological variables in non-TPE and TPE samples. Samples are clustered into two major groups, with variable names listed on the right side. Red indicates values greater than the corresponding cutoff point (1), and green indicates values smaller than the corresponding cutoff point (0). The variables include: (1) Liver function: Alkaline phosphatase (ALP), Total bile acids (TBA); (2) Electrolytes: Osmolality, Sodium (Na); (3) Tumor markers: Carcinoembryonic antigen (CEA); (4) Hematological analysis: Red cell distribution width-CV (RDW-CV), Neutrophil count, Neutrophil percentage, Lymphocyte count, Monocyte percentage; (5) Hospitalization: Length of hospital stay after pleural biopsy; (6) Lipid profile: Triglycerides; (7) PE biochemistry: Adenosine deaminase (ADA), Albumin, Total protein; (8) Age. The figure illustrates the distribution differences of these variables between the TPE and non-TPE groups. **C** The bar chart displays differences in the levels of ADA, monocytes, and total protein. ADA (PE biochemistry): The proportion differences between the TPE and non-TPE groups when ADA values are either below or above the cutoff point (25.85). Monocyte proportion (blood cell analysis): The proportion differences between the TPE and non-TPE groups when the monocyte proportion is either below or above the cutoff point (6.95). Total protein (PE biochemistry): The proportion differences between the TPE and non-TPE groups when total protein values are either below or above the cutoff point (47.35). **D** ROC curves for distinguishing non-TPE and TPE using pleural fluid ADA levels, blood routine monocyte percentage, and total protein in pleural fluid as biomarkers. The Area Under the Curve (AUC) and 95% CI for each biomarker are provided. These were used to evaluate the diagnostic ability of ADA (PE biochemistry), monocyte proportion (blood cell analysis), and total protein (PE biochemistry) in distinguishing between the TPE and non-TPE groups

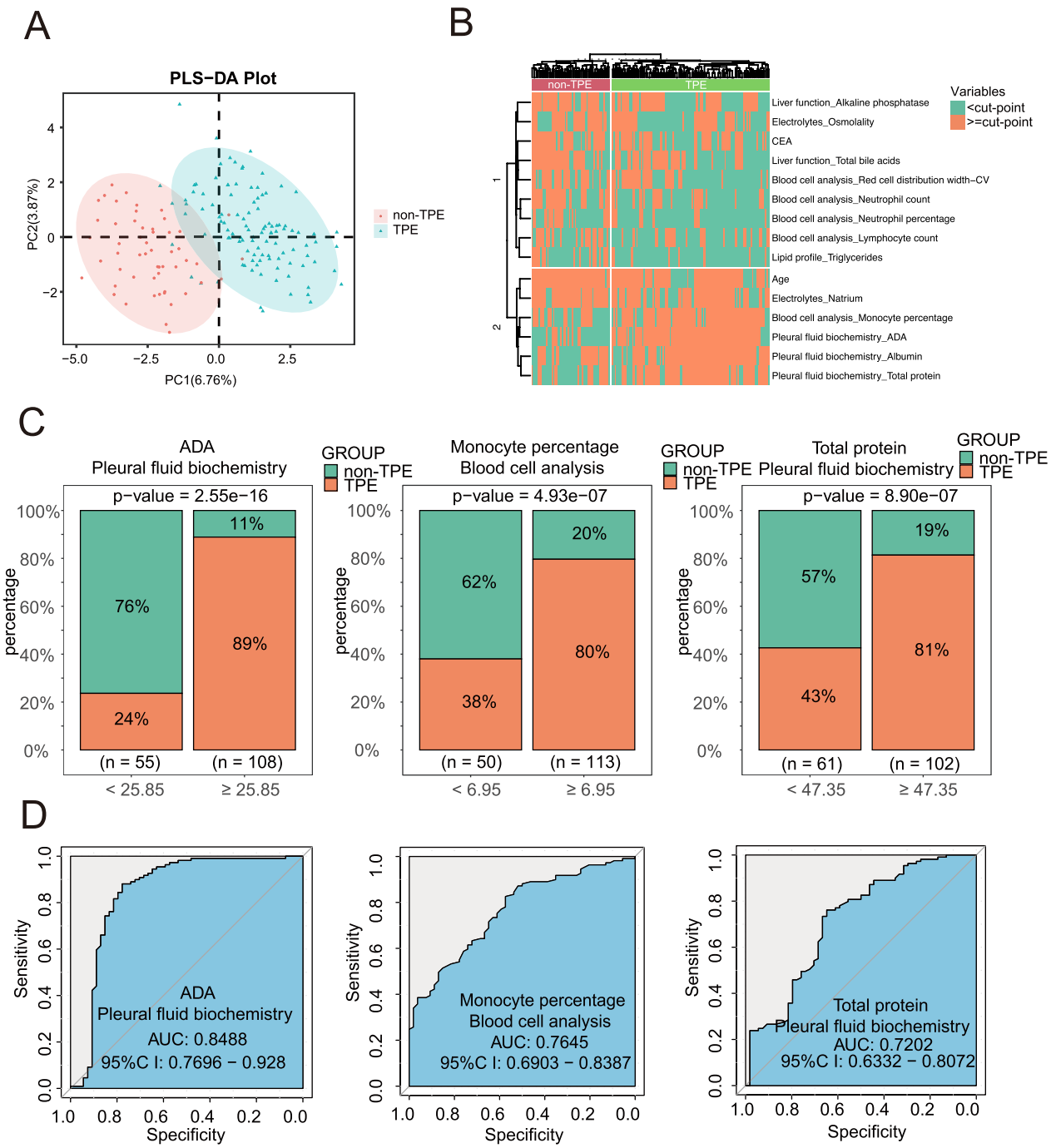

Fig. 2 (See legend on previous page.)

