Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Integrating ensemble and machine learning models for early prediction of pneumonia mortality using laboratory tests

Seung Min Baik [a], Kyung Sook Hong [b], Jae-Myeong Lee [c], Dong Jin Park [d,*]

[a] Division of Critical Care Medicine, Department of Surgery, Ewha Womans University Mokdong Hospital, Ewha Womans University College of Medicine, Seoul, South Korea
[b] Division of Critical Care Medicine, Department of Surgery, Ewha Womans University Seoul Hospital, Ewha Womans University College of Medicine, Seoul, South Korea
[c] Department of Acute Care Surgery, Korea University Anam Hospital, Seoul, South Korea
[d] Department of Laboratory Medicine, Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea

ARTICLE INFO

Keywords:
Artificial intelligence
Laboratory test
Pneumonia
Mortality
Prediction

ABSTRACT

Background: The recent use of artificial intelligence (AI) in medical research is noteworthy. However, most research has focused on medical imaging. Although the importance of laboratory tests in the clinical field is acknowledged by clinicians, they are undervalued in medical AI research. Our study aims to develop an early prediction AI model for pneumonia mortality, primarily using laboratory test results.

Materials and methods: We developed a mortality prediction model using initial laboratory results and basic clinical information of patients with pneumonia. Several machine learning (ML) models and a deep learning method—multilayer perceptron (MLP)—were selected for model development. The area under the receiver operating characteristic curve (AUROC) and F1-score were optimized to improve model performance. In addition, an ensemble model was developed by blending several models to improve the prediction performance. We used 80,940 data instances for model development.

Results: Among the ML models, XGBoost exhibited the best performance (AUROC = 0.8989, accuracy = 0.88, F1-score = 0.80). MLP achieved an AUROC of 0.8498, accuracy of 0.86, and F1-score of 0.75. The performance of the ensemble model was the best among the developed models, with an AUROC of 0.9006, accuracy of 0.90, and F1-score of 0.81. Several laboratory tests were conducted to identify risk factors that affect pneumonia mortality using the "Feature importance" technique and SHapley Additive exPlanations. We identified several laboratory results, including systolic blood pressure, serum glucose level, age, aspartate aminotransferase-to-alanine aminotransferase ratio, and monocyte-to-lymphocyte ratio, as significant predictors of mortality in patients with pneumonia.

Conclusions: Our study demonstrates that the ensemble model, incorporating XGBoost, CatBoost, and LGBM techniques, outperforms individual ML and deep learning models in predicting pneumonia mortality. Our findings emphasize the importance of integrating AI techniques to leverage laboratory test data effectively, offering a promising direction for advancing AI applications in medical research and clinical decision-making.

* Corresponding author. 1021, Tongil-ro, Eunpyeong-gu, Seoul, 03312, South Korea.
E-mail address: parkdj1280@gmail.com (D.J. Park).

https://doi.org/10.1016/j.heliyon.2024.e34525
Received 10 August 2023; Received in revised form 9 July 2024; Accepted 10 July 2024
Available online 14 July 2024

1. Introduction

Pneumonia is generally recognized as a disease with high morbidity. According to a report from OECD countries, pneumonia accounts for 30 % of all respiratory deaths [1]. Pneumonia was the ninth leading cause of death in the United States in 2020, and it has been demonstrated that COVID-19 contributed significantly to this statistic [2]. However, it is important to recognize that COVID-19 pneumonia has unique characteristics that may not fully represent the full spectrum of pneumonia cases. This distinction is critical in interpreting the findings and applying them to a broader range of pneumonia types. Furthermore, early assessment of the severity of pneumonia, such as through mortality prediction, can help ensure that patients receive optimal care and that limited healthcare resources are allocated appropriately. In fact, when the COVID-19 pandemic was declared, an unprecedented medical overload occurred worldwide, resulting in many casualties as patients requiring immediate intensive care were not admitted to intensive care units in a timely manner [3,4].

Recent advances in artificial intelligence (AI) have had a significant impact on medical research, notably on imaging data such as x-rays, computed tomography (CT), and magnetic resonance imaging [5,6]. While these imaging tests are highly accurate and informative, their frequency is limited owing to potential side effects associated with patient transportation [7]. Laboratory tests, by contrast, are easily accessible, frequently used, and can provide an initial critical assessment of a patient's condition prior to imaging. Early laboratory testing offers several distinct advantages in terms of predicting pneumonia mortality. First, laboratory tests are generally accessible and can be performed quickly, providing timely insight into the patient's physiologic status. This immediacy is useful in critical care settings, where rapid decision-making can have a significant impact on patient outcomes. Second, because laboratory tests cover a wide range of physiologic parameters, from blood counts to indicators of organ function, they provide a comprehensive overview of a patient's health status. This breadth of data can be extremely useful for risk stratification for premature death. Third, laboratory tests are less invasive and less risky than other imaging tests, making them more suitable for repeated assessment to monitor disease progression. Despite these advantages, the potential of laboratory tests in AI research is relatively unexplored and their application in healthcare AI has not been studied extensively [8]. Several factors contribute to this relative neglect. First, the sheer volume and complexity of laboratory data present significant challenges for data processing and model development. Second, models developed solely from laboratory data are perceived as lacking interpretability and clinical relevance, which may discourage researchers from focusing on this area. Finally, AI applications in healthcare have historically been focused on areas with more immediate visual or diagnostic impact, such as imaging [9]. Despite these challenges, the potential of laboratory tests to provide valuable insights into patient care is increasingly recognized, highlighting the need for more focused research in this area [10,11].

In this study, we aimed to develop an AI model for the early prediction of pneumonia mortality; early prediction is defined as the model's ability to predict outcomes based on the initial set of laboratory tests taken within the first 24 h of hospital admission. This approach is designed to assist clinicians in making timely and informed decisions regarding the urgency and nature of treatment required, potentially improving patient outcomes by facilitating early interventions.

2. Materials and methods

2.1. Patients and data collection

The study included patients aged 19 years and older who were diagnosed with pneumonia and required inpatient treatment at Ewha Womans University Mokdong Hospital, a tertiary care hospital, from September 2020 to March 2022. Pneumonia was diagnosed through confirmatory diagnostic tests. The following diagnostic tests were used.

