Preventive Medicine Reports 45 (2024) 102841

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context

Luyao Zhou [a], Min Shao [b], Cui Wang [b], Yu Wang [a,*]

[a] School of Biomedical Engineering, Anhui Medical University, Hefei, China
[b] Department of Critical Care Medicine, First Affiliated Hospital of Anhui Medical University, Hefei, China

ARTICLE INFO

Keywords:
Machine learning
Prediction model
Sepsis
Data imbalance
Shapley additive explanation
Clinical decision

ABSTRACT

Background: Early and accurate diagnoses of sepsis patients are essential to reduce the mortality. However, the sepsis is still diagnosed in a traditional way in China despite the increasing number of related studies, which may to some extent lead to delays in the treatment.
Methods: The study included 2,385 patients, including 364 with sepsis, collected from the First Affiliated Hospital of Anhui Medical University and partner hospitals from April to July 2022. External validation was conducted using the MIMIC-III database (over 60,000 patients from 2001 to 2012) and the eICU Collaborative Research Database (139,000 patients from 2014 to 2015). Multiple algorithm models, along with the SHapley Additive exPlanations (SHAP) analysis, are applied to explore the main risk factors for the accurate prediction of the sepsis. Multiple Imputations for filling missing data and the Synthetic Minority Oversampling (SMOTE) balancing method for balancing data are used for the data processing.
Result: Eighteen diagnostic features are used in the predictive model for early sepsis. The Random Forest model has the best performance among all the models, with an Area Under the Curve (AUC) of 87% and an F1-score (F1) of 77%. Moreover, the interpretation from the SHAP analysis is generally consistent with the current clinical situation.
Conclusion: The study revealed the relationship between these 18 clinical features and diagnostic outcomes. The results indicate that patients with laboratory values of Systolic Blood Pressure, Albumin, and Heart Rate exceeding certain thresholds are at a high likelihood of developing sepsis.

1. Introduction

Sepsis is a systemic inflammatory response syndrome caused by the invasion of pathogenic microorganisms such as bacteria. And the response of host to the infection is dysregulated, leading to life-threatening organ dysfunction. The sepsis is one of the leading causes of death in the Intensive Care Unit (ICU) (Verdonk et al., 2017; Shankar-Hari et al., 2016). Proper employments of supportive medications can reduce sepsis mortality and improve patient conditions (Hu et al., 2023), but they heavily rely on an accurate diagnosis as earlier as possible. Most hospitals utilize the traditional method like biomarkers for sepsis diagnoses (Faix, 2013). This approach takes longer and the diverse symptoms of sepsis make it challenging to diagnose clinically (Singer et al., 2016). Therefore, research on the early and accurate prediction of the sepsis is necessary and has been emphasized (Ocampo-Quintero et al., 2022; Schinkel et al., 2019). However, previous researches have some limitations. Some studies that focus on a single criterion, such as Calcitonin, Albumin, Platelets, interleukins, etc., are insufficient to accurately represent the condition of a patient and may yield biased results (Hernandez et al., 2018; Leśnik et al., 2023). Many other studies focus on the analysis of risks associated with the sepsis mortality rather than on the development of early predication models (Jiang et al., 2021; Bao et al., 2023; Ziyang et al., 2022; He et al., 2023), while those focused on the predication models are implemented based on public datasets (Johnson et al., 2016). For instance, J.S. Calvert developed a predictive model for the diagnosis of the sepsis by employing the InSight algorithm based on the MIMIC-II database in 2016 (Calvert et al., 2016). S. Nemati developed a model for predicting the episode of the sepsis 4 to 12 h prior to the clinical recognition in 2019, which is based on the MIMIC-III database (Nemati et al., 2018). In the same year, M. Scherpf et al. constructed an early prediction model for the sepsis using a neural network based on the MIMIC-III database with the AUC=0.81 (Scherpf

* Corresponding author at: School of Biomedical Engineering, Anhui Medical University, Hefei 230032, China.
E-mail address: wangyu@ahmu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.pmedr.2024.102841
Received 15 September 2023; Received in revised form 25 July 2024; Accepted 26 July 2024
Available online 2 August 2024

Table 1
Descriptive statistics of sepsis versus non-sepsis in clinical samples of adults collected for the study (April to July 2022) in Anhui Province, China (N=1968).