in PE, particularly ADA and total protein, received the highest scores. The percentage of neutrophils and lymphocyte counts in blood cell analysis also demonstrated high importance, which aligns with prior research findings [27]. Next, we compared the performance of four models: the best machine learning model (MLbest)-XGBoost, ChatGPT-4, ChatGPT-4o, and logistic regression. We evaluated these models using various metrics, including AUC, specificity, sensitivity, accuracy, F1 score, negative predictive value (NPV), and positive predictive value (PPV) (Fig. 5B). The results showed that ChatGPT-4o and MLbest (XGBoost) outperformed the others across all metrics. Both achieved AUCs approaching 1.00, with high sensitivity and specificity,

outperforming the traditional logistic regression model. This indicates that they hold significant application potential for diagnosing TPE. Although ChatGPT-4 slightly underperformed compared to ChatGPT-4o, it still demonstrated strong diagnostic capabilities across all metrics. We analyzed the diagnostic results of the best ML (XGBoost) model and the LLM-ChatGPT model in this study. We found that the diagnoses of the two models were consistent for 43 cases in the test set. However, there were discrepancies in 5 cases, including 4 cases of TPE and 1 case of non-TPE. This suggests that the two models showed a higher discrepancy rate for cases of TPE than for non-TPE cases (Supplementary Fig. 2). Furthermore, we developed the Python package "tpeai" (version 0.2.0) (https://pypi.org/project/tpeai/) (Fig. 5C). This package integrates ChatGPT-4 for distinguishing between TPE and non-TPE groups. By combining biochemical and hematological data, this tool effectively supports clinical diagnosis. We showed the specific output results of the "tpeai" package in diagnosing tuberculous PE, along with a detailed display of ChatGPT-4's logical reasoning and thought process during the analysis (Fig. 5D). In summary, the LLM-based ChatGPT-4 model demonstrates excellent performance in diagnosing TPE. By integrating multiple biochemical and hematological indicators, it can effectively diagnose TPE and provide valuable support for clinical decision-making.

Discussion

This study innovatively employed large language model (LLM) to diagnose TPE and compared its performance with traditional machine learning models and logistic regression models. The aim was to explore the potential application of LLM in the diagnosis of TPE. The

(See figure on next page.)

Fig. 3 Machine learning models effectively diagnose TPE. **A** The plot illustrates the path of coefficients for different biochemical and hematological variables in lasso regression as the regularization parameter λ (Log Lambda) varies. The x-axis represents Log Lambda, the logarithm of the regularization parameter, and the y-axis represents the regression coefficients for each variable. As λ increases, the coefficients gradually approach zero, indicating that lasso regression performs variable selection by shrinking the coefficients. Curves of different colors represent different variables, showing the changes in their coefficients during the model regularization process. **B** The plot displays the cross-validation process of lasso regression, where the y-axis represents Binomial Deviance, and the x-axis shows different values of Log(λ). The shaded gray area indicates the standard error range, and the red curve represents the mean of the binomial deviance. Through cross-validation, it is observed that as λ changes, the binomial deviance decreases, reaching a minimum, and the corresponding λ at this point represents the optimal regularization parameter. The optimal λ value, marked by the dashed line, is the best regularization parameter chosen by the model. **C** Heatmap comparing AUC and F1 scores of the top 93 models out of 453 machine learning models. This heatmap compares the AUC and F1 scores of various machine learning models, including those using XGBoost, GBM, DeepLearning, and other algorithms. The results include AUC and F1 scores for the training set, test set, and average values. Each model's performance is ranked according to its AUC and F1 scores, with higher values indicating better performance. The color bar in the table represents the values of AUC and F1 scores, with the intensity of color reflecting the level of performance. AUC: AUC measures the model's classification ability. An AUC value closer to 1 indicates better performance. In this heatmap, AUC values are presented for the training set, test set, and average, showing the performance variation of different models across different datasets. F1 Score: The F1 score is an indicator of a classification model's accuracy, balancing precision and recall. A higher F1 score suggests better balance in the model's performance across positive and negative classes. This heatmap displays the F1 scores of each model for the training and test sets and provides average values to facilitate comparisons of model performance at different stages. Model Name: Each row represents a different machine learning model, including various configurations of algorithms such as XGBoost, GBM, DeepLearning, etc. (e.g., model names like XGBoost_grid_1_model_108), and their corresponding AUC and F1 scores on the training and test sets. **D** The plot illustrates the importance of each variable in the best-performing XGBoost-based machine learning model on the test set. The importance of each variable is represented by the length of the corresponding bar, with longer bars indicating a greater contribution of that variable to the model's diagnostic capability. The most important variables include ADA, alkaline phosphatase, PE biochemical markers (albumin, total protein), and hematological analysis variables (neutrophil count, monocyte percentage, etc.). **E** This plot displays the SHAP values for each feature in the best XGBoost-based machine learning model, representing the contribution of each feature to the model's output. Features are listed on the y-axis, and the corresponding SHAP values are plotted along the x-axis. Each point represents a data point, and the color of the point indicates the value of the feature (ranging from low to high, with the color scale displayed on the right). Positive SHAP values (to the right of the vertical line) increase the model's diagnostic value, while negative SHAP values (to the left of the vertical line) decrease the diagnostic value. The features with the greatest impact on the diagnosis are positioned at the top of the plot. **F** SHAP analysis showing the contribution of multiple features to the model's diagnosis of specific samples (TPE vs. non-TPE). The x-axis represents the SHAP values, reflecting each feature's contribution to the diagnosis. Movement to the right indicates an increase in the diagnostic value, while movement to the left indicates a decrease. The cumulative effect of the SHAP values determines the final model diagnosis. The difference between the final diagnostic value, $f(x) = 0.0963$, and the expected value, $E(f(x)) = -0.626$, is reflected by the SHAP values of each feature. **G** Training set confusion matrix. This matrix displays the model's diagnostic results on the training set. In the matrix, Control represents normal cases, and Case represents diseased cases. The model's correctly diagnostic categories and misclassifications are as follows: True positives (TP): 68; False positives (FP): 1; True negatives (TN): 42; False negatives (FN): 4; The model's performance metrics on the training set are: sensitivity = 0.977, specificity = 0.944, accuracy = 0.913, recall = 0.944, F1 score = 0.944, and Kappa value = 0.908. **H** Test set confusion matrix. This matrix shows the model's diagnostic results on the test set: True positives (TP): 35; False positives (FP): 1; True negatives (TN): 10; False negatives (FN): 2; The model's performance metrics on the test set are: sensitivity = 0.909, specificity = 0.946, accuracy = 0.833, recall = 0.909, F1 score = 0.87, and Kappa value = 0.829

