INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 20, Issue: 1, Month: February, Year: 2025
Article Number: 6830, https://doi.org/10.15837/ijccc.2025.1.6830

# An Interpretable Machine Learning Based Model for Traumatic Severe Pneumothorax Evaluation

Y. Lv, J. Weng, J. Li*, W. Chen*, H. Huang, Y. Zhao

**Yinzhen Lv**
School of Economics and Management
Beijing Jiaotong University, Beijing, China
3 Shangyuan Village, Xizhimenwai, Haidian District, Beijing, China
22711022@bjtu.edu.cn

**Jiayi Weng**
School of Economics and Management
Beijing Jiaotong University, Beijing, China
3 Shangyuan Village, Xizhimenwai, Haidian District, Beijing, China
21711068@bjtu.edu.cn

**Jing Li***
School of Economics and Management
Beijing Jiaotong University, Beijing, China
3 Shangyuan Village, Xizhimenwai, Haidian District, Beijing, China
*Corresponding author: jingli@bjtu.edu.cn

**Wei Chen***
Department of Emergency
The Third Medical Center to Chinese People's Liberation Army General Hospital, Beijing, China
69 Yongding Road, Haidian District, Beijing, China
*Corresponding author: 778027058@qq.com

**Yuzhuo Zhao**
Department of Cardiology
The 79th Military Hospital of the People's Liberation Army, Liaoning, China
148 Weiguo Road, Baita District, Liaoyang City, Liaoning Province, China
zhaoyuzhuo0421@126.com

**Heshan Huang**
Department of Emergency
the Third Medical Center of Chinese PLA General Hospital, Beijing, China
28 Fuxing Road, Haidian District, Beijing , China
doctor_hhs@163.com

**Abstract**

Traumatic pneumothorax is a complex condition that is challenging to diagnose, particularly in hospitals, underdeveloped areas, and during mass casualty events. This study aimed to evaluate the potential of machine learning (ML) for diagnosing and assessing traumatic pneumothorax. We extracted 33 vital signs and blood gas parameters from the MIMIC-IV database, selecting 12 clinically significant features as inputs to four ML algorithms: extreme gradient boosting (XGBoost), artificial neural network (ANN), support vector machine (SVM), and k-nearest neighbors (KNN). Five-fold cross-validation was used to train and test the models, with external validation performed on the EICU database. Model performance was evaluated using AUROC, recall, and accuracy, with SHAP interpretability employed to understand feature importance. In total, 3871 participants from the MIMIC-IV database and 22,022 participants from the EICU database were analyzed. Hemoglobin, Oxygenation Index, and pH were found to be key indicators of severe traumatic pneumothorax. XGBoost exhibited the best performance, achieving an AUROC of 0.979 (95% CI: [0.966, 0.989]) on the MIMIC-IV dataset and 0.806 (95% CI: [0.740, 0.864]) on the EICU dataset. The results suggest that ML, particularly XGBoost, is faster and more convenient than traditional imaging methods, making it well-suited for emergency or mass casualty situations. ML algorithms show promise for initial diagnosis of traumatic pneumothorax, with XGBoost demonstrating strong interpretability and robust external validation.

**Keywords:** Traumatic Pneumothorax; Interpretable machine learning; XGBoost; Evaluation Model.

# 1 Introduction

According to recent reports, approximately 7.4 males and 1.2 females for every 100,000 persons suffer from pneumothorax [21, 22], and death due to lack of timely treatment or rapid control of various types of pneumothorax accounts for 1.26% of the total annual death rate [26] (male 0.64%, female 0.62%). Traumatic severe pneumothorax is characterized by rapid onset, rapid development, and high mortality caused by delayed treatment [27, 34]. For example, tension pneumothorax is responsible for about 33% of preventable deaths in the combat field [3]. However, such injuries can be quickly resolved with timely and effective treatment. Therefore, the rapid identification and treatment of severe traumatic pneumothorax has become vital for the survival of the wounded.

Traditional pneumothorax diagnostic and evaluation methods include symptom recognition, physical examination, and imaging examination [2]. In medical settings with advanced equipment, traumatic severe pneumothorax can often be treated quickly because of more experienced medical personnel and advanced examination equipment. However, there are many difficulties in the early diagnosis of traumatic severe pneumothorax in other environments, such as in traffic accidents, remote areas that lack medical equipment, and rescue scenes of battlefields or mass casualty events. Therefore, there is an urgent need for a simple, efficient, and accurate traumatic severe pneumothorax assessment tool to assist medical staff in the timely determination of possible high-risk patients in special circumstances.

Traditional medical assessment methods rely on the experience of first responders, some clinical indicators, and various medical equipment. These medical conditions often cannot be assessed outside the hospital or in non-medical settings. There are increasing opportunities for the intersection of artificial intelligence and medical care. With in-depth research using big data technology, many serious injuries have been closely related to the objective core examination index set [17, 18, 35, 40, 43].

Matthew Moll et al. [15] proposed the application of machine learning to predict COPD in all-cause mortality based on clinical characteristics. They found that such methods had better disease prediction than traditional methods for subjects with moderate to severe COPD. Azodi CB et al. [15] showed that interpretable machine learning models have better effects on related research in the medical field, from which we can find more accurate and meaningful data and judgment results.

This study seeks to address two core research questions:

1. Can machine learning models accurately diagnose and evaluate severe traumatic pneumothorax using parameters that are highly related to severe traumatic pneumothorax and can be easily obtained?

2. Which machine learning model offers the most interpretable and clinically applicable insights for traumatic pneumothorax diagnosis?

We hypothesize that AI models will perform well in terms of both accuracy and speed, and that SHAP (SHapley Additive exPlanations) will enhance the interpretability of these models, making them more useful in clinical decision-making.

In this study, we employ a retrospective analysis using the MIMIC-IV and EICU databases to develop machine learning models for traumatic pneumothorax diagnosis. We extracted 33 vital signs and blood gas parameters and selected 12 key indicators to train four machine learning algorithms: XGBoost, artificial neural network (ANN), support vector machine (SVM), and k-nearest neighbors (KNN). The models were trained using five-fold cross-validation and externally validated on the EICU database. Model performance was evaluated based on AUROC, accuracy, recall, etc., with SHAP employed for interpretability. This approach allows us to identify the most significant clinical features contributing to pneumothorax diagnosis while ensuring robust model performance.

# 2 Methods

## 2.1 Participants

Patients with traumatic severe pneumothorax in the existing dataset were selected as the main research objects. The study population was drawn from the MIMIC-IV (Medical Information Mart for Intensive Care) database, a collaborative effort made by emergency physicians, intensivists, and computer science experts at the Massachusetts Institute of Technology (MIT), University of Oxford, Massachusetts General Hospital (MGH), Beth Israel Deacons Medical Center, and other institutions, containing data from 180,733 inpatients from 2008 to 2019 [14]. External validation data were obtained from the EICU database, consisting of more than 200,000 medical records of intensive care unit (ICU) patients used to support medical research. These data can be used to analyze disease treatment, clinical decision-making, and epidemiological studies [30]. The above two datasets are publicly available and have high international recognition.

Based on these two databases, our inclusion criteria were adults 18 years or older with relatively complete demographic and hospital care data. Primary diseases that might have affected the study were excluded, including chronic bronchitis, chronic obstructive pulmonary disease, pulmonary fibrosis, pulmonary emphysema, primary pulmonary hypertension, chronic pneumothorax, cardiac insufficiency, congenital heart disease, atrial septal defect, ventricular septal defect, patent ductus arteriosus, liver and kidney failure, various advanced tumors, pneumothorax caused by previous cardiopulmonary surgery, and pneumothorax caused by chest surgery during hospitalization. We determined the experimental group and control group based on the remaining data. The following criteria were used for group setting and data extraction.