- Chest radiography: pneumonia was identified by the presence of infiltrates or consolidations on chest x-rays. Radiographic findings were reviewed and confirmed by an experienced radiologist.
- Chest CT: if chest radiography was inconclusive, a chest CT scan was utilized for more detailed evaluation. The diagnostic criteria for CT included identifying lung glass opacities, consolidation, or other signs suggestive of pneumonia.
- Sputum culture: microbiologic confirmation was sought through sputum culture, particularly if bacterial pneumonia was suspected. Positive culture results indicating the presence of pathogenic bacteria were used to confirm the diagnosis. For patients with suspected COVID-19 pneumonia, polymerase chain reaction (PCR) testing for the SARS-CoV-2 virus was used to confirm the diagnosis.

Patients diagnosed with pneumonia but not requiring hospitalization were excluded based on exclusion criteria. We also excluded patients with coexisting conditions that could independently affect mortality risk, such as advanced cancer and tuberculosis. This retrospective data collection involved extracting relevant information from patient records, with a particular focus on data obtained on the day of hospitalization. The following data were collected. Clinical information, which includes basic demographic and clinical data such as sex, age, and COVID-19 PCR results, are fundamental to understanding the patient population and potential risk factors associated with pneumonia mortality. Vital signs are essential in assessing the overall condition of the patient and include parameters such as blood pressure and heart rate. Hematologic parameters include white blood cell count (WBC, × 10^3/μL), red blood cell count (× 10^6/μL), hemoglobin (g/dL), hematocrit (%), and related metrics. Hematologic parameters are crucial for indicating infection or inflammation, which are common in pneumonia, and can provide insights into its severity and type. Biochemical markers used encompass liver function tests, such as aspartate aminotransferase (AST, IU/L) and alanine aminotransferase (ALT, IU/L); renal function tests, such as blood urea nitrogen (BUN, mg/dL) and creatinine (mg/dL); and markers, such as lactate dehydrogenase (LDH, IU/L) and C-reactive protein (mg/dL). Biochemical markers help assess the systemic impact of pneumonia and identify complications including liver or kidney involvement. Respiratory and metabolic parameters include the arterial blood gas analysis (ABGA), which provides essential information about respiratory efficiency and gas exchange, a key concern in pneumonia. The data included a wide range of parameters, as summarized in Supplementary Table S1. While the retrospective nature of this study is advantageous for accessing a large amount of existing patient data, some challenges and limitations exist. One major challenge is the potential for incomplete or inconsistent record keeping, which can affect the completeness and accuracy of the collected data. In addition, historical data may not fully capture the dynamic nature of clinical practice and patient care, which changes over time. To mitigate these limitations, we applied strict criteria for data inclusion and conducted a thorough review to ensure the reliability and consistency of the data used in the model.

2.2. *Data preprocessing*

The 80,940 data instances used in this study consisted of 76 parameters. We acquired 64,388 laboratory results and clinical information samples and 16,552 (20.4 %) missing values were preprocessed as the median value. During the preprocessing step, the range of parameters was standardized and scaled using “scikit-learn,” which is a Python library.

2.3. *Feature extraction*

Various features were extracted using the collected data to improve the prediction performance of the model. Feature extraction provides new features (parameters) that can be used for AI training, thereby improving prediction performance. We obtained the following parameters through feature extraction: absolute monocyte count (AMC, $\times$ 10^3/μL), monocyte-to-lymphocyte ratio (monocyte/lymphocyte), AST-to-ALT ratio (AST/ALT), and pulse rate-to-respiratory rate ratio (PR/RR).

2.4. *Model selection and development*

We used several machine learning (ML) models—CatBoost, eXtreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), random forest (RF), support vector machine (SVM), and K-nearest neighbor (KNN)—with good classification performance. In addition, multilayer perceptron (MLP), a deep learning (DL) method, was used. All collected data, including those generated through feature extraction, were split into a training set (80 %) and test set (20 %). K-fold cross validation (n_split:5) was performed to avoid data loss during model training and to improve model prediction performance.

The performance of all the models was evaluated in terms of the area under the receiver operating characteristic curve (AUROC), accuracy, precision, recall, and F1-score. In developing the prediction models, AUROC and F1-score were optimized for each individual model. AUROC was optimized by tuning the hyperparameters, considering the validation loss during the model training process. In the case of MLP, AUROC was optimized by stacking two hidden layers and using a dropout technique. In addition, the F1-score was optimized on the data used to develop our model because of the data imbalance between the survival and nonsurvival groups. The F1-score is the harmonic average of precision and recall, and it is an important predictive performance evaluation index, as important as accuracy and AUROC in evaluating model performance. In this study, F1-score was optimized through a cut-off adjustment for each AUROC-optimized model.

2.5. *Utility and calibration*

We used the expected calibration error (ECE) and standardized net benefit (SNB) in model development. The ECE score for calibration evaluation is the expected value of the difference between confidence and actual accuracy [12]. A higher ECE score indicates a larger difference between the output reliability (pseudo-probability) and the actual model accuracy of the predictions. SNB is a state-of-the-art utility metric used to evaluate the performance of a decision model or process in specific situations [13].

2.6. *Development of ensemble model*

We attempted to improve performance by developing an ensemble model based on a combination of the developed models. The AUROC-optimized ensemble model was developed using three models (XGBoost, CatBoost, and LGBM) that exhibited a high AUROC and a weighted soft-voting technique using the probability value. To develop the F1-score-optimized ensemble model, we used three models (XGBoost, CatBoost, and RF) with a high F1-score, and a hard-voting technique was applied.

2.7. *SHapley Additive exPlanations (SHAP) method*

The SHapley Additive exPlanations (SHAP) method was used to analyze laboratory results to evaluate the impact of feature parameters on pneumonia mortality. The SHAP method is a novel technique that estimates the impact of each feature using a probabilistic game rule [14,15]. The SHAP method for MLP decomposes the output prediction of a neural network for a specific input by backpropagating all features to extract the contributions of all neurons [16]. Therefore, we obtained the feature impact using the SHAP method for the ML models and MLP.

2.8. Feature importance

We extracted the features of the ML models with excellent performance based on their importance using the "Feature importance" technique, which allocates a score to the input parameters (features) based on the importance of predicting a target variable (mortality).

3. Results

This study was conducted between September 2020 and March 2022; it involved 1065 patients diagnosed with pneumonia and categorized into survival (877 patients) and nonsurvival (188 patients) groups.

3.1. Performances of ML and DL models by AUROC optimization

The AUROC of the developed XGBoost model was the best (0.8989), followed by LGBM (0.8968), CatBoost (0.8962), RF (0.8586), and DL (0.8498). In terms of accuracy, XGBoost, CatBoost, LGBM, and RF were the best (0.88), followed by SVM (0.86), DL (0.86), and RF (0.85). For the F1-score, XGBoost (0.76) was the best, followed by CatBoost, LGBM, RF (0.74), and DL (0.73). The overall performance of XGBoost was the best (Table 1 and Fig. 1).