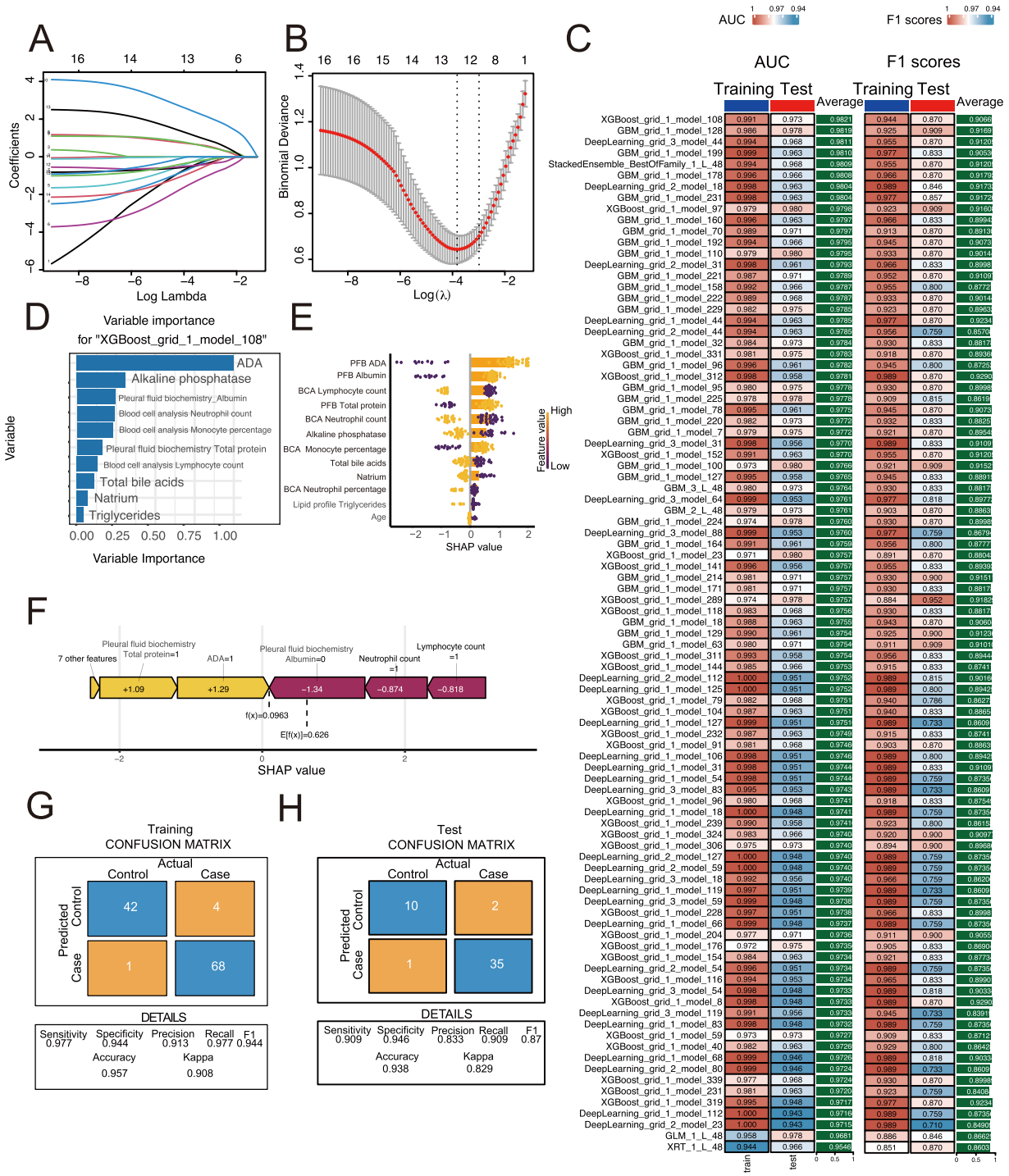

Fig. 3 (See legend on previous page.)

results demonstrate that the LLM model effectively integrates various clinical variables and distinguishes between TPE and non-TPE. Compared to traditional logistic regression models and common machine learning algorithms, the LLM model performed similarly to standard machine learning methods. It outperformed the logistic regression model in terms of sensitivity, specificity, and other evaluation metrics. These findings validate the potential application value of LLM for early diagnosis of TPE.

Traditional diagnostic methods for TPE primarily rely on pleural effusion (PE) cultures and pleural biopsy. However, these methods have a high risk of missed diagnoses, and PE culture results often take up to 8 weeks to be available [28]. Therefore, early and accurate identification of TPE remains a pressing challenge. In this study, we successfully developed a diagnostic model based on an LLM artificial intelligence framework using clinical data from TPE patients. This model enables faster and more effective diagnosis in clinical settings. Compared to traditional diagnostic workflows, this model not only offers higher diagnostic efficiency but also demonstrates greater practical value for non-surgical diagnostic tools.

From a clinical perspective, the LLM-based TPE diagnostic model offers a novel approach to addressing the limitations of existing diagnostic methods for TPE. As a manifestation of tuberculosis on the pleura, the accuracy and efficiency of TPE diagnosis are crucial for timely treatment, particularly in regions with a high incidence of tuberculosis. Traditional TPE diagnosis primarily relies on pleural biopsy; however, this is surgical and time-consuming. Previous studies have validated the role of biomarkers such as adenosine deaminase (ADA) and lymphocyte ratio in TPE diagnosis [29], but these individual biochemical or cytological indicators are insufficient to capture the full complexity of TPE presentations. In this study, the LLM model integrates various indicators, including ADA, total protein, and monocyte percentage. This non-surgical, cost-effective diagnostic approach helps clinicians diagnose TPE more quickly and accurately under non-surgical conditions, supporting early diagnosis. Furthermore, variables such as ADA in PE, total protein, and monocyte percentage in the blood count were found to be strong diagnostic markers for distinguishing TPE from non-TPE. The key role of ADA levels in TPE diagnosis was further validated, consistent with previous research findings [30].

The LLM model developed in this study demonstrates performance comparable to the machine learning models established in our article. It outperforms previous machine learning-based models and traditional logistic regression models for diagnosing TPE, showing superior diagnostic ability. Previous studies using machine learning for TPE diagnosis have primarily relied on algorithms such as Random Forests (RF) and Support Vector Machines (SVM). For example, Li et al.

(See figure on next page.)

Fig. 4 Traditional Logistic Model Effectively Diagnoses TPE. **A** Forest plot of the TPE diagnostic model. This plot presents the odds ratios (OR) and p-values of the variables derived from the multivariate logistic regression model. The OR for each variable is displayed with horizontal error bars, where the length of the bars reflects the confidence interval of the OR, and the point represents the estimated OR value. The variables include age, various biochemical markers, and hematological parameters. PFB, pleural fluid biochemistry; BCA, blood cell analysis; Neu, Neutrophil; Mono, monocyte percentage; RDW, red cell distribution width; CV, coefficient of variation. **B** This nomogram illustrates