**Experimental group:**

1. Admission to hospital due to trauma;
2. Meets one of the following two points:
   (a) Imaging tips are “Large pneumothorax” (Reference standards of BTS: More than 2 cm from the chest wall to the outer pulmonary edge; or Reference standards of CHEST: More than 3 cm from apex to cupola);
   (b) Treatment for pneumothorax, including thoracentesis aspiration, thoracostomy exhaust, thoracostomy, or thoracostomy drainage.

**Control group:**

1. Non-traumatic wounded;
2. Imaging without pneumothorax-related prompts;
3. Did not receive lung-related treatment or surgery during hospitalization;
4. Wounded are alive at the time of discharge.

**Data extraction:** The cross-sectional data of the experimental and control groups were compared and evaluated. In the experimental group, the cross-section was the imaging examination time or pneumothorax treatment; the cross-section data collection time was earlier than the cross-section, but not by more than one hour. The cross-section of the control group is the time of discharge of the wounded patients, and the cross-section data is the last dataset within 24 hours before discharge.

According to the above criteria, samples with serious missing data or data with outliers were excluded. We obtained a severe traumatic pneumothorax group ($n = 174$) and a non-traumatic non-pneumothorax group ($n = 3,697$) in the MIMIC-IV database, and a severe traumatic pneumothorax group ($n = 105$) and a non-traumatic non-pneumothorax group ($n = 21,917$) in the EICU database. The above data for the pneumothorax group are the experimental groups of the two datasets.

In selecting the 33 features related to pneumothorax diagnosis, we focused on vital signs and blood gas analysis that can be dynamically and continuously monitored in real-time. Traumatic pneumothorax significantly affects patients' vital signs, which can immediately reflect physiological changes after trauma. Blood gas analysis provides rapid information about oxygen and carbon dioxide levels, as well as pH, which are closely related to the occurrence of pneumothorax and are easily accessible in clinical practice. Therefore, the selection of these features not only aligns with the pathological mechanisms of traumatic pneumothorax but also has high clinical operability. Features with more than 50% missing values were excluded (Figure 1), and multiple imputation was used for data imputation.

Multiple imputation (MI) is a robust statistical method used to handle missing data. Unlike traditional methods, such as complete case analysis (which discards observations with missing values) or single imputation (which fills missing values with a single estimate), MI recognizes the uncertainty inherent in missing data. The basic idea is to create multiple different versions of the dataset by imputing missing values multiple times, each imputation drawn from a distribution that reflects the uncertainty of the missing data. These multiple datasets are then analyzed separately, and the results are pooled to give a final estimate [1]. Ultimately, 12 key features were selected as input for the machine learning model.

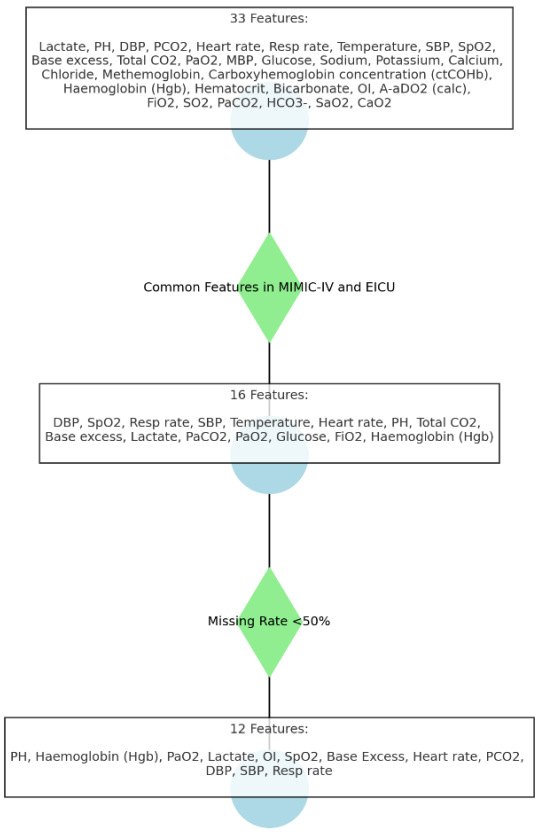



Figure 1: The process of selecting indicators.

To avoid biased model evaluation and improve the robustness of the model, we used the sample balance method. A 1 to 5 ratio of the experimental group to the control group was adopted for downsampling the original data of the extracted control group. The control group data used for training and validation were a MIMIC-IV non-trauma non-pneumothorax group ($n = 870$) and an EICU non-trauma non-pneumothorax group ($n = 525$) (Figure 2).

## 2.2 Analysis techniques

To construct the diagnostic model of Traumatic severe pneumothorax, we collected and extracted demographic characteristics. A total of 33 indicators, including vital signs and blood gas analysis, were examined. According to the significance of individual indicators in the diagnostic evaluation of pneumothorax and the missing data (samples with missing values greater than 50% were excluded [15], and missing values were filled by multiple imputation), 12 key indicators were selected as the input of the machine learning model.

SPSS (version 27.0) software was used to analyze the demographic characteristics and 12 indicators of the Traumatic severe pneumothorax group and non-traumatic non-pneumothorax group at baseline. The Mann-Whiney U test was used to compare the two groups for continuous variables. For categorical variables, the chi-square test was used for statistical analysis. At baseline, P values for most variables were two-sided and considered to indicate a significance of less than 0.05. The difference between the pneumothorax and non-pneumothorax groups was statistically significant.

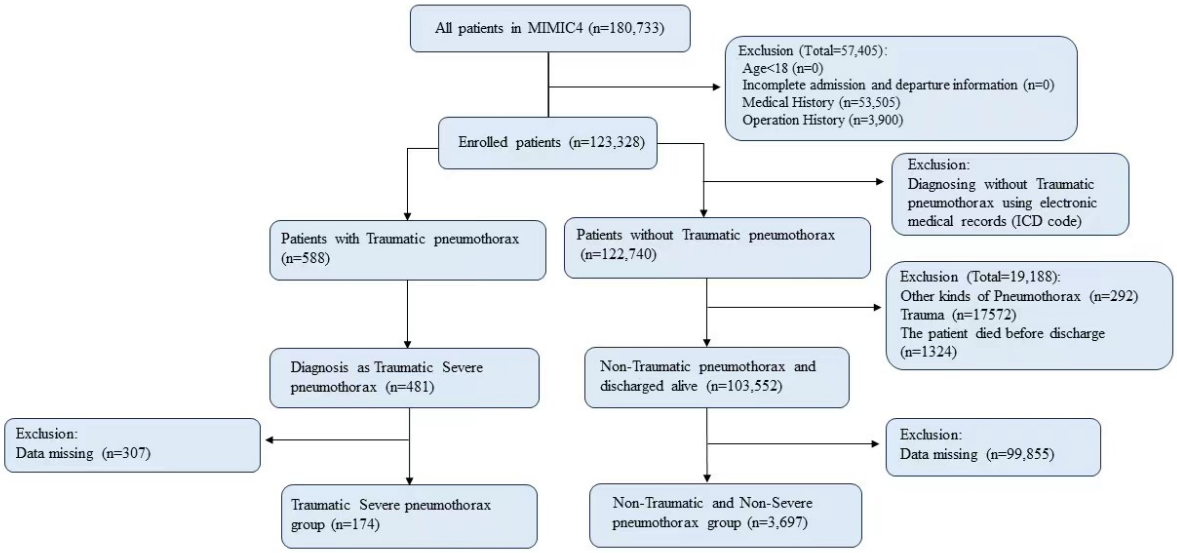
Considering that our research is based on small sample data, we chose XGBoost, ANN, KNN, and SVM for comparison. XGBoost is an algorithm based on enhancing integration, characterized by many optimization terms, such as second-order Taylor formula expansion, regular terms, and a block storage structure [5]. ANN is a neural network algorithm based on the interconnection of artificial neurons [11]. SVM is a multi-classification model whose basic model is a linear classifier defined in the feature space with the largest interval, distinguishing it from the perceptron. The basic idea is to solve the separation hyperplane that can correctly divide the training dataset and has the largest geometric interval [12]. KNN is a supervised classifier model that judges the category to which a new value $x$ belongs according to the category of the nearest $K$ points [6]. XGBoost was chosen for its good interpretability, and ANN, KNN, and SVM were chosen because of their fast, good performance on simple data. The parameters needed to debug the four models were set. Several gradients were set for each parameter, and a hyperparameter grid search was used for parameter tuning. The `sklearn` library in Python was used to build the overall model, and Jupyter Notebook was used as the compiler.