XGBoost achieved the best ECE score (0.026), followed by MLP (0.027), LGBM (0.033), CatBoost (0.035), RF (0.037), SVM (0.038), and ensemble model (0.040). KNN (0.798) was the best in terms of SNB, followed by CatBoost (0.781), LGBM (0.781), RF (0.781), XGBoost (0.776), and ensemble models (0.776).

3.2. Improved performance of ML and DL models by F1 score optimization

Cut-off adjustments were performed for each model to optimize the F1-score. Consequently, the AUROC and accuracy were the same, and as the precision and recall changed, the F1-score increased. The F1-score increased from 0.76 to 0.80 in XGBoost, 0.74 to 0.78 in LGBM, 0.74 to 0.80 in CatBoost, 0.72 to 0.75 in SVM, 0.74 to 0.80 in RF, and 0.63 to 0.70 in KNN. The F1-score of MLP also slightly increased from 0.74 to 0.75 (Table 2 and Fig. 1).

3.3. Performance of ensemble model

To improve AUROC, we developed an ensemble model using three models (XGBoost, CatBoost, and LGBM) with excellent AUROC. We used a weighted soft-voting technique that differentially assigns weights to the results of the three models, resulting in an AUROC of 0.9006, which was the highest among the models we developed (Table 1 and Fig. 1). In addition, to improve the F1-score, the three models (XGBoost, CatBoost, and RF) with high F1-score were used. A hard-voting method based on the voting results of the three models of the output class was used. An ensemble model was developed using the voting results of two or more of the above three models. The F1-score was 0.81, and the accuracy was 0.90, which was the best among the models we developed (Table 2 and Fig. 1).

3.4. Feature impact by SHAP for ML and DL models

Fig. 2 shows the top 20 feature impacts for the prediction performance of each model using the SHAP method. Systolic blood pressure (SBP) had the most impact on the performance of XGBoost, followed by serum glucose level, age, AST/ALT, and body temperature. For CatBoost, SBP was the highest, followed by serum glucose level, AST/ALT ratio, mean arterial blood pressure (MBP), and age. For LGBM, SBP was the highest, followed by serum glucose level, AST/ALT ratio, total protein, and age. AST/ALT and AMC,

Table 1
Performance of developed models by AUROC optimization.

Model	AUROC	Accuracy	Precision	Recall	F1-score	ECE	SNB
XGBoost	0.8989	0.88	0.84	0.72	0.76	0.026	0.776
SVM	0.8224	0.86	0.79	0.69	0.72	0.038	0.764
CatBoost	0.8962	0.88	0.84	0.70	0.74	0.035	0.781
LGBM	0.8968	0.88	0.84	0.70	0.74	0.033	0.781
KNN	0.7732	0.85	0.87	0.60	0.63	0.086	0.798
RF	0.8586	0.88	0.84	0.70	0.74	0.037	0.781
MLP	0.8498	0.86	0.78	0.70	0.73	0.027	0.759
Ensemble	0.9006[a]	0.90	0.84	0.79	0.81	0.040	0.776

Abbreviations: AUROC, Area under the receiver operating characteristic curve; ECE, Expected calibration error; SNB, standardized net benefit; XGBoost, eXtreme gradient boosting, SVM: Support vector machine, LGBM: Light gradient boosting machine, KNN: K-nearest neighbor, RF: Random forest, MLP: Multilayer perceptron.

[a] Improved AUROC by blending of XGBoost, CatBoost, and LGBM.

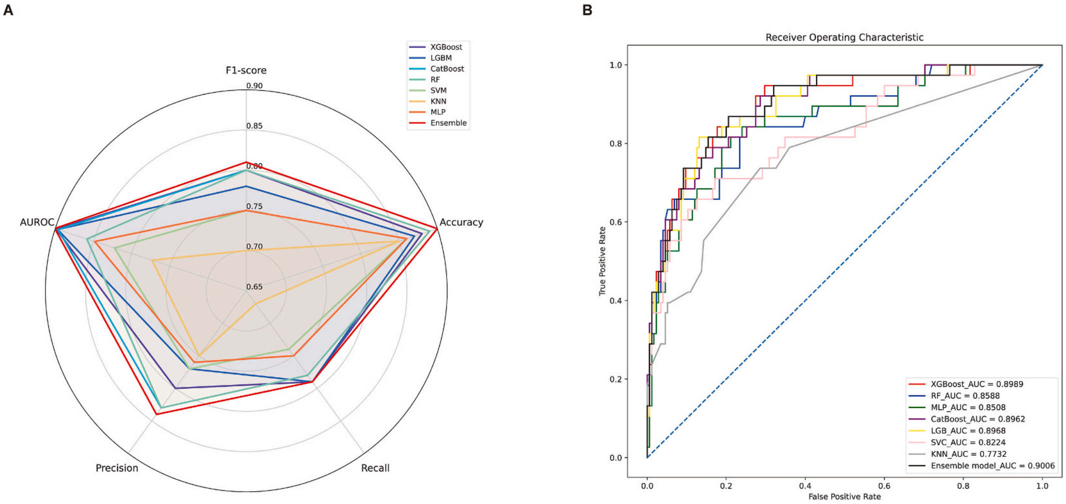

Fig. 1. Performance of each model. (A) Area under the receiver operating characteristic curve (AUROC) (B) Rador plot visualization for each model.

Table 2
Performance of developed models by F1-score optimization.

Model	AUROC	Accuracy	Precision	Recall	F1-score
XGBoost	0.8989	0.88	0.80	0.79	0.80
SVM	0.8224	0.86	0.77	0.74	0.75
Catboost	0.8962	0.88	0.83	0.78	0.80
LGBM	0.8968	0.88	0.77	0.79	0.78
KNN	0.7732	0.85	0.75	0.67	0.70
RF	0.8586	0.88	0.83	0.78	0.80
MLP	0.8498	0.86	0.76	0.75	0.75
Ensemble	0.9006	0.90	0.84	0.79	0.81[a]

Abbreviations: AUROC, area under the receiver operating characteristic curve; XGBoost, eXtreme gradient boosting, SVM: support vector machine, LGBM: Light gradient boosting machine, KNN: K-nearest neighbor, RF: Random forest, MLP: Multilayer perceptron.