the scores of various variables calculated by the TPE diagnostic model and their corresponding probabilities of TPE (Pr(GROUP)). Each variable in the model is assigned a score, with the variable's value weighted according to its corresponding points, which influences the overall score and subsequently diagnoses the likelihood of TPE. The plot lists several clinical variables (e.g., Triglycerides, Total bile acids, Age, Albumin, etc.) and their corresponding scores. Each variable's score is indicated by a red circle, while the blue box displays the score range for that variable. Some variables (e.g., ADA, Adenosine deaminase) have higher scores, indicating a larger contribution to the model. The lower part of the figure shows the total score, derived from the weighted sum of all variable scores, along with the diagnostic probability of TPE (Pr(GROUP)). As the total score increases, the diagnostic probability of TPE also rises significantly. **C** ROC curve for the logistic regression model used to identify TPE in the training set, showing the model's performance at various specificity and sensitivity levels, including the AUC and 95% CI. **D** ROC curve for the logistic regression model in the test set, showing the diagnostic performance and statistical data. **E** This plot illustrates the net benefit of the TPE logistic regression model in the training set. The x-axis represents the treatment threshold probability, and the y-axis represents the net benefit. Different decision thresholds influence the effectiveness of the treatment strategy, with the model demonstrating a good net benefit in both low and high probability ranges. The red curve represents the net benefit of treating all patients, the green curve represents the net benefit of treating no patients, and the blue curve represents the model's diagnostic effectiveness. **F** This plot shows the net benefit evaluation of the TPE logistic regression model in the test set. Similar to the training set, the model exhibits good diagnostic performance across varying thresholds. The red and green curves again represent the net benefit of treating all patients and treating no patients, respectively, while the blue curve represents the net benefit of the treatment strategy predicted by the logistic regression model. **G** This figure displays the ROC curve results from 1000 bootstrap samples, assessing the model's classification performance in the training set. The blue curve represents the average ROC curve based on the training set data, while the gray shaded area indicates the variability of the results from the 1000 bootstrap samples, showing the changes in sensitivity and specificity across different samples. The figure demonstrates the stability and performance of the model assessed through bootstrap sampling on the training set, with the blue curve showing a high classification accuracy, indicating the model's robust diagnostic ability. **H** This plot presents the ROC curves of the TPE logistic regression model and various stepwise regression methods, along with their corresponding AUC values and 95% confidence intervals. The TPE logistic regression model using forward stepwise regression (red curve) performed the best, with an AUC of 0.96. **I, J** ROC curve comparison of different models in the training set **(I)** and test set **(J)**. This plot displays the ROC curve comparison between the TPE logistic regression model and other published models (including Wu, Li, Zhou, Liu, Li2, Ren, Lei, and Wang models) in the training set. The TPE logistic regression model (green curve) demonstrates the best AUC value (0.96) in the training set compared to the other models. The plot highlights the AUC values of the different models, emphasizing the superiority of the TPE model in diagnosing TPE