SHAP [23] is a method used to explain how machine learning models work. It belongs to the method of post hoc explanation of the model. Its core idea is to calculate the marginal contribution of features to the model output based on game theory and then explain the "black box model" from the global and local levels. We also performed SHAP analysis to understand some relationship between the metrics and the model. We performed SHAP analysis on the XGBoost model with the best performance, calculated the importance of indicators to the model using the SHAP Value, and analyzed their influence on the model.

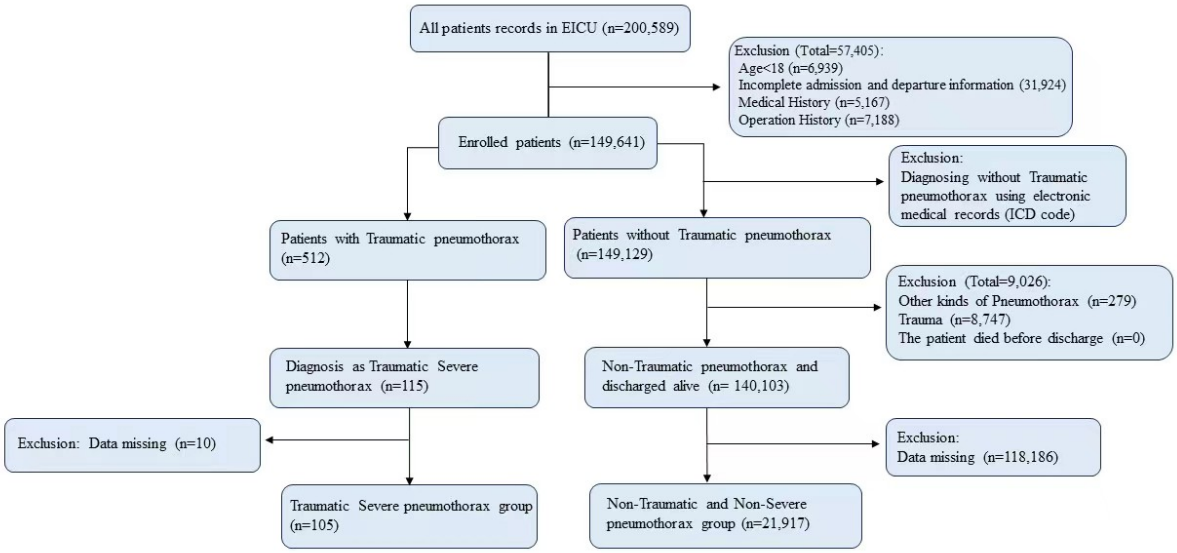
We built ROC curves for each model to evaluate the models and test their differences. AUROC, AUPRC, accuracy, F1.5, Recall, Precision, Sensitivity, Specificity, NPV, PPV, and the Youden index of each model were calculated as evaluation indexes. The Delong test was performed on all models, and it could be concluded that XGBoost was significantly different from the other three models.

## 2.3 Development of The Model

Figure 3 shows a schematic diagram of the study design. We randomly divided 1,044 people from MIMIC-IV into a training set ($n = 835$) and a test set ($n = 209$). The machine learning model received input consisting of 12 filtered features, and the index weights were calculated and sorted based on the Weight method of the XGBoost model.




(a) **(A)**




(b) **(B)**

Figure 2: Extraction process for study cohort: (A) MIMIC-IV; (B) EICU.