[a] Improved F1-score by blending of XGBoost, CatBoost, and RF.

which were generated through feature extraction in the three ML models, were included in the top 10. The feature impact of the MLP model using the SHAP method was highest in COVID-19 positive finding (positive correlation), followed by age, SBP, MBP, diastolic blood pressure (DBP), and total protein. For the MLP model, AMC and AST/ALT were in the top 20. For the four models using the SHAP method, O_2 saturation, serum glucose level, total protein, BUN, LDH, AMC, and AST/ALT were included in the top 20.

3.5. *Feature importance for ML models*

The “Feature importance” technique was performed for four ML models with excellent performance among the developed models (top 20, Fig. 3). XGBoost exhibited feature importance in the order of SBP, MBP, PT (INR), DBP, O_2 saturation, and COVID-19 PCR results. AMC and AST/ALT, which were generated through feature extraction, were also included in the top 20. LGBM exhibited feature importance in the order of SBP, monocyte (%), AMC, body temperature, AST, age, and serum glucose level. RF exhibited feature importance in the order of SBP, MBP, DBP, pulse pressure, AMC, base excess, and monocyte (%); AST/ALT was also included in the top 20. Among the laboratory results, O_2 saturation, AST/ALT, AMC, monocyte (%), and BUN were commonly included in the top 20 among the four ML models.

4. Discussion

We developed a high-performance pneumonia mortality prediction model using only the most easily accessible laboratory results and basic clinical information (sex, age, vital signs, etc.) in the clinical field. In addition, we confirmed that the performance was improved by implementing an ensemble model using the ML and MLP models. It is important to acknowledge that our model was specifically developed and validated on a patient cohort primarily composed of pneumonia cases, including COVID-19 patients, within a specific time frame and healthcare setting. This specialization may contribute to the model’s high performance in terms of the metrics and also suggests that the model may be finely tuned to the characteristics of this particular patient population. By contrast, the Acute Physiology and Chronic Health Evaluation (APACHE) II, Simplified Acute Physiology Score (SAPS) 3, and Sequential Organ Failure

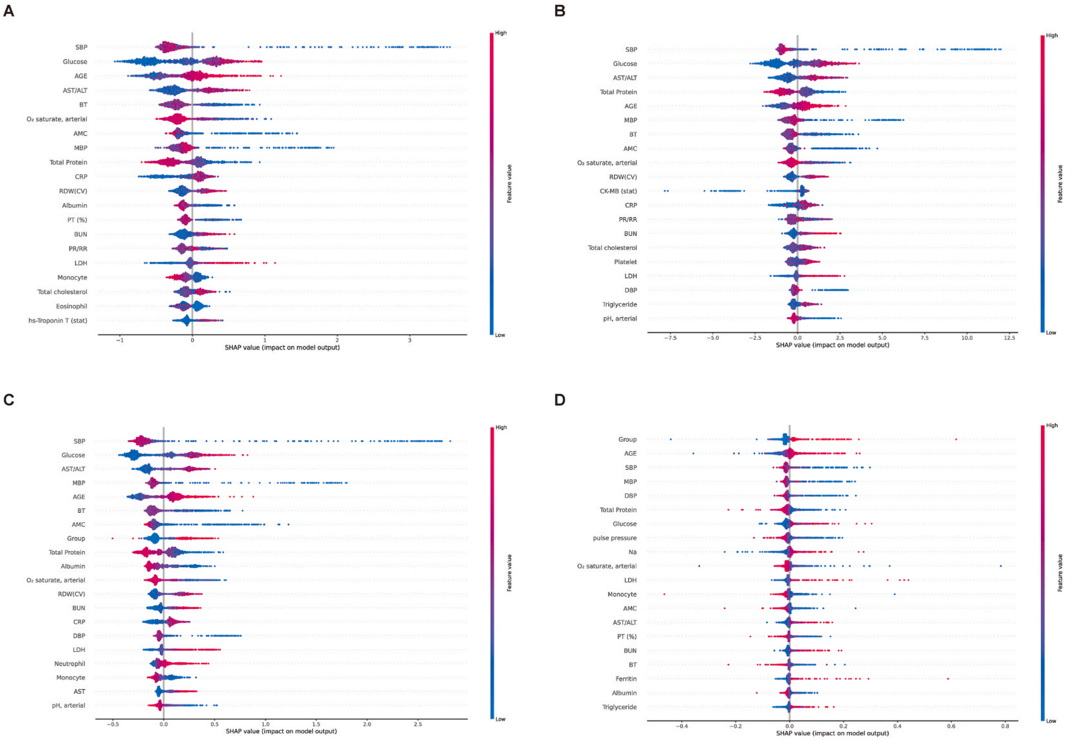

Fig. 2. Analysis of features contributing to pneumonia mortality by Shapley Additive exPlanations (SHAP) method. (A) XGBoost (B) LGBM (C) CatBoost (D) MLP.