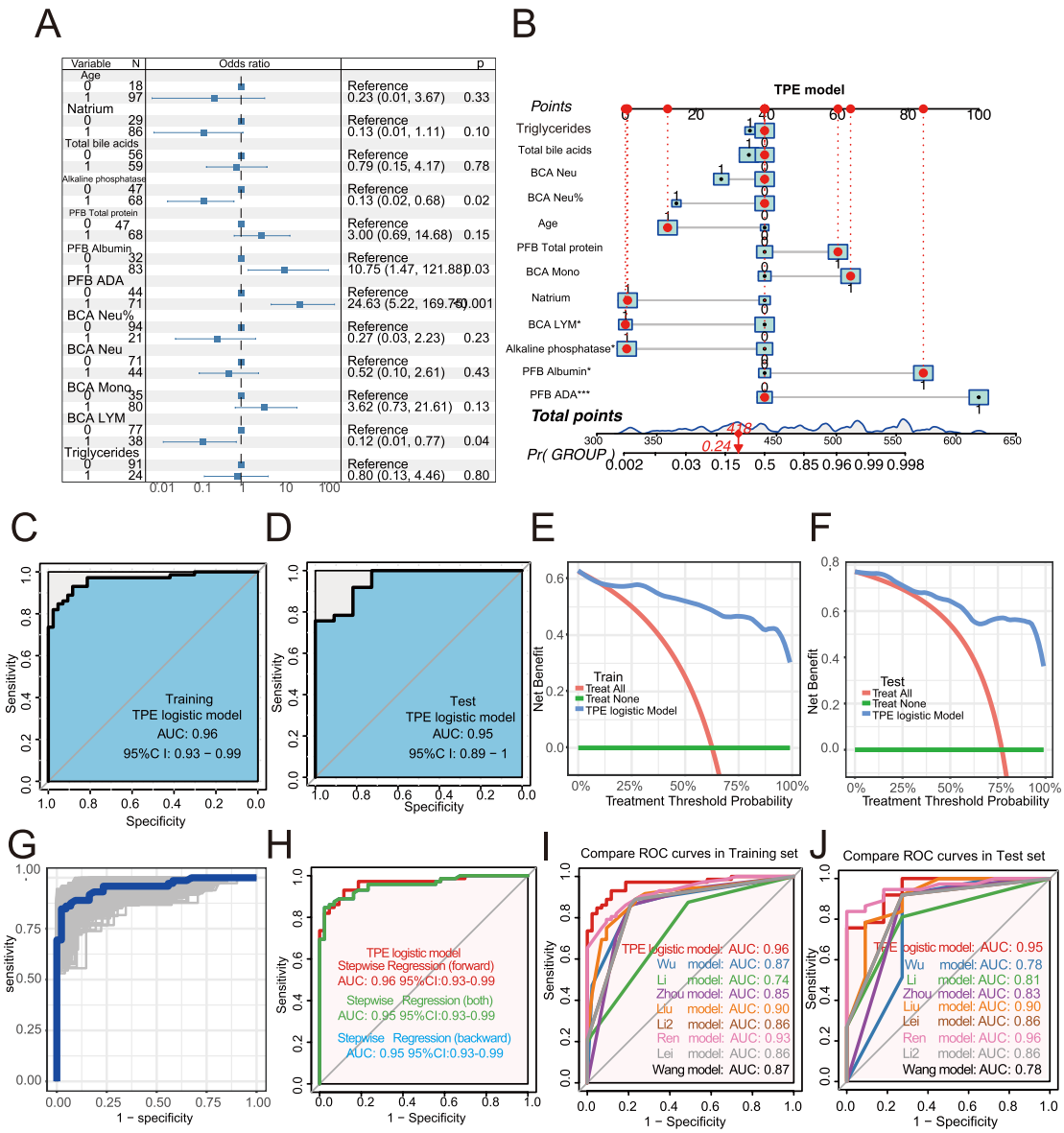

Fig. 4 (See legend on previous page.)