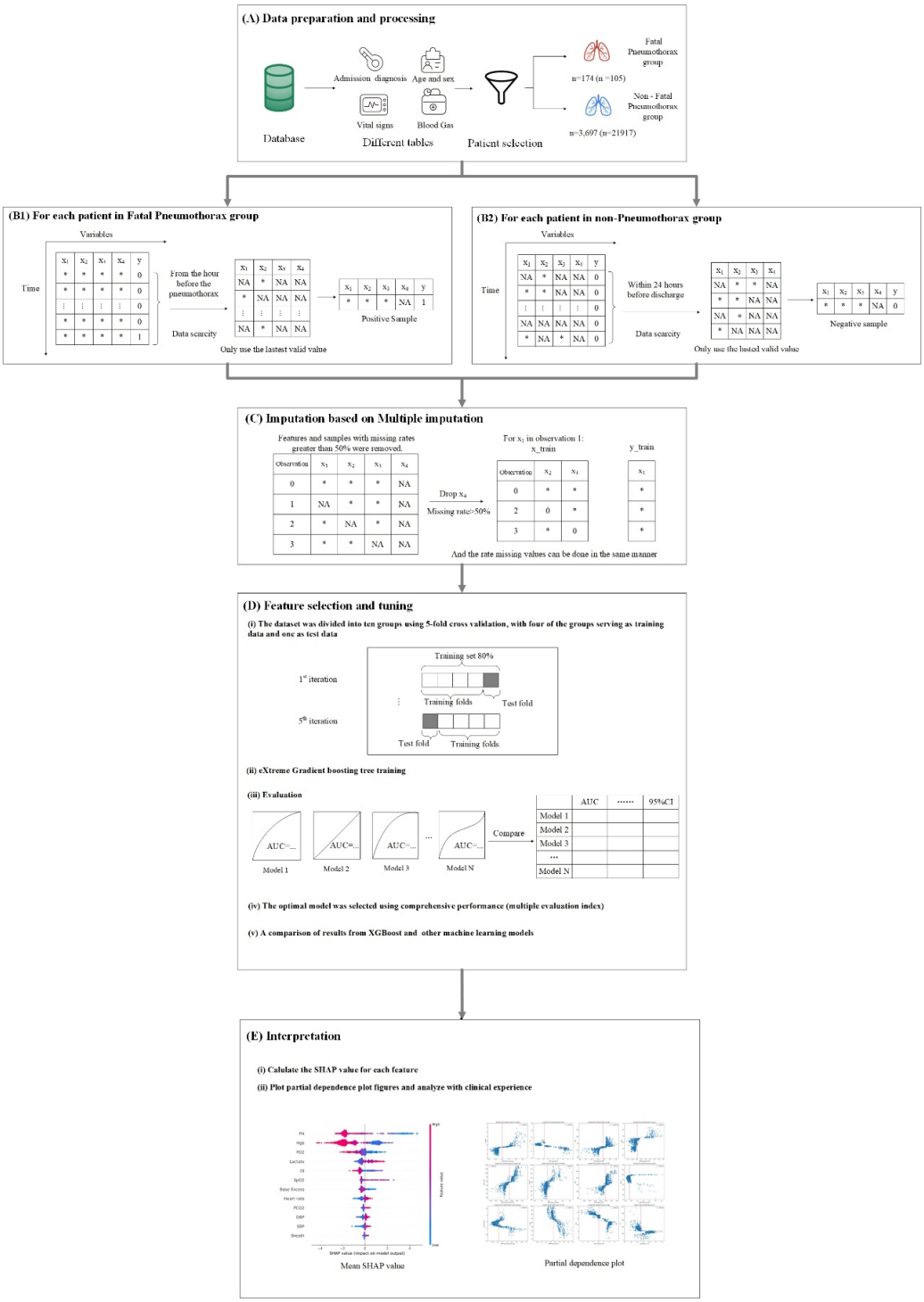



Figure 3: Schematic of study design. (A) Data extraction and processing. Data including diagnosis, demographic information (e.g., sex and race), blood gas analysis, and vital signs were extracted from MIMIC-IV and EICU; (B1) Data extraction in the experimental group; (B2) Data extraction for the control group; (C) Culling and imputation of missing values; (D) XGBoost and other machine learning model training, validation, and testing; (E) Interpretation of machine learning models. Feature importance values were calculated based on the weight method of XGBoost, Overall SHAP dependency plots, and partial dependence plots (PDPs) for 12 features. SHAP: SHapley Additive exPlanations; Hgb: Hemoglobin; OI: Oxygenation Index; SpO2; Oxygen Saturation; pH: Potential of hydrogen; PO2: Partial pressure of oxygen; SBP: Systolic blood pressure; PCO2: Partial pressure of carbon dioxide; DBP: Diastolic blood pressure.

Table 1: P-value of DeLong's test, comparing significant differences in performance of machine learning models.

Delong Test for MIMIC-4

| Model | Z-score | P Value |
|---|---|---|
| XGBoost vs ANN | 3.892 | <0.001 |
| XGBoost vs SVM | 2.072 | <0.001 |
| XGBoost vs KNN | 3.693 | <0.001 |

Delong Test for External Validation

| Model | Z-score | P Value |
|---|---|---|
| XGBoost vs ANN | 4.406 | <0.001 |
| XGBoost vs SVM | 7.867 | <0.001 |
| XGBoost vs KNN | 6.302 | <0.001 |

*XGBoost: eXtreme Gradient Boosting; ANN: Artificial Neural Network; SVM: Support Vector Machine; KNN: k-Nearest Neighbors.*

## 3 Results

The demographic characteristics of the subjects in MIMIC-IV and EICU were relatively similar, with minor differences in age and ethnic distribution (Table 2). Most of the single indexes in the experimental and control groups were statistically significant ($P < 0.05$). Specifically, the comparison of SPO2 and PO2 in the EICU database showed no significant difference, whereas OI exhibited a significant difference between the two groups.

We conducted a detailed comparative analysis of the demographics of the database. First, the demographic baseline of the MIMIC-IV database is presented in Table 2.1. This analysis revealed no significant difference in most variables, except for male sex and some racial categories, supporting the rationality of the sample division. Next, Table 2.2 illustrates the demographic baseline for the EICU database. It shows no significant differences between the experimental and control groups, except for male sex and age, again demonstrating reasonable sample division.

Finally, a comprehensive comparison of demographic characteristics between the MIMIC-IV and EICU databases is provided in Table 2.3. As these datasets are constructed from populations in different regions, certain differences, such as racial composition, are evident. These findings confirm the suitability of the datasets for external validation purposes.

Percentages do not necessarily total 100% because of rounding. P value from Student t-test (for continuous data), $\chi^2$ (for categorical data). P value in **boldface** if $< 0.05$. In Table 3, we observe that among the four machine learning classifier models, the multiple evaluation metrics of XGBoost outperform the other three models, with 94.7% accuracy and 73.0% recall rate. The AUROC of all four models exceeded 90%, with XGBoost achieving an AUROC of 0.979 [0.966–0.989].

To verify the externality of the model's performance, we took the MIMIC data as the training set and the EICU data as the test set. According to an 80:20 ratio, the training and test sets were screened, and the model was verified. XGBoost still performed better in the external validation. The ROC curves of the four machine learning models under the two datasets were compared (Figure 4). In general, the AUC value of XGBoost always ranked first, and the other three models performed less well than XGBoost on the external data.

Figure 5 shows the relevance of SHAP interpretability, with red indicating high impact and blue indicating low impact. It can be seen that the influence of PH, Hgb, and PO2 are the top three factors for determining "is pneumothorax." This shows that these three indicators also have a significant impact on the model output. At the same time, it can be seen that the indicators of evaluation and monitoring of human respiratory and metabolic status after lung injury will change significantly, which has a more significant impact on the model output results, and the change of its influence with the value is also in line with clinical common sense.

Table 2: Demographic Characteristics of Subjects in MIMIC-4 and EICU Included in Analysis

Table2.1 Baseline tables of the EICU populations

| Variable | MIMIC-4 Dataset | | P Value |
|---|---|---|---|
| | Experiment group | Control Group | |
| No. of subjects | 174 | 3,697 | |
| Age, mean(SD), y | 55.3 (21.9) | 53.5 (19.3) | 0.306 |
| BMI, mean(SD), kg/m$^2$ | 25.9 (6.5) | 28.2 (6.9) | 0.006 |
| Male sex | 123 (70.7%) | 1,856 (50.2%) | <0.001 |
| Race | | | |
| Caucasian | 105 (60.3%) | 2,380 (64.4%) | 0.278 |
| African American | 9 (5.2%) | 408 (11%) | 0.015 |
| Asian | 1 (1%) | 119 (3.2%) | 0.049 |
| Other | 59 (33.9%) | 790 (21.4%) | <0.001 |

Table2.2 Baseline table of the MIMIC population

| Variable | EICU Dataset | | P Value |
|---|---|---|---|
| | Experiment group | Control Group | |
| No. of subjects | 105 | 21,917 | |
| Age, mean(SD), y | 52.4 (19.2) | 62.8 (15.8) | <0.001 |
| BMI, mean(SD), kg/m$^2$ | 27.1 (6.4) | 29.6 (8.6) | 0.002 |
| Male sex | 81 (77.1%) | 12,172 (55.5%) | <0.001 |
| Race | | | |
| Caucasian | 79 (75.2%) | 17,122 (78.1%) | 0.476 |
| African American | 16 (15.2%) | 2191 (9.9%) | 0.074 |
| Asian | 4 (3.8%) | 247 (1.1%) | 0.009 |
| Other | 6 (5.7%) | 2,357 (10.8%) | 0.096 |

Table2.3 Baseline table for MIMIC and EICU population

| Variable | MIMIC-4 | EICU | P Value |
|---|---|---|---|
| No. of subjects | 3,871 | 22,022 | |
| Age, mean(SD), y | 53.6 (19.4) | 62.7 (15.8) | <0.001 |
| BMI, mean(SD), kg/m$^2$ | 28.1 (6.9) | 29.6 (8.5) | <0.001 |
| Male sex | 1,979 (51.1%) | 12,253 (55.6%) | <0.001 |
| Race | | | |
| Caucasian | 2,485 (64.2%) | 17,201 (78.1%) | <0.001 |
| African American | 417 (10.8%) | 2,207 (10%) | 0.154 |
| Asian | 120 (3.1%) | 251 (1.1%) | <0.001 |
| Other | 849 (21.9%) | 2,363 (10.7%) | <0.001 |