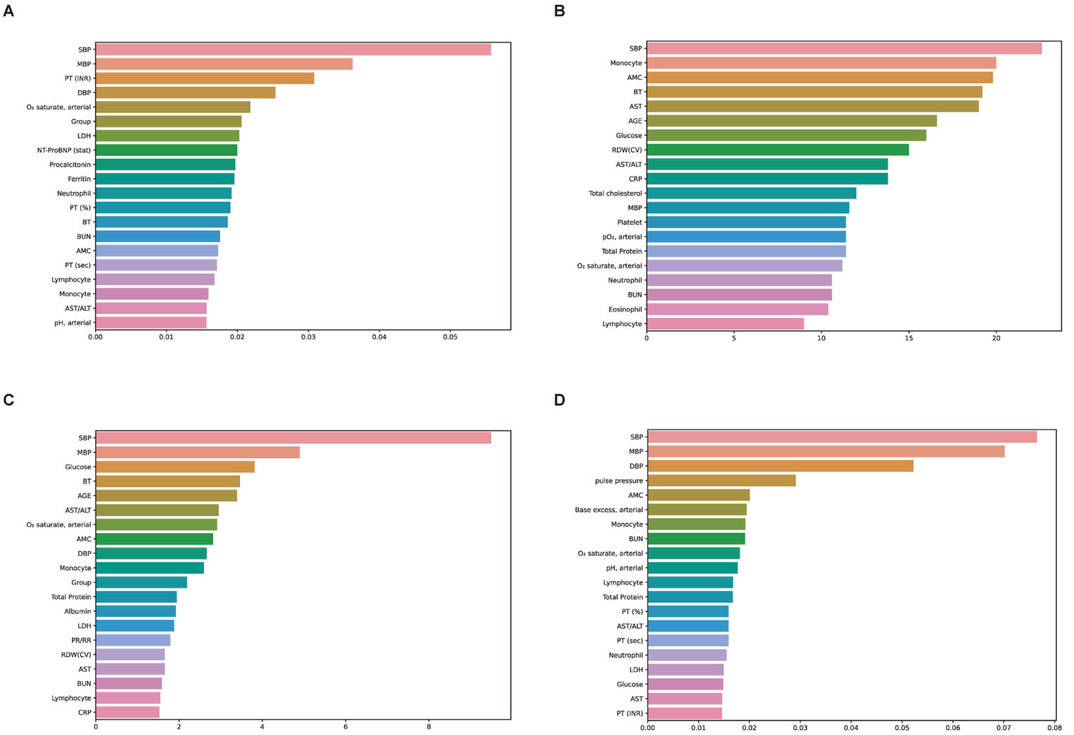

Fig. 3. Analysis of features contributing to pneumonia mortality by “Feature importance technique.” (A) XGBoost (B) LGBM (C) CatBoost (D) RF.

Assessment (SOFA) scores have been validated across a broader range of patient populations and clinical settings [17]. This extensive validation imparts a level of generality and robustness to these models, which may not yet be fully established for our AI model. Additionally, differences in data collection protocols, such as the time frame of data collection and the specific parameters measured, can also influence the performance metrics. Another consideration is the evolution of medical practice and patient demographics over time, which may affect the comparability of these models. For instance, the patient population and clinical practices during the COVID-19 pandemic may differ significantly from those in the periods when APACHE II, SAPS 3, and SOFA were developed and validated. Therefore, while our AI model shows promising results, we must be cautious in directly comparing its performance with those of these well-established models. Further validation studies, particularly in diverse clinical settings and with varied patient populations, are necessary to fully ascertain the model's applicability and effectiveness in comparison with existing standards.

The MLP used in this study consisted of an input layer, hidden layer, and output layer; the best results were obtained when two hidden layers were stacked rather than one, and overfitting with more than three or four hidden layers worsened the performance. In addition, when 1024 nodes of the first hidden layer and 128 nodes of the second hidden layer were used to set the number of nodes in the layer, more diverse features could be extracted compared with the 512 nodes of the first hidden layer and 64 nodes of the second hidden layer, thereby improving performance. In addition, we used the dropout technique when stacking the hidden layers. The dropout method can be used to avoid overfitting and secure diversity in DL, such as MLPs using neural networks [18–20]. In general, among the tree-based models, the boosting-type ML model is known to exhibit good classification [21–23]. In this study, the latest boosting models, XGBoost, CatBoost, and LGBM, exhibited better AUROC than RF, which is a classic tree model bagging technique. Tree-based models solve this problem by creating trees through various questions; however, XGBoost, CatBoost, and LGBM are somewhat different in terms of tree depth. While XGBoost and CatBoost build a tree with all the data and deepen the tree branch by increasing the number of questions, LGBM tends to build a deeper tree with only the data corresponding to the question that comes out in favor of asking the question. This LGBM learning method is known as "Greedy algorithm." In this study, the AUROCs of XGBoost, CatBoost, and LGBM did not show significant differences (Tables 1 and 2). However, because of optimizing the F1-score through cut-off adjustment, LGBM's F1-score was slightly lower than those of the other two models, which is thought to be due to the above-mentioned "Greedy algorithm" of LGBM (Table 2). In this study, which mainly used numerical data, including laboratory tests, the performance of MLP was not as good as that of the latest boosting models; however, it exhibited better results than KNN and SVM. Therefore, the use of an MLP in the development of an AI model using formal hospital data composed mainly of numerical data is also worthy of consideration.

An ensemble model refers to the blending of various AI models to improve performance [24,25]. In this study, an ensemble model was developed using two methods. First, a weighted soft-voting technique was used for AUROC improvement; it assigned weight to each probability value using XGBoost, CatBoost, and LGBM, which exhibited good AUROC. As a result, we obtained an AUROC of 0.9006, which was the best among our models. Second, the class result values of XGBoost (0.80), CatBoost (0.80), and RF (0.80), which exhibited excellent F1-score by adjusting the cut-off, were used, and a hard-voting method that selects the results voted by two or more out of the three models was used to obtain the optimal F1-score (0.81) (Table 2). The choice of MLPs along with ML models such as XGBoost, CatBoost, and LGBM for the ensemble AI model in this study was to leverage the strengths of each model. MLPs are good at identifying complex patterns in data, which is essential in a healthcare application like ours. MLPs excel at recognizing complex relationships that are not immediately apparent, making them well-suited to the subtle task of predicting pneumonia mortality. For the ML component, we chose models like XGBoost, CatBoost, and LGBM because of their ability to handle diverse datasets robustly and effectively, and each has a proven track record in classification tasks, particularly in healthcare prediction. These models bring different advantages: XGBoost for speed and performance on imbalanced data, CatBoost for excellent performance on categorical data, and LGBM for efficiency on large datasets. By combining these models in an ensemble approach, we were able to build a system that not only provides accurate predictions but is also reliable and versatile in a variety of clinical scenarios. We believe that this combination can complement the strengths of each model to improve the overall predictive ability of the AI system for pneumonia mortality. This combination of DL and ML models was driven by the goal of developing a comprehensive and effective tool that can accurately predict pneumonia mortality and meet the complexity of medical data analysis.

Although the ranking of "Feature importance" in the laboratory results for each model was different, the types of parameters included were similar (Fig. 3). For four models (XGBoost, LGBM, CatBoost, and RF), O_2 saturation, AST/ALT, AMC, monocyte (%), and BUN were included in the top 20. In general, ABGA results are important for respiratory diseases such as pneumonia [26,27]. Information such as pH, $PaCO_2$, PaO_2, and O_2 saturation can be obtained using ABGA. In general, PaO_2 is considered more important than O_2 saturation in respiratory diseases such as pneumonia because PaO_2 is more sensitive to hypoxemia than O_2 saturation, as shown by the oxyhemoglobin dissociation curve [28]. Nevertheless, it is very characteristic that the most important feature affecting the performance of our prediction model was O_2 saturation and not PaO_2. The importance of BUN has also been confirmed, and BUN levels are closely related to kidney function. Pneumonia is a major cause of sepsis and may require adequate fluids [29]. The fact that early BUN levels are an important feature of mortality from respiratory diseases such as pneumonia suggests that clinicians should consider BUN levels when initially evaluating patients with pneumonia. Feature extraction AMC and AST/ALT, which were performed to improve the predictive performance of the model, were also included in the top 20 ML models. Therefore, AMC and AST/ALT may be considered when assessing severity, such as pneumonia mortality, and related studies can be conducted.