Feature	Nonsepsis (N=1658)	Sepsis (N=310)	*P*-value[1]
Age	68(18,90)	73 (18, 90)	< 0.001
Sex			0.05
Male	1042(62.84 %)	200(64.51 %)	
Female	616(37.15 %)	110(35.48 %)	
BMI	22.8 (12.4, 48.4)	22.5 (13.1, 38.2)	0.19
SOFA	5.1(0.1, 20)	8 (1, 20)	< 0.001
Heart Rate	95(30,200)	119 (43, 220)	< 0.001
SBP	120(30,260)	98.5 (50, 235)	< 0.001
SPO2	97(50, 100)	95 (35, 100)	< 0.001
Breath	19 (5,78)	23 (5, 58)	< 0.001
Temp	36.8 (34, 41.9)	37.3 (35, 41)	< 0.001
PaCO2	37.3 (12, 420)	37 (10.1, 220)	0.008
LympCount	1 (0.2, 80)	0.67 (0.1, 13.2)	<0.001
Albumin	33.1 (9.3, 300)	29.3 (9.6, 48.1)	0.001
Chlorine	105 (1.8, 190.6)	103.6 (64.1, 147)	0.02
Lactate	2.1 (0.5, 19.4)	3.1 (0.3, 34.3)	< 0.001
Pao2fio2	265 (38.6, 840)	224.5 (31, 840)	< 0.001
Platelets	167.9 (8, 527)	151 (6, 614)	0.01
BUN	9.05 (0.4, 457)	11,7 (2.86, 368)	0.001
Creatinine	84 (1.7, 792)	113.8 (9.34, 776)	0.004

Abbreviations: BMI,Body Mass Index; SBP, Systolic Blood Pressure; SPO2,Blood Oxygen Saturation; Temp,temperature; PaCO2,Partial Pressure of Carbon Dioxide in Arterial Blood; Pao2fio2,Oxygenation index; BUN,Blood Urea Nitrogen.
[1] The chi-square test was used for sex in the p-value, and the *t*-test was used for other features.

Table 2
Results of model performance (unbalanced/balanced) after applying multiple imputation and mean imputation for adult participants in Anhui Province, China (April-July 2022).

Model	Process	Accuracy	Precision	Recall	F1	AUC
LR	Multiple	0.69/ **0.74**	0.86/ **0.77**	0.46/ **0.69**	0.60/ **0.73**	0.84/ **0.84**
	Mean	0.67/ **0.72**	0.88/ **0.71**	0.38/ **0.73**	0.54/ **0.72**	0.77/ **0.78**
RF	Multiple	0.71/ **0.75**	0.91/ **0.88**	0.45/ **0.66**	0.60/ **0.77**	0.84/ **0.87**
	Mean	0.71/ **0.75**	0.90/ **0.83**	0.47/ **0.61**	0.62/ **0.70**	0.81/ **0.83**
KNN	Multiple	0.51/ **0.61**	0.79/ **0.59**	0.21/ **0.74**	0.16/ **0.65**	0.66/ **0.72**
	Mean	0.54/ **0.63**	0.77/ **0.63**	0.11/ **0.66**	0.19/ **0.64**	0.74/ **0.73**
DT	Multiple	0.68/ **0.70**	0.75/ **0.70**	0.54/ **0.72**	0.63/ **0.71**	0.70/ **0.76**
	Mean	0.65/ **0.65**	0.88/ **0.67**	0.38/ **0.60**	0.53/ **0.63**	0.66/ **0.73**
XGB	Multiple	0.73/ **0.74**	0.86/ **0.85**	0.54/ **0.57**	0.66/ **0.68**	0.85/ **0.86**
	Mean	0.70/ **0.74**	0.88/ **0.83**	0.47/ **0.61**	0.62/ **0.70**	0.81/ **0.79**
NN	Multiple	0.67/ **0.72**	0.86/ **0.86**	0.41/ **0.53**	0.55/ **0.65**	0.75/ **0.76**
	Mean	0.63/ **0.65**	0.76/ **0.70**	0.37/ **0.51**	0.50/ **0.59**	0.76/ **0.72**