Table 3: Table 3 - Diagnostic Performance of Machine Learning Models in MIMIC-IV and EICU External Validation

| Model | Dataset | Accuracy | F1 Score | Precision | Recall | AUROC (95% CI) | AUPRC | Sensitivity | NPV | PPV |
|---|---|---|---|---|---|---|---|---|---|---|
| XGBoost | MIMIC-IV | 0.947 | 0.822 | 0.941 | 0.730 | 0.979 (0.966, 0.989) | 0.926 | 0.991 | 0.730 | 0.941 |
| ANN | MIMIC-IV | 0.882 | 0.529 | 0.793 | 0.397 | 0.915 (0.893, 0.936) | 0.717 | 0.979 | 0.397 | 0.793 |
| SVM | MIMIC-IV | 0.934 | 0.761 | 0.957 | 0.632 | 0.967 (0.946, 0.985) | 0.911 | 0.994 | 0.632 | 0.957 |
| KNN | MIMIC-IV | 0.943 | 0.816 | 0.891 | 0.753 | 0.948 (0.923, 0.969) | 0.900 | 0.982 | 0.753 | 0.891 |
| XGBoost | 80%MIMIC-IV+20%EICU | 0.718 | 0.512 | 0.360 | 0.886 | 0.806 (0.740, 0.864) | 0.359 | 0.886 | 0.684 | 0.967 |
| ANN | 80%MIMIC-IV+20%EICU | 0.627 | 0.426 | 0.287 | 0.829 | 0.703 (0.628, 0.780) | 0.238 | 0.829 | 0.586 | 0.944 |
| SVM | 80%MIMIC-IV+20%EICU | 0.636 | 0.415 | 0.284 | 0.771 | 0.705 (0.621, 0.780) | 0.247 | 0.771 | 0.609 | 0.930 |
| KNN | 80%MIMIC-IV+20%EICU | 0.679 | 0.247 | 0.204 | 0.314 | 0.655 (0.565, 0.740) | 0.215 | 0.314 | 0.753 | 0.845 |

XGBoost: eXtreme Gradient Boosting; ANN: Artificial Neural Network; SVM: Support Vector Machine; KNN: k-Nearest Neighbors; AUROC: Area Under Receiver Operating Characteristic Curve; AUPRC: Area Under Precision-Recall Curve; NPV: Negative Predictive Value; PPV: Positive Predictive Value.

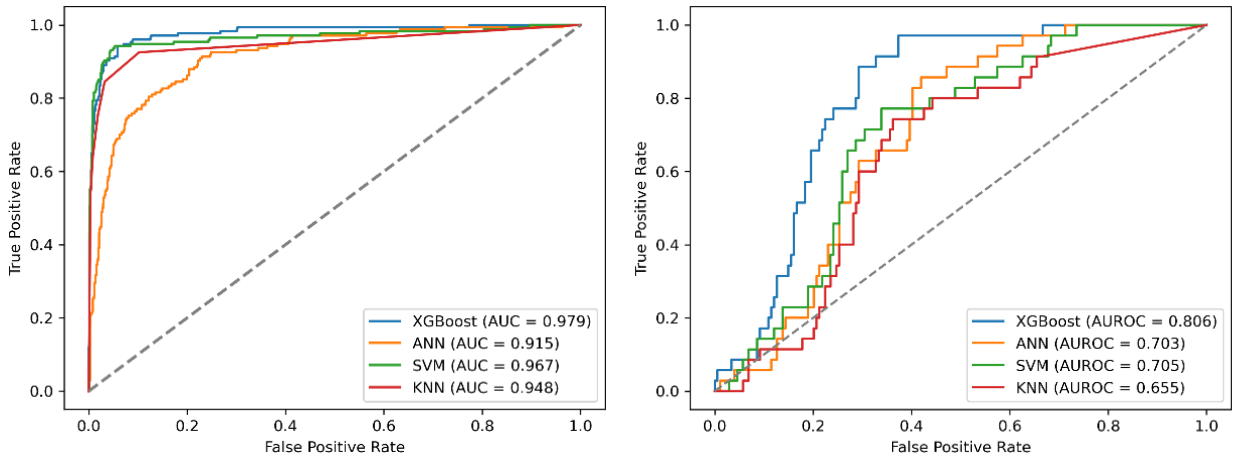



Figure 4: Receiver operating characteristic curve in MIMIC-IV and EICU external validation. (A) MIMIC-IV has 80% for training (n = 835), 20% for testing (n = 209); (B) External Test has 80% for training (n = 835 from MIMIC), 20% for testing (n = 209 from EICU). XGBoost: eXtreme Gradient Boosting; ANN: Artificial Neural Network; SVM: Support Vector Machine; KNN: k-Nearest Neighbors; AUROC: area under receiver-operating curve.

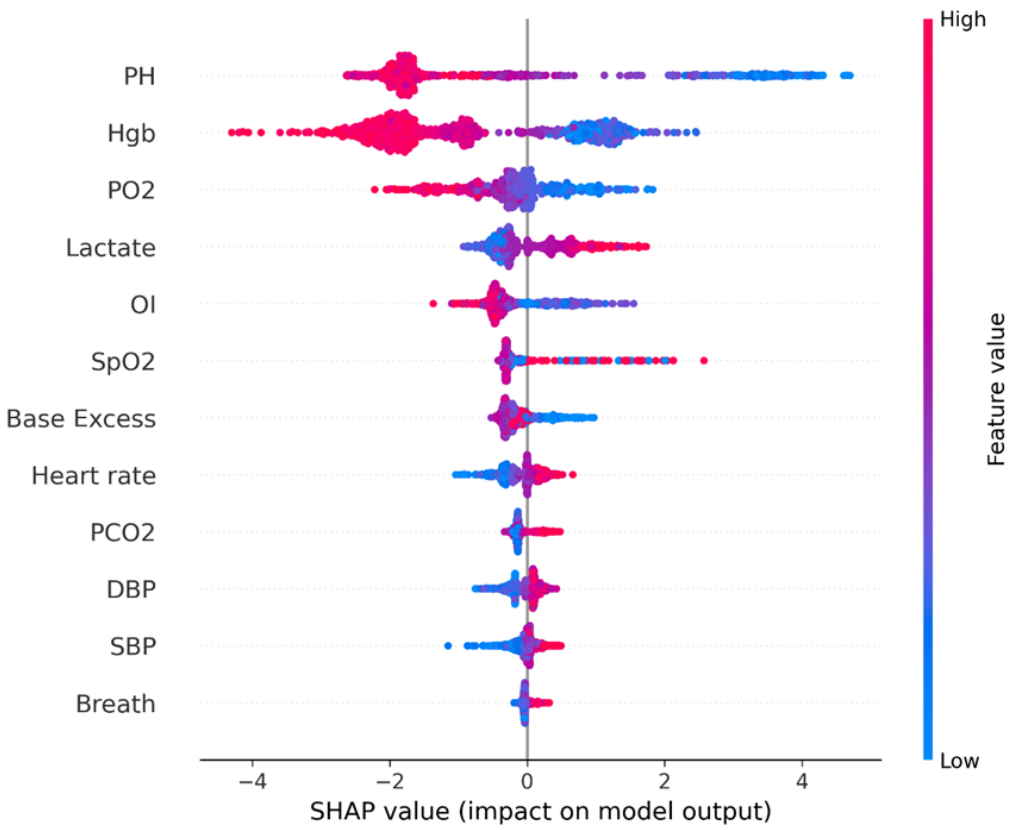



Figure 5: SHAP summary plot of each metric in the XGBoost model. XGBoost: eXtreme Gradient Boosting; SHAP: SHapley Additive exPlanations; Hgb: Hemoglobin; OI: Oxygenation Index; SpO2: Oxygen Saturation; pH: Potential of hydrogen; PO2: Partial pressure of oxygen; SBP: Systolic blood pressure; PCO2: Partial pressure of carbon dioxide; DBP: Diastolic blood pressure.