Although laboratory results of risk factors affecting pneumonia mortality can be confirmed through "Feature importance," positive or negative correlation could not be confirmed. Therefore, we obtained a more detailed feature impact by using the SHAP method (Fig. 2). In addition, "Feature importance" could not be applied to MLP; however, the feature impact was obtained using the SHAP method. Consequently, we obtained the feature impact using the SHAP method by targeting XGBoost, LGBM, CatBoost, and MLP. Among the laboratory results of the four models using the SHAP method, O_2 saturation, serum glucose level, total protein, BUN, LDH, AMC, and AST/ALT were commonly included in the top 20. Except for O_2 saturation, laboratory tests for serum glucose, total protein, BUN, LDH, AMC, and AST/ALT levels were not specific for respiratory diseases, such as pneumonia. High serum glucose (positive correlation), low total protein (negative correlation), and high BUN levels (positive correlation) are commonly observed in stressful situations, such as infection and sepsis [29–31]. LDH is a laboratory test that is not specific to a particular disease because it can be elevated in a wide variety of diseases, such as sepsis and cancer [32,33]; however, LDH showed an important feature impact in the pneumonia mortality prediction model. Furthermore, the discovery of AMC and AST/ALT generated through feature extraction, which contributed to the performance improvement of the predictive model, was a characteristic result of this study. In general, because monocytes (%) are checked in the clinical field, the actual AMC may vary depending on the WBC count, even if the % value is the same. In clinical practice, monocyte values tend to be overlooked when evaluating pneumonia severity. However, since the declaration of the COVID-19 pandemic, several studies have reported that monocytes are related to COVID-19 pneumonia [34,35]. Therefore, AMC may be related to the initial evaluation of the severity of other bacterial or viral pneumonia, as well as COVID-19 pneumonia, and hence a follow-up study on this should be considered. It is also a feature of our study that the feature extracted AST/ALT used to improve the prediction performance of the model showed high feature impact in both "Feature importance" and the SHAP method. In general, liver enzymes such as AST and ALT are easily overlooked during the initial evaluation of respiratory disease. The APACHE II, SAPS 3, and SOFA scores measure total bilirubin levels instead of AST and ALT levels for liver function assessment [36–38]. Although there are few studies on the correlation between AST and ALT and pneumonia mortality, some studies on the correlation between COVID-19 and AST and ALT levels have been reported [39]. In our study, the AST/ALT levels were positively correlated with mortality. This result indicates that the elevation of AST was more pronounced than that of ALT. Some studies have reported that there is a significant increase in AST levels in patients with severe COVID-19; however, these studies did not target all types of pneumonia cases [40–43]. Therefore, studies on liver enzymes, such as AST and ALT, should be considered when evaluating pneumonia severity.

In our study, the identification of features such as O_2 saturation, AST/ALT ratio, and AMC as significant predictors of mortality in pneumonia patients is both novel and clinically insightful. Discussing the implications of these findings enhances the interpretability and applicability of our AI model in the context of pneumonia severity and patient management. O_2 saturation, a critical parameter for respiratory function, is essential in assessing the severity of pneumonia. Low O_2 saturation levels can indicate severe respiratory compromise, which is common in advanced pneumonia stages. Clinically, this parameter's prominence in our model suggests the need for vigilant monitoring of respiratory status in pneumonia patients, guiding timely interventions to prevent severe outcomes. The AST/ALT ratio's relevance highlights the impact of pneumonia on systemic inflammation and liver function. This finding may prompt clinicians to consider broader systemic involvement in pneumonia, especially in cases of severe infection or complications such as COVID-19 pneumonia. Regular monitoring of liver function tests, including AST/ALT ratios, could become an important aspect of comprehensive patient assessment, aiding in the early detection of systemic complications. Moreover, the significance of AMC in our model sheds light on the role of the immune system in pneumonia progression. Elevated AMC levels could signify an intensified immune response, often seen in severe infections. This insight can be instrumental in identifying high-risk patients early on, enabling clinicians to tailor treatment strategies more effectively. These novel findings, particularly concerning AMC and AST/ALT ratios, enhance our understanding of pneumonia severity and have practical implications for clinical practice. Incorporating these parameters into routine patient evaluations could lead to more personalized and effective treatment approaches, aligning with the goals of precision medicine. Such integration of AI model predictions with clinical insights holds the promise for improving patient outcomes in pneumonia care, offering a more nuanced approach to disease management and treatment decision-making.

To ensure the practical application of our AI model in clinical settings, it is essential to outline how it can be integrated into the existing healthcare infrastructure to assist health professionals. Our AI model leverages routinely collected laboratory data and patient information to predict mortality risk of patients with pneumonia upon their initial presentation. This early prediction capability allows healthcare providers to identify high-risk patients early, optimizing care prioritization and potentially improving survival rates. Additionally, the model aids in optimizing resource allocation in resource-limited settings by identifying high-risk patients, which helps in the management of intensive care resources and personnel. Furthermore, the model's output can inform tailored treatment strategies based on predicted risk levels, enhancing personalized care practices. The model can also be part of a dynamic system that reassesses risk as new laboratory results become available during a patient's hospital stay, allowing for ongoing optimization of treatment plans.

Although our study presents significant findings in the development of an AI model for pneumonia mortality prediction, it is important to acknowledge its limitations and suggest directions for future research. One of the primary limitations is the bias potential inherent in the dataset. Our study relies on retrospective data from a single tertiary care hospital, which may not fully represent the broader population of patients with pneumonia. This limitation could impact the generalizability of our model to different healthcare settings or patient demographics. Future studies should aim to validate and refine the model using multicenter data, encompassing a more diverse patient population. Another concern is the quality and completeness of the data used. As with any retrospective study, there is the risk of missing or inaccurately recorded data, which can affect the model's accuracy and reliability. Efforts to minimize data quality issues through rigorous data cleaning and validation processes were undertaken; however, these concerns cannot be entirely eliminated in retrospective analyses. Furthermore, although our model showed promising results in predicting pneumonia mortality, its real-world applicability requires further exploration. This includes addressing practical challenges in integrating the model into clinical workflows and ensuring its adaptability to different electronic health record systems. Future research should also focus on expanding the model's capabilities, such as incorporating additional relevant variables, including patient comorbidities, treatment regimens, and longer-term outcomes. This expansion would enhance the model's comprehensiveness and applicability in clinical practice. In addition, although this study provides important insights into the predictive value of laboratory tests in determining pneumonia mortality, we acknowledge the potential impact of post-COVID-19 status as a limitation. The inclusion of COVID-19 infection status as an input variable was an important decision to accurately reflect the impact of the pandemic; however, potential bias may arise given the unique pathophysiologic mechanisms and clinical presentation associated with COVID-19 pneumonia. This factor may affect the generalizability of our findings to non-COVID-19 pneumonia cases. We recommend that future studies further explore the distinctive features of COVID-19 pneumonia and their impact on AI-based prediction models to ensure a comprehensive understanding of risk factors for pneumonia mortality post-pandemic. Finally, as a minor limitation, the exclusion of individuals under 19 years of age limits the applicability of our findings to the pediatric population. Future studies could aim to develop and validate similar predictive models specific to pediatric pneumonia, which would be extremely useful for improving care in all age groups.