[12] improved diagnosis accuracy with an SVM model (bGACO-SVM), and Zhou et al. [10] proposed a new algorithm, CFDE, with an SVM model for feature selection in TPE diagnosis. However, the performance of these models was weaker than the LLM model developed in this study. Logistic regression models are widely used in disease classification. For example, Li et al. [13] reported an area under the curve (AUC) of 0.92 for TPE diagnosis using logistic regression. However, logistic regression models are limited by linear relationships and struggle to accurately capture complex interactions between nonlinear features. In contrast, the LLM model is better suited for handling large datasets and nonlinear feature data. In this study, the LLM model outperformed the logistic regression model across multiple metrics, including AUC, F1 score, and sensitivity.

The LLM model's excellent performance can be attributed to its ability to handle complex data and integrate multiple variables [16]. In contrast, while machine learning is advantageous in certain specific applications (e.g., small sample data scenarios and lightweight real-time applications), it generally suffers from poor interpretability and limited generalizability. Traditional logistic regression models, although advantageous in interpretability, are constrained by their assumption of

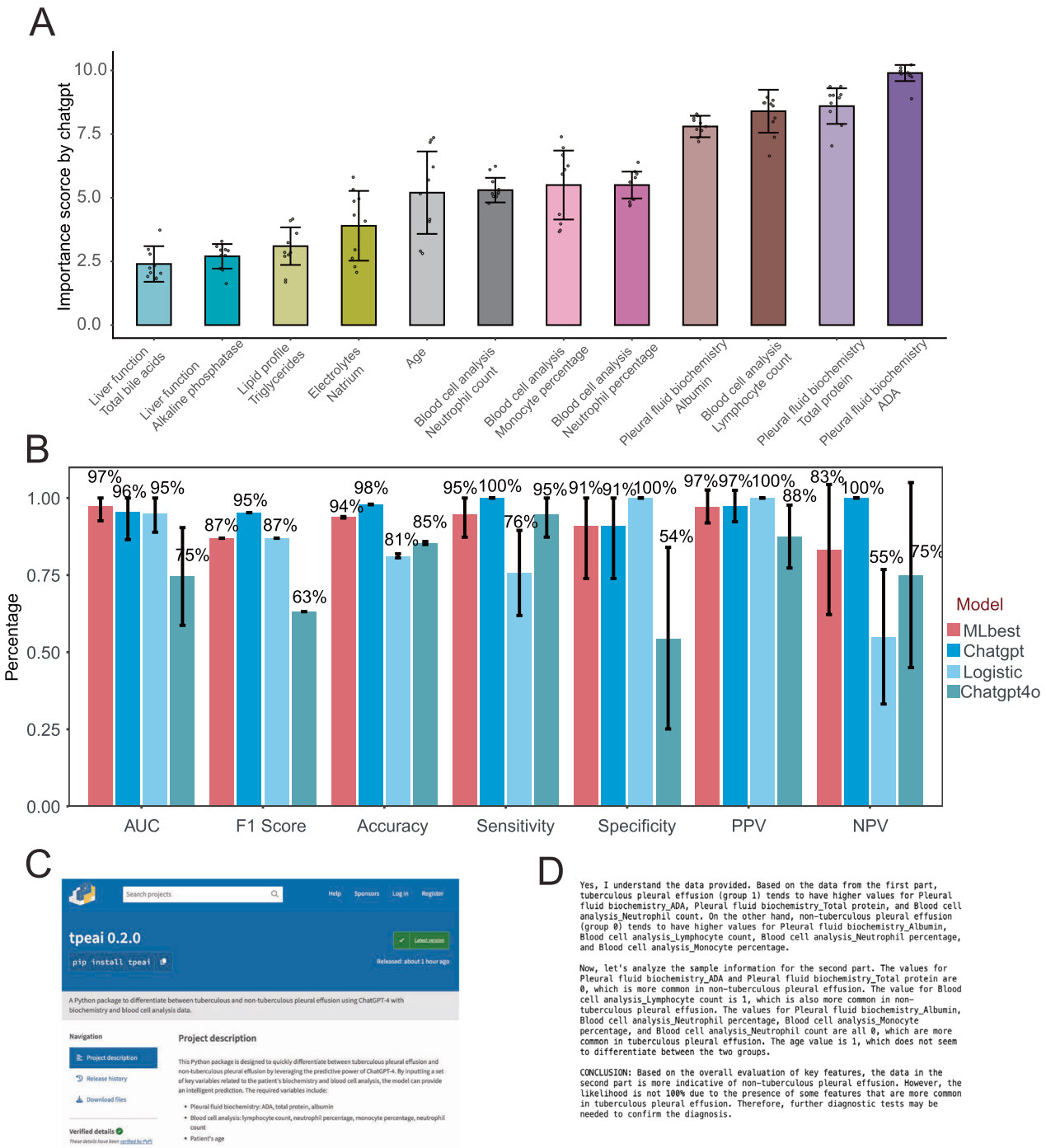

Fig. 5 LLM (ChatGPT-4, ChatGPT-4o) for Diagnosing TPE. **A** This plot presents the importance scores of various clinical variables computed by the ChatGPT model. The importance scores are represented by bar charts, with the length of each bar indicating the relative importance of the variable. The error bars on each bar represent the range of variability in the importance scores of the variables. **B** This plot presents the performance of each model (MLbest, ChatGPT-4, Logistic Regression, and ChatGPT-4o) across various metrics, including AUC, F1 score, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The performance of each model is displayed using bar charts, with black error bars representing the standard error for each metric. **C** Information about the Python package "tpeai" (version 0.2.0) created in this study, which uses ChatGPT-4 to distinguish between TPE and non-TPE groups. **D** Output from using the "tpeai" package to diagnose TPE. The figure includes a discussion of how ChatGPT-4 interprets the data and model diagnosis, detailing how different biochemical and hematological markers help differentiate TPE group from non-TPE group, along with the logical reasoning and thought process in the diagnosis

linear relationships, limiting their applicability in complex clinical settings.

Despite its advantages, the LLM model has certain limitations. Small sample sizes may lead to overfitting, compromising the model's generalizability. Its reliance on specific biomarkers further restricts its applicability across different regions. Additionally, the random feature selection and increased complexity can complicate result interpretation. While this study demonstrates the LLM model's potential in diagnosing TPE, further validation is necessary for real-world settings. The limited sample size, lack of multi-center data, and absence of external validation may affect the model's stability. Future research could explore integrating multimodal data, such as genomic and imaging information, with the current model to enhance diagnostic accuracy and address these limitations.

Conclusion