In the overall dependence graph of SHAP, the importance of features is reflected by the ranking, and the effect direction of features can also be reflected [24]. To determine how individual features affect the results of the evaluation model, we plotted the SHAP partial dependence of individual features of the XGBoost model (Figure 6). In the single-feature SHAP dependence plot, the X-axis represents the SHAP value of the feature, and the Y-axis represents the value of the feature. The SHAP value of a single feature exceeding zero indicates a positive contribution to evaluating and diagnosing traumatic severe pneumothorax.

This study involved two main characteristics of the experimental group, namely, trauma and hypoxia caused by severe traumatic pneumothorax, which complement each other and lead to blood loss and microcirculation disorders of the wounded, manifested as decreased blood *Hgb*, decreased *pH* value, increased *Lactate*, and increased negative *Base Excess* value. Hypoxia can lead to decreased *PO2* and increased *PCO2* of the wounded. *SpO2* decreased (a small number of wounded patients had a great impact on this index due to oxygen inhalation in the hospital and other reasons, resulting in the index being in the normal range), and *OI* decreased. Combined with the above factors and the pain after trauma, the vital signs of the wounded can be manifested as increased heart rate, increased *SBP* and *DBP*, and increased respiratory rate (Figure 6).

## 4 Discussion

In this study, a variety of scenarios for the definitive diagnosis of pneumothorax are mentioned in the above content, which are not suitable for the application of large medical imaging equipment. In such scenarios, there is a clear contradiction between the urgency of evaluation and diagnosis of traumatic pneumothorax and medical conditions. The pneumothorax assessment strategy provided in this study relies only on vital signs and blood gas analysis indicators. With the miniaturization and non-invasive development of indicator acquisition devices, the types of vital signs collected by

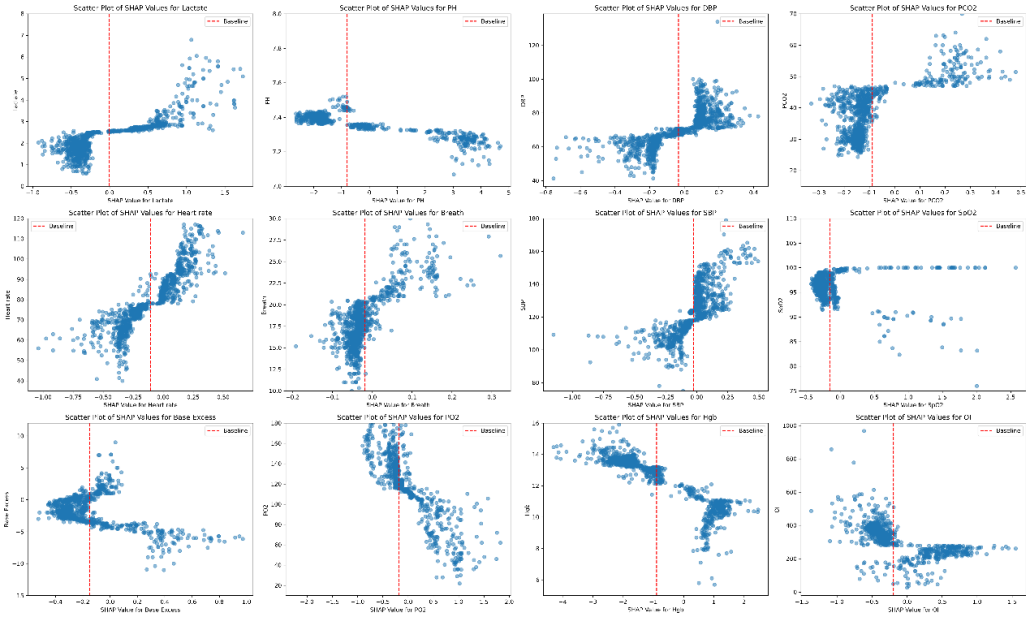


Figure 6: Partial SHAP dependence plots for 12 features of blood gas analysis and vital signs. SHAP: SHapley Additive exPlanations; Hgb: Hemoglobin; OI: Oxygenation Index; SpO2: Oxygen Saturation; pH: Potential of hydrogen; PO2: Partial pressure of oxygen; SBP: Systolic blood pressure; PCO2: Partial pressure of carbon dioxide; DBP: Diastolic blood pressure. Baseline: average impact of each metric on model output.

wearable vital sign acquisition devices fully cover the requirements of this study, and can be monitored in real time. The small non-invasive blood gas analyzer can also conveniently collect all the blood gas analysis indicators needed in this study. With the gradual application and continuous promotion of such equipment, the research results will surely have stronger practical advantages in complex scenes, especially in the case of mass incidents.

At present, the application of various artificial intelligence technologies in the medical field is still in the development stage. The results of this study aim to provide an auxiliary assessment tool for severe traumatic pneumothorax. In practical clinical work, the diagnosis of traumatic pneumothorax needs to rely on the chief complaint of the wounded (including local pain, dyspnea, etc.), physical examination signs (chest percussion, etc.), and even diagnostic puncture of pneumothorax. Non-invasive definitive diagnosis relies on imaging. The implementation of the former is affected by the consciousness state of the wounded, and the accuracy of the assessment is limited by the technical level of on-site medical staff and medical devices. Our study is just to provide a simple and objective assessment tool to assist medical staff to complete the final comprehensive judgment combined with field conditions. With the supplement of data volume and the improvement of analysis technology, its auxiliary role will be gradually enhanced.

We used clinically available vital signs and blood gas analysis indicators to construct high-performance traumatic severe pneumothorax evaluation models. We applied *XGBoost*, *KNN*, *ANN*, and *SVM* to classify patients. The results showed that *XGBoost* had the best effect ($AUROC = 0.979$, 95% CI [0.966–0.989]). Its accuracy and recall rate ranked first compared with the other three machine learning models. According to the learning curve of each model, there was no significant difference in the performance of our model between the training and test sets, indicating that it did not overfit. With the application of an advanced medical-grade noninvasive blood gas analyzer or bedside blood gas analyzer, blood gas analysis index results can be repeatedly and dynamically obtained [28], so the aging assessment and dynamic monitoring of pneumothorax can be obtained.

We applied different datasets in the research and external validation of the model. Johannes Hofmanninger et al. [13] employed external validation verification. Comparing their work with other studies' verification, externality verification will make the research more medically meaningful and the results more universal and convincing. Other studies show that internal validation alone cannot prove the stable performance of the model [7]. External verification demonstrates the robustness and stability of our training model and shows the strong universality of our research among different

individuals.