5. Conclusion

The development of an AI model that primarily utilizes laboratory test results to predict pneumonia mortality represents an important step forward in the application of AI in clinical settings. Our study showed that an ensemble model incorporating techniques such as XGBoost, CatBoost, and LGBM significantly outperformed other tested AI models in predicting pneumonia mortality. This model not only provided the highest accuracy but also showed robustness in handling a wide variety of clinical data, making it a particularly useful tool for clinicians seeking to improve treatment outcomes. Our findings support the adoption of this ensemble approach as a reliable predictor of mortality in patients with pneumonia and highlight the potential of advanced ML techniques in improving patient care.

The immediate implication for clinicians is the potential of AI-based tools to improve decision-making in the treatment of pneumonia. This model demonstrates how the risk of death can be predicted with high accuracy by utilizing routine laboratory data that are readily available in clinical settings. This can help clinicians stratify risk early on, enabling more personalized and timely interventions, especially for high-risk patients. Integrating these AI models into clinical workflows can augment traditional diagnostic and prognostic methods to improve patient outcomes.

Researchers in the field of healthcare AI can leverage our approach to model development, including using ensemble techniques that combine multiple AI methodologies to improve predictive accuracy. Emphasizing explainability and clinical relevance in AI model design is another key aspect that can guide future research efforts. Moreover, this study highlights the importance of focusing on lesser-known data types, such as laboratory results, in healthcare AI research to open new avenues for innovation.

Ultimately, the study suggests a shift to a more data-driven, precise, and patient-centered approach when managing pneumonia. The insights gained here have the potential to influence current clinical practice by introducing AI as a complementary tool in patient care. However, it is important to note that AI should augment clinical judgment not replace it. Future research should focus on further validating these models in different clinical settings and exploring integration with existing healthcare systems.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics statement

This study was approved by the Institutional Review Board of the Ewha Womans University Mokdong Hospital (approval number: EUMC 2022-01-031-001). The patient records were reviewed and published in accordance with the Declaration of Helsinki. The requirement for informed consent was waived owing to the retrospective nature of the study.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Seung Min Baik: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Kyung Sook Hong:** Writing – review & editing, Validation. **Jae-Myeong Lee:** Writing – review & editing, Validation. **Dong Jin Park:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e34525.