et al., 2019), which outperformed the InSight algorithm model by J.b. Calvert et al. In 2021, R. Margherita developed an early sepsis prediction model using the MIMIC-III database, which was superior to other early prediction models available at the time (Margherita and Vincent, 2021). Across various retrospective studies, clinicians in different regions possess distinct insights into sepsis, and the developed models also consider varying clinical features. Moreover, most of studies are based on the data from a single center, thus lack the ability of generalizations for ensuring the clinical practice. Therefore, they may have difficulty interpreting final results when faced with different complex clinical situations. This study addressed the limitations by aggregating clinical data from multiple centers to construct a comprehensive database. Local clinical experts and relevant literature guided the selection of key features for the model (Elfeky et al., 2017; Madushani et al., 2022). Various established methods were employed for data processing, including machine learning techniques such as SHAP analysis (Jiang et al., 2021; Nordin et al., 2023); machine learning has been widely used in medical diagnostics such as breast cancer and brain infarction (Wang et al., 2023; Ouyang et al., 2023; Sharma et al., 2022; Singh et al., 2023; Yagin et al., 2023). Missing values were handled using multiple imputation techniques, ensuring data balance and model performance, as verified in this study.

2. Materials and methods

The entire experimental process has roughly the following steps. Firstly, the apriori analysis algorithm is used for the feature extraction from the collected dataset. The apriori algorithm identifies those most closely related features, and then, taking into account the recommendations of clinical experts as well as those used in the previous studies, eighteen features among them are identified for the construction of the model classifiers and assignment of weights. Finally, six algorithm models are introduced in order to perform predictive analyses.

2.1. Data source and study population

In this study, information was collected from 2,385 patients in the First Affiliated Hospital of Anhui Medical University and partner hospitals:

a) the basic information;
b) the life support employed;
c) the result of blood test;
d) the infection and the use of antibiotics;
e) the immunomodulatory nutritional support;
f) the use of the analgesia/sedation.

Detailed parameters including demographics values, laboratory test values, disease scores, and basic physical signs are covered. The system was divided into sepsis information collection and non-sepsis information collection columns based on whether the selection was sepsis or not. The Third International Consensus Definition of Sepsis-3 (Sepsis) provided two main criteria for the formal diagnosis of sepsis: 1) an infection must be suspected (identified by the administration of a prescription for antibiotics and sampling of body fluids for microbiological cultures); and 2) confirmation of infection, organ dysfunction manifested by a mandatory increase of 2 points or more in the Sequential Organ Failure Assessment (SOFA) score (Singer et al., 2016). The data from a total of 1968 patients, with 310 sepsis patients, are finally included in the model analysis. For external validation, the MIMIC-III and eICU databases were utilized (Johnson et al., 2016; Pollard et al., 2018). Both internal validation and external validation captured information about the features of the ICU in the 24 h prior to hospitalization. MIMIC-III served as a single-center external validation source, containing more than 60,000 hospital admissions between 2001 and 2012 (inclusive), with a total of 7,230 cases of data included in the test set. while eICU provided multicenter external validation, containing more than 139,000 hospital admissions between 2014 and 2015 (inclusive), with a total of 11,900 cases of data included in the test set. The inclusion criteria were as follows:

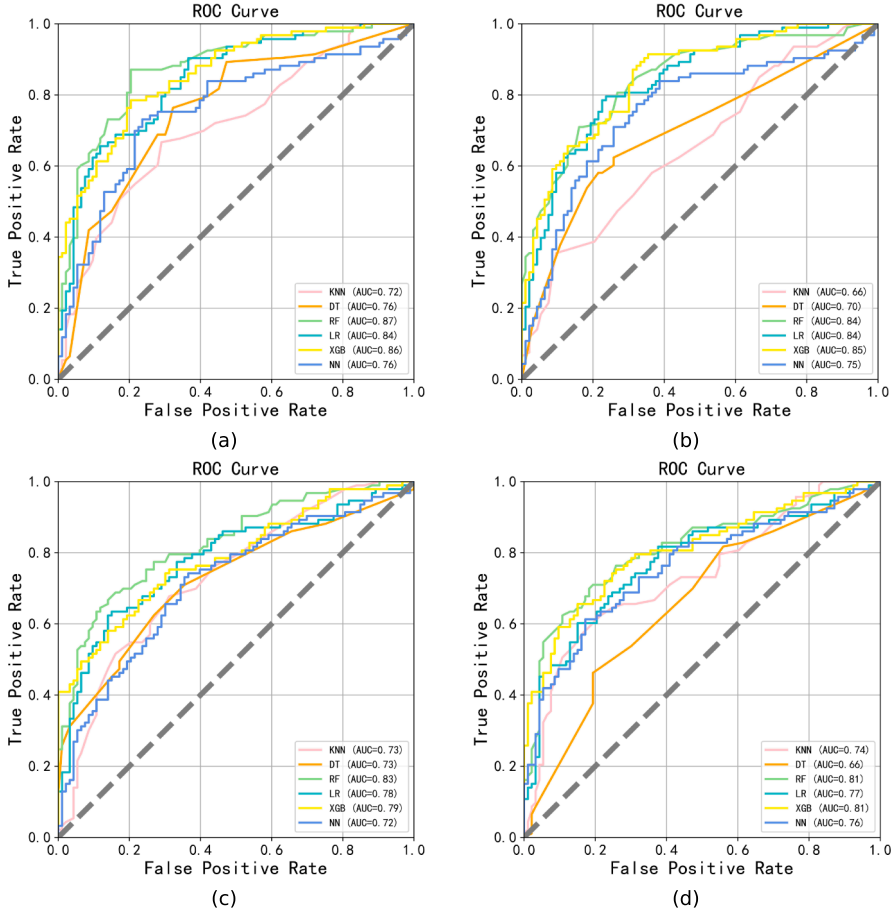

Fig. 1. The ROC curves for adult participants in Anhui Province, China (April-July 2022). (a/b) present multiple imputation (with SMOTE/without SMOTE); while (c/d) present mean imputation (with SMOTE/without SMOTE).

1) adult patients aged 18 years or older; and
2) hospitalization in the ICU for more than 24 h, with sufficient data; and
3) Patients with SOFA>=2 and suspected infection diagnosis of sepsis in the internally validated and externally validated datasets (MIMIC-III and eICU) were included according to the Third International Consensus Definition of Sepsis (Sepsis-3). One author (ZLY, ID: 11706576) accessed the databases and handled data extraction.

2.2. Feature selection and outlier handling

One of the most critical procedures before the development of the Machine Learning (ML) model is the feature selection (Kanyongo and Ezugwu, 2023), which is of beneficial to the enhance model accuracy, improve the model performance, and increase the speed of machine learning. The irrelevant features like patient ID, admission time, educational level and others which are not clearly distinguishable and cannot be coded, as well as those anomalistic features like mean arterial pressure and procalcitonin, are excluded. The handling of the remaining features is based on previous studies and ICU clinician recommendations (Ullrich et al., 2020; Li et al., 2018). For the risk factors, the probability value (p-value) is calculated by the spa chi-square test. P-value smaller than 0.05 rejects the null hypothesis and the confidence interval is satisfied.

2.3. The process of the imbalanced data

The number of non-sepsis patients is almost 6 times that of sepsis patients, and such an imbalance in the dataset will lead to a decrease in the accuracy of the model. Therefore, the resampling method is introduced for the correction. The SMOTE algorithm proposed by V. Chawla is one of the most widely used method for the resampling and has the advantage of effectively reducing overfitting compared to the simple random sampling (Aceña et al., 2022; Mutasa et al., 2020). In our study, the dataset of the sepsis patient is the minority. For a patient sample (PS-A) in this dataset, another random closest patient sample (PS-B) is selected, a new patient sample (PS-C) is generated by randomly selecting a point on the line connecting the PS-A and the PS-B, thus the PS-C is an outright new patient sample.