This study suggests that the LLM-based diagnostic model provides a novel approach for the early non-surgical diagnosis of TPE. The ChatGPT-4 Python package, named "tepai" (https://pypi.org/project/tpeai/), developed in this study, provides a simple and user-friendly tool for clinicians. It allows for the rapid generation of diagnostic recommendations based on basic biochemical and hematological data inputs. Further optimization of this tool will enhance its ability to support precise diagnosis and personalized treatment for TPE. As artificial intelligence technology advances, the application of LLM is expected to expand. When combined with multimodal data such as imaging and genomic data, it could significantly improve the diagnostic efficiency and accuracy of TPE. Future research should focus on validating the LLM model in larger, multi-center datasets to ensure its broad applicability and robustness.

Abbreviations

TPE	Tuberculous pleural effusion
LLM	Large language model
AUC	Area under the curve
TB	Tuberculosis
PE	Pleural effusion
ADA	Adenosine deaminase
AI	Artificial intelligence
MLA	Machine learning algorithms
KNN	K-nearest neighbors
RF	Random forests
SVM	Support vector machines
CT	Computed tomography
LDH	Lactate dehydrogenase
CRP	C-reactive protein
ESR	Erythrocyte sedimentation rate
WBC	White blood cells
CEA	Carcinoembryonic antigen
NSE	Neuron-specific enolase
ROC	Receiver operating characteristic
PLS-DA	Partial least squares discriminant analysis
VIP	Variable importance projection
VIF	Variance inflation factor
LASSO	Least absolute shrinkage and selection operator
CI	Confidence intervals
DCA	Decision curve analysis
OR	Odds ratios
SD	Standard deviation
PFB	Pleural fluid biochemistry
MLbest	The best machine learning model
NPV	Negative predictive value
PPV	Positive predictive value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-025-03130-y.

Additional file 1.
Additional file 2.
Additional file 3.
Additional file 4.
Additional file 5.
Additional file 6.
Additional file 7.
Additional file 8.
Additional file 9.

Author contributions

JLT and XLY conceived the project. Data analysis was performed by CLW, WYL, and PFM. The interpretation of the data involved contributions from CLW, WYL, PFM, YYL, JC, LL, JW, XFL, MXW, YYC, MBH, QH, QH, XLY and JLT. All authors contributed to writing and approving the final manuscript.

Funding

This study was supported by the Natural Science Foundation of Jiangxi Province (20202BABL206116), the Startup Fund for scientific research, Fujian Medical University (XJ2021018101), and Key Research and Development Program of Ganzhou City, Jiangxi Province (2023LNS37008).

Data Availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Affiliated Hospital of Jiujiang University (Approval No.: jjumer-b-2024-0405) and was conducted in accordance with the Declaration of Helsinki. As retrospective data were used, written informed consent was waived. This article does not include any research involving human participants by the authors. All data were analyzed anonymously to ensure the privacy of the participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Respiratory Medicine, Affiliated Hospital of Jiujiang University, No. 57 East Xunyang Road, Xunyang District, Jiujiang 332000, China. [2]Department of Hematology, The Second Affiliated Hospital of Fujian Medical University, Quanzhou 362000, China. [3]Department of Gastroenterology, Affiliated Hospital of Jiujiang University, Jiujiang 332000, China. [4]Department of Cardiology, Affiliated Hospital of Jiujiang University, Jiujiang 332000, China. [5]Department of Respiratory Medicine, First Affiliated Hospital of Gannan

Medical University, No. 23, Qingnian Road, Zhanggong District, Ganzhou 341000, China.