References

[1] OECD, E. Union, Health at a Glance: Europe 2018, 2018.
[2] Centers for Disease Control and Prevention, Age-adjusted death rates for the 10 leading causes of death in 2020: United States, 2019 and 2020, June, 2022, https://www.cdc.gov/nchs/nvss/deaths.htm?CDC_AA_refVal=, 2020.
[3] R. Li, C. Rivers, Q. Tan, M.B. Murray, E. Toner, M. Lipsitch, The demand for inpatient and ICU beds for COVID-19 in the US: lessons from Chinese cities, medRxiv (2020), https://doi.org/10.1101/2020.03.09.20033241.
[4] G. Grasselli, A. Pesenti, M. Cecconi, Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response, JAMA 323 (2020) 1545–1546, https://doi.org/10.1001/jama.2020.4031.
[5] R. Fusco, R. Grassi, V. Granata, et al., Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: machine learning and deep learning approaches for diagnosis and treatment, J Pers Med 11 (2021), https://doi.org/10.3390/jpm11100993.
[6] W. Zouch, D. Sagga, A. Echtioui, et al., Detection of COVID-19 from CT and chest X-ray images using deep learning models, Ann. Biomed. Eng. 50 (2022) 825–835, https://doi.org/10.1007/s10439-022-02958-5.
[7] C. Lyphout, J. Bergs, W. Stockman, et al., Patient safety incidents during interhospital transport of patients: a prospective analysis, Int Emerg Nurs 36 (2018) 22–26, https://doi.org/10.1016/j.ienj.2017.07.008.
[8] T.U. Blatter, H. Witte, C.T. Nakas, A.B. Leichtle, Big data in laboratory medicine-FAIR quality for AI? Diagnostics 12 (2022) 1923, https://doi.org/10.3390/diagnostics12081923.
[9] L.R. Baltazar, M.G. Manzanillo, J. Gaudillo, et al., Artificial intelligence on COVID-19 pneumonia detection using chest xray images, PLoS One 16 (2021) e0257884, https://doi.org/10.1371/journal.pone.0257884.
[10] M.M. Islam, H.C. Yang, T.N. Poly, Y.J. Li, Development of an artificial intelligence-based automated recommendation system for clinical laboratory tests: retrospective analysis of the National Health Insurance database, JMIR Med Inform 8 (2020) e24163, https://doi.org/10.2196/24163.
[11] G. Cardozo, S.F. Tirloni, A.R. Pereira Moro, J.L.B. Marques, Use of artificial intelligence in the search for new information through routine laboratory tests: systematic review, JMIR Bioinform Biotech 3 (2022) e40473, https://doi.org/10.2196/40473.
[12] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.
[13] A. Campagner, F. Sternini, F. Cabitza, Decisions are not all equal. Introducing a utility metric based on case-wise raters' perceptions, Comput. Methods Progr. Biomed. (2022) 106930.
[14] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888 (2018).
[15] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[16] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences. arXiv Preprint arXiv:1704.02685, 2017.
[17] A.L.E. Falcao, A.G.A. Barros, A.A.M. Bezerra, et al., The prognostic accuracy evaluation of SAPS 3, SOFA and Apache II scores for mortality prediction in the surgical ICU: an external validation study and decision-making analysis, Ann. Intensive Care 9 (2019) 18, https://doi.org/10.1186/s13613-019-0488-9.
[18] H. Li, J. Weng, Y. Mao, et al., Adaptive dropout method based on biological principles, IEEE Trans Neural Netw Learn Syst 32 (2021) 4267–4276, https://doi.org/10.1109/TNNLS.2021.3070895.
[19] L. Li, C. Zhang, S. Liu, H. Guan, Y. Zhang, Age prediction by DNA methylation in neural networks, IEEE/ACM Trans Comput Biol Bioinform 19 (2022) 1393–1402, https://doi.org/10.1109/TCBB.2021.3084596.
[20] J. Xie, Z. Ma, J. Lei, et al., Advanced dropout: a model-free methodology for Bayesian dropout optimization, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 4605–4625, https://doi.org/10.1109/TPAMI.2021.3083089.
[21] C. Freeman, D. Kuli, O. Basir, Feature-selected tree-based classification, IEEE Trans. Cybern. 43 (2013) 1990–2004, https://doi.org/10.1109/TSMCB.2012.2237394.
[22] S.A. Parikh, R. Gomez, M. Thirugnanasambandam, et al., Decision tree based classification of abdominal aortic aneurysms using geometry quantification measures, Ann. Biomed. Eng. 46 (2018) 2135–2147, https://doi.org/10.1007/s10439-018-02116-w.
[23] N.J. Rhodes, J.N. O'Donnell, B.D. Lizza, M.M. McLaughlin, J.S. Esterly, M.H. Scheetz, Tree-based models for predicting mortality in gram-negative bacteremia: avoid putting the CART before the horse, Antimicrob. Agents Chemother. 60 (2016) 838–844, https://doi.org/10.1128/AAC.01564-15.
[24] M. Cannas, B. Arpino, A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting, Biom. J. 61 (2019) 1049–1072, https://doi.org/10.1002/bimj.201800132.
[25] E. Yaman, A. Subasi, Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification, BioMed Res. Int. 2019 (2019) 9152506, https://doi.org/10.1155/2019/9152506.
[26] V.M. Manuilov, A.V. Suvorov, S.V. Kurkin, et al., Evaluation of the efficiency of oxygen–helium therapy for patients with Covid-19-associated pneumonia, Hum. Physiol. 48 (2022) 863–870, https://doi.org/10.1134/s0362119722070143.
[27] K.P. Levin, B.H. Hanusa, A. Rotondi, et al., Arterial blood gas and pulse oximetry in initial management of patients with community-acquired pneumonia, J. Gen. Intern. Med. 16 (2001) 590–598, https://doi.org/10.1046/j.1525-1497.2001.016009590.x.
[28] A. Madan, Correlation between the levels of $SpO_2$and PaO_2, Lung India 34 (2017) 307–308, https://doi.org/10.4103/lungindia.lungindia_106_17.
[29] L. Evans, A. Rhodes, W. Alhazzani, et al., Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021, Crit. Care Med. 49 (2021) e1063–e1143, https://doi.org/10.1097/CCM.0000000000005337.
[30] W.K. Yu, W.Q. Li, N. Li, J.S. Li, Influence of acute hyperglycemia in human sepsis on inflammatory cytokine and counterregulatory hormone concentrations, World J. Gastroenterol. 9 (2003) 1824–1827, https://doi.org/10.3748/wjg.v9.i8.1824.
[31] G.S. Martin, M. Moss, A.P. Wheeler, M. Mealer, J.A. Morris, G.R. Bernard, A randomized, controlled trial of furosemide with or without albumin in hypoproteinemic patients with acute lung injury, Crit. Care Med. 33 (2005) 1681–1687, https://doi.org/10.1097/01.ccm.0000171539.47006.02.
[32] M.S. Rahi, V. Jindal, S.P. Reyes, K. Gunasekaran, R. Gupta, I. Jaiyesimi, Hematologic disorders associated with COVID-19: a review, Ann. Hematol. 100 (2021) 309–320, https://doi.org/10.1007/s00277-020-04366-y.
[33] M. Maekawa, M. Inomata, M.S. Sasaki, et al., Electrophoretic variant of a lactate dehydrogenase isoenzyme and selective promoter methylation of the LDHA gene in a human retinoblastoma cell line, Clin Chem. 48 (2002) 1938–1945.
[34] M. Merad, J.C. Martin, Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages, Nat. Rev. Immunol. 20 (2020) 355–362, https://doi.org/10.1038/s41577-020-0331-4.
[35] A. Jafarzadeh, P. Chauhan, B. Saha, S. Jafarzadeh, M. Nemati, Contribution of monocytes and macrophages to the local tissue inflammation and cytokine storm in COVID-19: lessons from SARS and MERS, and potential therapeutic interventions, Life Sci. 257 (2020) 118102, https://doi.org/10.1016/j.lfs.2020.118102.
[36] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, Apache II: a severity of disease classification system, Crit. Care Med. 13 (1985) 818–829.
[37] R.P. Moreno, P.G. Metnitz, E. Almeida, et al., SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission, Intensive Care Med. 31 (2005) 1345–1355, https://doi.org/10.1007/s00134-005-2763-5.

[38] M. Cardenas-Turanzas, J. Ensor, C. Wakefield, et al., Cross-validation of a sequential organ failure assessment score-based model to predict mortality in patients with cancer admitted to the intensive care unit, J. Crit. Care 27 (2012) 673–680, https://doi.org/10.1016/j.jcrc.2012.04.018.

[39] N. Ali, K. Hossain, Liver injury in severe COVID-19 infection: current insights and challenges, Expert Rev Gastroenterol Hepatol 14 (2020) 879–884, https://doi.org/10.1080/17474124.2020.1794812.

[40] P.P. Bloom, E.A. Meyerowitz, Z. Reinus, et al., Liver biochemistries in hospitalized patients with COVID-19, Hepatology 73 (2021) 890–900, https://doi.org/10.1002/hep.31326.

[41] W.J. Guan, Z.Y. Ni, Y. Hu, et al., Clinical characteristics of coronavirus disease 2019 in China, N. Engl. J. Med. 382 (2020) 1708–1720, https://doi.org/10.1056/NEJMoa2002032.

[42] S. Richardson, J.S. Hirsch, M. Narasimhan, et al., Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area, JAMA 323 (2020) 2052–2059, https://doi.org/10.1001/jama.2020.6775.

[43] F. Lei, Y.M. Liu, F. Zhou, et al., Longitudinal association between markers of liver injury and mortality in COVID-19 in China, Hepatology 72 (2020) 389–398, https://doi.org/10.1002/hep.31301.