Our study highlights interpretability based on previous studies, which is of great significance in the clinical application of the model. The use of SHAP values can make the results of our machine learning algorithm interpretable, which can help doctors understand the evaluation basis of the model better and facilitate more accurate diagnosis. By using SHAP, we can obtain the feature contribution ranking and the contribution of individual features to evaluate specific injuries. In the diagnosis of traumatic severe pneumothorax, this means that the doctors can understand how each clinical feature influences the diagnostic outcome. As seen in Figure 6, the top three SHAP features of the XGBoost model were pH, Hgb, and PO2, which further reflect the obvious characteristics of trauma, blood loss, and hypoxia in traumatic pneumothorax. The analysis of individual indicators by SHAP plots is also consistent with general clinical cognition.

The research on the combination of artificial intelligence technology and pneumothorax can be traced back to the research related to pneumothorax images by Sanada's team [36]. However, at that time, the effect was not as good as expected because of the limitation of technology. We can see more and more researches on machine learning and clinical application. Ali Reza Khoshdel's team and Chiao-Chin Lee's team [25] combined machine learning with pneumothorax prediction to predict the occurrence of pneumothorax. Rajpurkar [32], Rubin [33], Yuchi Tian [39], Bharati [4], Wang Yaqi [41], Jakhar [16], and Sebastian [13] conducted research on the diagnosis of pneumothorax by computer vision and medical imaging, and Song Yang [44], Yanfeng Zhao [45], and Saibin Wang [42] applied machine learning algorithms to study pneumothorax prediction based on patient data. The latter approach is similar to our team's; their models performed at or near our level, but few teams specialized in data mining research on traumatic pneumothorax. At the same time, with the emergence of interpretable machine learning, there are some examples such as Jabal Mohamed Sobhi's team [37] in the field of ischemic Stroke Outcome Prediction, Diptesh Das [8] in the field of diagnosis of Alzheimer's disease, Interpretable and clinically relevant research by Michael Fanton's team [10] in the field of predicting the optimal day of trigger during ovarian stimulation. The ideas of these studies have given some inspiration to the study we are conducting now. Kuo, W [20] et al. conducted a review study.

Traditional diagnostic methods rely on medical staff for the patient's assessment, physical examination and imaging, of which imaging is the most important diagnostic method. Its advantage lies in that it can obtain the highest accuracy rate (almost 100%) through imaging examination, and at the same time, it can evaluate the location and severity of pneumothorax. For example, in this study, one of the most important screening bases for patients with severe pneumothorax is the imaging report of the patients during their hospital stay. However, this diagnosis method relies on imaging equipment, which seriously restricts its use environment and takes a long time.

Compared with the traditional diagnostic evaluation method of pneumothorax, the model has stronger environmental adaptability, timeliness, and repeatability. This is especially useful in Settings such as battlefields or disaster relief sites. The assessment method does not rely on large instruments and equipment and can achieve synchronous monitoring in mass medical emergencies. Vital signs and blood gas analysis are more readily available. Through wearable devices, vital signs can be dynamically monitored in real time.

Compared with traditional medical imaging studies, ours has the following advantages. First, our model does not need to rely on medical imaging equipment. According to Bharati, the medical imaging method using Basic CapsNet takes about 1815 seconds (only model training time), and it requires a high level of medical equipment, which is unsuitable for areas that lack medical tools [4]. Second, in recent years, hand-held ultrasound diagnostic equipment has brought great convenience and high accuracy to diagnosing pneumothorax in special environments. However, how to conduct group evaluation and monitoring is still an urgent problem to be solved, which our model can overcome to a certain extent. Third, our model has a short running time, data acquisition is convenient, and data can be dynamically and repeatedly evaluated, facilitating the determination of whether a person suffers from Traumatic severe pneumothorax within the golden rescue time.

Our study has some limitations. Firstly, Due to the limitation of the accuracy of the current model, in order to prevent the missed diagnosis in the application of clinical medicine, we need to improve the recall, because it may cause the occurrence of false positives in the application process.

With the continuous improvement of the model, this problem can be gradually solved. Second, vital signs and blood gas analysis indicators are not the gold standard for diagnosing severe traumatic pneumothorax in clinical practice. For example, large-area pulmonary contusion can sometimes show similar index values and trends to severe traumatic pneumothorax. However, the model can still provide auxiliary decision-making for severe or fatal pulmonary trauma. If relevant evidence, such as injury factors and clinical symptoms, is supplemented in later studies, satisfactory evaluation results can still be obtained. Third, compared with other diseases or trauma, pneumothorax is still a niche area in clinical practice, especially when the scope of research is limited to more severe traumatic pneumothorax. Therefore, this study was limited by the sample size, and related studies have the same problem [19, 29, 31]. However, the ideas and methods of this study are also applicable to large sample data, and with the continuous accumulation of data, this problem can be solved.

The model is constructed by MIMIC database, and directly applying it to another database itself will lead to a decline in the performance of the model. At the same time, by analyzing other reasons, we find that there is some difference in the demography of the wounded extracted from the two groups of databases. What's more? the ethnicity of the two databases is significantly different. This is the objective reason for the decrease in performance. In general, the performance of external validation can prove that our model is very robust.

## 5 Conclusion

We constructed an evaluation model for severe traumatic pneumothorax based on the XGBoost algorithm and verified its performance through external validation. It has wide application prospects and can assist medical staff in realizing the dynamic pneumothorax assessment of mass casualties in various special environments. The overall research ideas and methods can provide a basis for solutions for many other medical fields. Our algorithm has achieved good performance in both internal and external validation, and the results of the diagnostic model can also provide some decision support for clinical diagnosis. However, prospective clinical validation is still needed to determine whether the model can accurately identify traumatic pneumothorax in actual clinical scenarios, and we can include more data in different ethnic groups for future research.

### Funding

This research was supported by the Beijing Natural Science Foundation, grant number M22023 and the National Key Research and Development Plan grant number 2022YFC3006202.

### Ethic

This study has obtained ethics approval from the Ethics Committee Of Chinese PLA General Hospital review board S2020-129-02.

### Author contributions

The authors contributed equally to this work.

### Conflict of interest

The authors declare no conflict of interest.

## References

[1] Austin, Peter C., et al. "Missing data in clinical research: a tutorial on multiple imputation." *Canadian Journal of Cardiology* 37.9 (2021): 1322-1331.

[2] Anderson DE, Kocik VI, Rizzo JA, et al. A Narrative Review of Traumatic Pneumothorax Diagnoses and Management. *Med J (Ft Sam Houst Tex)*. 2023;(Per 23-1/2/3):3-10.

[3] Beckett A, Savage E, Pannell D, Acharya S, Kirkpatrick A, Tien HC. Needle decompression for tension pneumothorax in Tactical Combat Casualty Care: do catheters placed in the midaxillary line kink more often than those in the midclavicular line?. *J Trauma*. 2011;71(5 Suppl 1):S408-S412. doi:10.1097/TA.0b013e318232e558.

[4] Bharati S, Podder P, Mondal MRH. Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*. 2020;20:100391.

[5] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Published online 2016. doi:10.1145/2939672.2939785.

[6] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967;13(1):21-27. doi:10.1109/TIT.1967.1053964.

[7] Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40. doi:10.1186/1471-2288-14-40.

[8] Das D, Ito J, Kadowaki T, Tsuda K. An interpretable machine learning model for diagnosis of Alzheimer's disease. *PeerJ*. 2019;7:e6543. doi:10.7717/peerj.6543.

[9] Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Datasets. *Computational Intelligence*. 2004;20(1):18-36. doi:10.1111/j.0824-7935.2004.t01-1-00228.x.