2.4. The choice of machine learning models

According to the previous research on the early prediction model of the sepsis (Le et al., 2019; Kijpaisalratana et al., 2022; Wang et al., 2021), a set of algorithms are identified for the application, which are Neural Network Algorithm (NN), Random Forest Algorithm (RF), K Nearest Neighbor Algorithm (KNN), Logistic Regression Algorithm (LR), Extreme Gradient Boosting Tree Algorithm (XGB), and DecisionTree (DT) Algorithm. Then models generated by above algorithms are trained using the processed dataset as described in section 2.3. Randomly selected data, which accounted for 70 % of the total data, is used as the training set, and the rest data is used as the testing set for checking the accuracy of the model. The nested resampling method is employed with using 5-fold cross validation for avoiding overfittings in the machine learning (Ounpraseuth et al., 2012; Kucheryavskiy et al., 2023). In addition, we also utilize a number of publicly available datasets as control trials to validate that models trained on our own dataset are more realistic in the clinical practice.

Table 3
Results of model performance (unbalanced/balanced) for MIMIC-III (2001–2012) and eICU (2014–2015) adult participants using multiple imputation and mean imputation..

Model	Source	Process	Accuracy	Precision	Recall	F1	AUC
LR	MIMIC-III	Multiple	0.70/**0.76**	0.78/**0.77**	0.54/**0.73**	0.64/**0.75**	0.83/**0.85**
		Mean	0.72/**0.75**	0.85/**0.75**	0.52/**0.74**	0.65/**0.75**	0.84/**0.84**
	eICU	Multiple	0.69/**0.72**	0.82/**0.71**	0.48/**0.74**	0.60/**0.73**	0.79/**0.79**
		Mean	0.69/**0.73**	0.80/**0.72**	0.50/**0.72**	0.61/**0.72**	0.78/**0.77**
RF	MIMIC-III	Multiple	0.67/**0.74**	0.85/**0.74**	0.41/**0.74**	0.55/**0.74**	0.83/**0.86**
		Mean	0.70/**0.73**	0.82/**0.78**	0.51/**0.63**	0.63/**0.70**	0.83/**0.83**
	eICU	Multiple	0.65/**0.74**	0.82/**0.73**	0.38/**0.76**	0.52/**0.74**	0.80/**0.82**
		Mean	0.73/**0.76**	0.83/**0.77**	0.58/**0.74**	0.69/**0.75**	0.79/**0.80**
KNN	MIMIC-III	Multiple	0.56/**0.62**	0.79/**0.61**	0.15/**0.66**	0.16/**0.63**	0.73/**0.75**
		Mean	0.54/**0.66**	0.88/**0.68**	0.08/**0.59**	0.15/**0.62**	0.75/**0.75**
	eICU	Multiple	0.56/**0.64**	0.77/**0.64**	0.16/**0.65**	0.27/**0.64**	0.71/**0.72**
		Mean	0.57/**0.66**	0.77/**0.65**	0.21/**0.69**	0.33/**0.67**	0.72/**0.71**
DT	MIMIC-III	Multiple	0.68/**0.71**	0.73/**072**	0.58/**0.69**	0.64/**0.70**	0.76/**0.82**
		Mean	0.70/**0.76**	0.76/**0.76**	0.59/**0.76**	0.67/**0.75**	0.70/**0.75**
	eICU	Multiple	0.64/**0.69**	0.72/**0.69**	0.46/**0.73**	0.56/**0.70**	0.74/**0.75**
		Mean	0.72/**0.76**	0.80/**0.77**	0.58/**0.74**	0.67/**0.75**	0.75/**0.76**
XGB	MIMIC-III	Multiple	0.72/**0.71**	0.79/**0.78**	0.60/**0.59**	0.68/**0.67**	0.83/**0.84**
		Mean	0.76/**0.76**	0.83/**0.80**	0.65/**0.69**	0.73/**0.74**	0.83/**0.85**
	eICU	Multiple	0.69/**0.70**	0.77/**0.74**	0.54/**0.62**	0.64/**0.68**	0.80/**0.81**
		Mean	0.75/**0.76**	0.80/**0.79**	0.66/**0.70**	0.73/**0.75**	0.84/**0.85**
NN	MIMIC-III	Multiple	0.69/**0.70**	0.79/**0.75**	0.50/**0.61**	0.62/**0.67**	0.81/**0.83**
		Mean	0.67/**0.70**	0.81/**0.80**	0.43/**0.52**	0.57/**0.63**	0.82/**0.82**
	eICU	Multiple	0.65/**0.68**	0.79/**0.63**	0.41/**0.84**	0.53/**0.72**	0.75/**0.75**
		Mean	0.71/**0.69**	0.72/**0.65**	0.69/**0.83**	0.71/**0.73**	0.74/**0.76**