Received: 20 November 2024 Accepted: 30 January 2025
Published online: 12 February 2025

References

1. Shaw JA, Diacon AH, Koegelenberg CFN. Tuberculous pleural effusion. Respirology. 2019;24(10):962–71.
2. Xu H-Y, Li C-Y, Su S-S, et al. Diagnosis of tuberculous pleurisy with combination of adenosine deaminase and interferon-γ immunospot assay in a tuberculosis-endemic population: a prospective cohort study. Medicine (Baltimore). 2017;96(47): e8412.
3. Zhai K, Lu Y, Shi H-Z. Tuberculous pleural effusion. J Thorac Dis. 2016;8(7):E486–94.
4. Choi H, Chon HR, Kim K, et al. Clinical and Laboratory differences between lymphocyte- and neutrophil-predominant pleural tuberculosis. PLoS ONE. 2016;11(10): e0165428.
5. Lee SJ, Kim HS, Lee SH, et al. Factors influencing pleural adenosine deaminase level in patients with tuberculous pleurisy. Am J Med Sci. 2014;348(5):362–5.
6. Li D, Shen Y, Fu X, Li M, Wang T, Wen F. Combined detections of interleukin-33 and adenosine deaminase for diagnosis of tuberculous pleural effusion. Int J Clin Exp Pathol. 2015;8(1):888–93.
7. Kirsch CM, Kroe DM, Azzi RL, Jensen WA, Kagawa FT, Wehner JH. The optimal number of pleural biopsy specimens for a diagnosis of tuberculous pleurisy. Chest. 1997;112(3):702–6.
8. Gruson D, Helleputte T, Rousseau P, Gruson D. Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. Clin Biochem. 2019;69:1–7.
9. Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, et al. Potential value and impact of data mining and machine learning in clinical diagnostics. Crit Rev Clin Lab Sci. 2021;58(4):275–96.
10. Zhou X, Chen Y, Gui W, et al. Enhanced differential evolution algorithm for feature selection in tuberculous pleural effusion clinical characteristics analysis. Artif Intell Med. 2024;153: 102886.
11. Ren Z, Hu Y, Xu L. Identifying tuberculous pleural effusion using artificial intelligence machine learning algorithms. Respir Res. 2019;20(1):220.
12. Li C, Hou L, Pan J, Chen H, Cai X, Liang G. Tuberculous pleural effusion prediction using ant colony optimizer with grade-based search assisted support vector machine. Front Neuroinform. 2022;16:1078685.
13. Li C, Hou L, Sharma BY, et al. Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. Comput Methods Programs Biomed. 2018;153:211–25.
14. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med. 2024;7(1):205.
15. Shao J, Ma J, Yu Y, et al. A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. Innovation (Camb). 2024;5(4): 100648.
16. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;2:642.
17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.
18. Blank IA. What are large language models supposed to model? Trends Cogn Sci. 2023;27(11):987–9.
19. Visibelli A, Roncaglia B, Spiga O, Santucci A. The impact of artificial intelligence in the odyssey of rare diseases. Biomedicines. 2023;11(3):887.
20. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt. 2023;43(6):1562–70.
21. Lapidus D. Strengths and limitations of new artificial intelligence tool for rare disease epidemiology. J Transl Med. 2023;21(1):292.
22. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ. 2024;10: e51391.
23. Zheng Y, Sun X, Feng B, et al. Rare and complex diseases in focus: ChatGPT's role in improving diagnosis and treatment. Front Artif Intell. 2024;7:1338433.
24. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for diagnosing rare eye diseases? A pilot study. Ophthalmol Ther. 2023;12(6):3395–402.
25. Clerici CA, Chopard S, Levi G. Rare disease in the age of artificial intelligence. Recenti Prog Med. 2024;115(2):67–75.
26. Ferreiro L, Toubes ME, San José ME, Suárez-Antelo J, Golpe A, Valdés L. Advances in pleural effusion diagnostics. Expert Rev Respir Med. 2020;14(1):51–66.
27. Jeon DS, Kim S-H, Lee JH, Choi C-M, Park HJ. Conditional diagnostic accuracy according to inflammation status and age for diagnosing tuberculous effusion. BMC Pulm Med. 2023;23(1):400.
28. Lo Cascio CM, Kaul V, Dhooria S, Agrawal A, Chaddha U. Diagnosis of tuberculous pleural effusions: a review. Respir Med. 2021;188: 106607.
29. Garcia-Zamalloa A, Taboada-Gomez J. Diagnostic accuracy of adenosine deaminase and lymphocyte proportion in pleural fluid for tuberculous pleurisy in different prevalence scenarios. PLoS ONE. 2012;7(6): e38729.
30. Chan C, Chan KKP. Pleural fluid biomarkers: a narrative review. J Thorac Dis. 2024;16(7):4764–71.

Publisher's Note