[10] Fanton M, Nutting V, Solano F, et al. An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation. *Fertil Steril*. 2022;118(1):101-108. doi:10.1016/j.fertnstert.2022.04.003.

[11] Hoehfeld M, Fahlman SE. Learning with limited numerical precision using the cascade-correlation algorithm. *IEEE Transactions on Neural Networks*. 1992;3(4):602-611. doi:10.1109/72.143374.

[12] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*. 1998;13(4):18-28. doi:10.1109/5254.708428.

[13] Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*. 2020;4:50-50.

[14] [Online] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. MIMIC-IV v2.0. June 12, 2022. Accessed September 21, 2023. Available at: https://physionet.org/content/mimiciv/2.0/.

[15] Jakobsen JC, Gluud C, Wetterslev J, et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17:162. doi:10.1186/s12874-017-0442-1.

[16] Jakhar K, Kaur A, Gupta M. Pneumothorax Segmentation: Deep Learning Image Segmentation to predict Pneumothorax. *arXiv.org*. Ithaca: Cornell University Library; 2021.

[17] Kapoor D, Xu C. Spinal Cord Injury AIS Predictions Using Machine Learning. *eNeuro*. 2023;10(1):ENEURO.0149-22.2022. Published 2023 Jan 4. doi:10.1523/ENEURO.0149-22.2022.

[18] Khairuddin MZF, Lu Hui P, Hasikin K, et al. Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *Int J Environ Res Public Health*. 2022;19(21):13962. Published 2022 Oct 27. doi:10.3390/ijerph192113962.

[19] Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. *Clin Imaging*. 2020;61:15-19. doi:10.1016/j.clinimag.2020.01.008.

[20] Kuo, W., et al. "Artificial Intelligence in Lung Disease Detection: A Review." *Journal of Thoracic Imaging* 35.3 (2020): 147-156.

[21] Levine DJ, Sako EY, Peters J. *Fishman's Pulmonary Diseases and Disorders*. 4th ed. McGraw-Hill; 2008. ISBN 978-0-07-145739-2. Page 1519.

[22] Light RW. *Pleural Diseases*. 5th ed. Lippincott Williams & Wilkins; 2007. ISBN 978-0-7817-6957-0. Page 306.

[23] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*. 2017;30.

[24] Lundberg SM, Erion GG, Lee SI. Consistent Individualized Feature Attribution for Tree Ensembles. *ArXiv*. 2018; abs/1802.03888.

[25] Lee CC, Lin CS, Tsai CS, et al. A deep learning-based system capable of detecting pneumothorax via electrocardiogram. *Eur J Trauma Emerg Surg*. 2022;48:3317–3326. doi:10.1007/s00068-022-01904-3.

[26] MacDuff A, Arnold A, Harvey J, et al. (BTS Pleural Disease Guideline Group). Management of spontaneous pneumothorax: British Thoracic Society Pleural Disease Guideline 2010. *Thorax*. 2010;65(8 Suppl 2):ii18-ii31. doi:10.1136/thx.2010.136986. PMID: 20696690.

[27] Nelson D, Porta C, Satterly S, et al. Physiology and cardiovascular effect of severe tension pneumothorax in a porcine model. *J Surg Res*. 2013;184(1):450-457. doi:10.1016/j.jss.2013.05.057.

[28] Nielsen MB, Cantisani V, Sidhu PS, et al. The Use of Handheld Ultrasound Devices - An EFSUMB Position Paper [published correction appears in *Ultraschall Med.* 2019 Feb;40(1):e1]. *Ultraschall Med.* 2019;40(1):30-39. doi:10.1055/a-0783-2303.

[29] Olthof AW, Shouche P, Fennema EM, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Comput Methods Programs Biomed*. 2021;208:106304. doi:10.1016/j.cmpb.2021.106304.

[30] [Online] Pollard T. eICU. eICU Collaborative Research Database. Accessed September 21, 2023. Available at: https://eICU-crd.mit.edu/.

[31] Park S, Lee SM, Kim N, et al. Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol*. 2019;29(10):5341-5348. doi:10.1007/s00330-019-06130-x.

[32] Rajpurkar, P., et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *arXiv preprint arXiv:1711.05225* (2017).

[33] Rubin, J., et al. "Classifying Pneumothorax in Chest X-Rays Using Convolutional Neural Networks." *Journal of Digital Imaging* 31 (2018): 379-384.

[34] Saiphoklang N, Kanitsap A. Prevalence, clinical manifestations and mortality rate in patients with spontaneous pneumothorax in Thammasat University Hospital. *J Med Assoc Thai*. 2013;96(10):1290-1297.

[35] Santos K, Dias JP, Amado C. A literature review of machine learning algorithms for crash injury severity prediction. *J Safety Res*. 2022;80:254-269. doi:10.1016/j.jsr.2021.12.007.

[36] Sanada S, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography: Automated detection of pneumothorax in chest images. *Med Phys*. 1992;19:1153-1160. doi:10.1118/1.596790.

[37] Sobhi JM, Joly O, Kallmes D, et al. Interpretable Machine Learning Modeling for Ischemic Stroke Outcome Prediction. *Front Neurol*. 2022;13:884693. doi:10.3389/fneur.2022.884693.

[38] Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: Experimental evaluation. *Information Sciences*. 2020;513:429-441. doi:10.1016/j.ins.2019.11.004.

[39] Tian Y, Wang J, Yang W, Wang J, Qian D. Deep multi-instance transfer learning for pneumothorax classification in chest X-ray images. *Med Phys*. 2022;49:231–243. doi:10.1002/mp.15328.

[40] Vallmuur K. Machine learning approaches to analysing textual injury surveillance data: a systematic review. *Accid Anal Prev*. 2015;79:41-49. doi:10.1016/j.aap.2015.03.018.

[41] Wang Y, Sun L, Jin Q. Enhanced Diagnosis of Pneumothorax with an Improved Real-Time Augmentation for Imbalanced Chest X-rays Data Based on DCNN. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2021;18:951-962.

[42] Wang S, Tu J, Chen W. Development and Validation of a Prediction Pneumothorax Model in CT-Guided Transthoracic Needle Biopsy for Solitary Pulmonary Nodule. *Biomed Res Int*. 2019;2019:7857310. doi:10.1155/2019/7857310.

[43] Ye G, Balasubramanian V, Li JK, Kaya M. Machine Learning-Based Continuous Intracranial Pressure Prediction for Traumatic Injury Patients. *IEEE J Transl Eng Health Med*. 2022;10:4901008. Published 2022 Jun 2. doi:10.1109/JTEHM.2022.3179874.

[44] Yang S, Lou L, Wang W, et al. Pneumothorax prediction using a foraging and hunting based ant colony optimizer assisted support vector machine. *Comput Biol Med*. 2023;161:106948. doi:10.1016/j.compbiomed.2023.106948.

[45] Zhao Y, Wang X, Wang Y, Zhu Z. Logistic regression analysis and a risk prediction model of pneumothorax after CT-guided needle biopsy. *J Thorac Dis*. 2017;9(11):4750-4757. doi:10.21037/jtd.2017.09.47.


Journal's webpage: http://univagora.ro/jour/index.php/ijccc/


This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

*Cite this paper as:*

Lv, Y., Weng, J., Li, J., Chen, W., Huang, H., Zhao, Y. (2025). An interpretable machine learning based model for traumatic severe pneumothorax evaluation, *International Journal of Computers Communications & Control*, 20(1), 6830, 2025.
https://doi.org/10.15837/ijccc.2025.1.6830