2.5. The assessment of model parameters

The following metrics are used for evaluating the model by which the best model is found: classification accuracy, precision, recall, F1, and AUC (Cabot John and Gyang, 2023; Perez-Melo and Kibria, 2020). The partial dependency plots (PDPs) of the SHAP analysis are plotted to describe how the individual prediction changed as the values of the patient feature parameters changed. In the predicted output, it is true negative (TN) if sepsis patients are categorized as "Sepsis", and false negative (FN) if non-sepsis patients are categorized as sepsis patients, or if sepsis patients are categorized as non-sepsis patients. Formulas for the performance evaluation are as follow:

$$Accuracy = (TN + TP)/(TP + TN + FP + FN) \times 100\%$$

$$Precision = TP/(TP + FP) \times 100\%$$

$$Recall = TP/(TP + FN) \times 100\%$$

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \times 100\% \quad (2\text{-}5\text{-}1)$$

2.6. Statement

The project has been reviewed and approved by the Medical Ethics Committee of the First Affiliated Hospital of Anhui Medical University (Approval No. PJ2022-01-09).

3. Result

3.1. Features selection

As mentioned in section 2, a total of eighteen features are identified and included in the baseline training dataset with the results determined by the apriori algorithm (Hassan et al., 2023; Ni et al., 2022). Interpretations of the identified features are list in Table 1.

3.2. Imbalance tests and the model training

The study assessed the identified features using Pearson's correlation coefficient and found that the features exhibited low correlation with each other. Six algorithms as mentioned in section 2.4 are used to develop predictive models with the clinical parameters obtained from non-sepsis and sepsis datasets. We use both Multiple Imputation and Mean Imputation for filling the missing value in the unbalanced data. As shown in Table 2 and Fig. 1(a-d) for tests on the unbalanced data, both the low F1 and recall indicated the essential of data balancing. To ensure experimental objectivity, both the preprocessed MIMIC-III and eICU datasets, which initially had differing sepsis to non-sepsis ratios, were included in the training set at a 1:3 ratio with the local datasets.

In order to validate the generalizability of the model, external validation was implemented in this study using the MIMIC-III single-center dataset and the eICU multi-center dataset. The same method as the one used in internal validation was followed. The RF model, processed with SMOTE after Multiple Imputations, showed improved performance during external validation, achieving AUC scores of 0.86 on the MIMIC-III dataset and 0.82 on the eICU dataset. This demonstrated a certain degree of generalization and applicability for the model in this study. Table 3 and Fig. 2(a-h) presents the evaluation results of the six models based on the externally validated MIMIC-III dataset and the eICU multi-center dataset.

3.3. Analysis of sepsis risk factors

The RF model has the best performance considering all the criteria. In order to better emphasize the contribution of each feature to the model, the importance of different features is analyzed. Three most important variables of eighteen clinical features are SBP, Heart Rate and

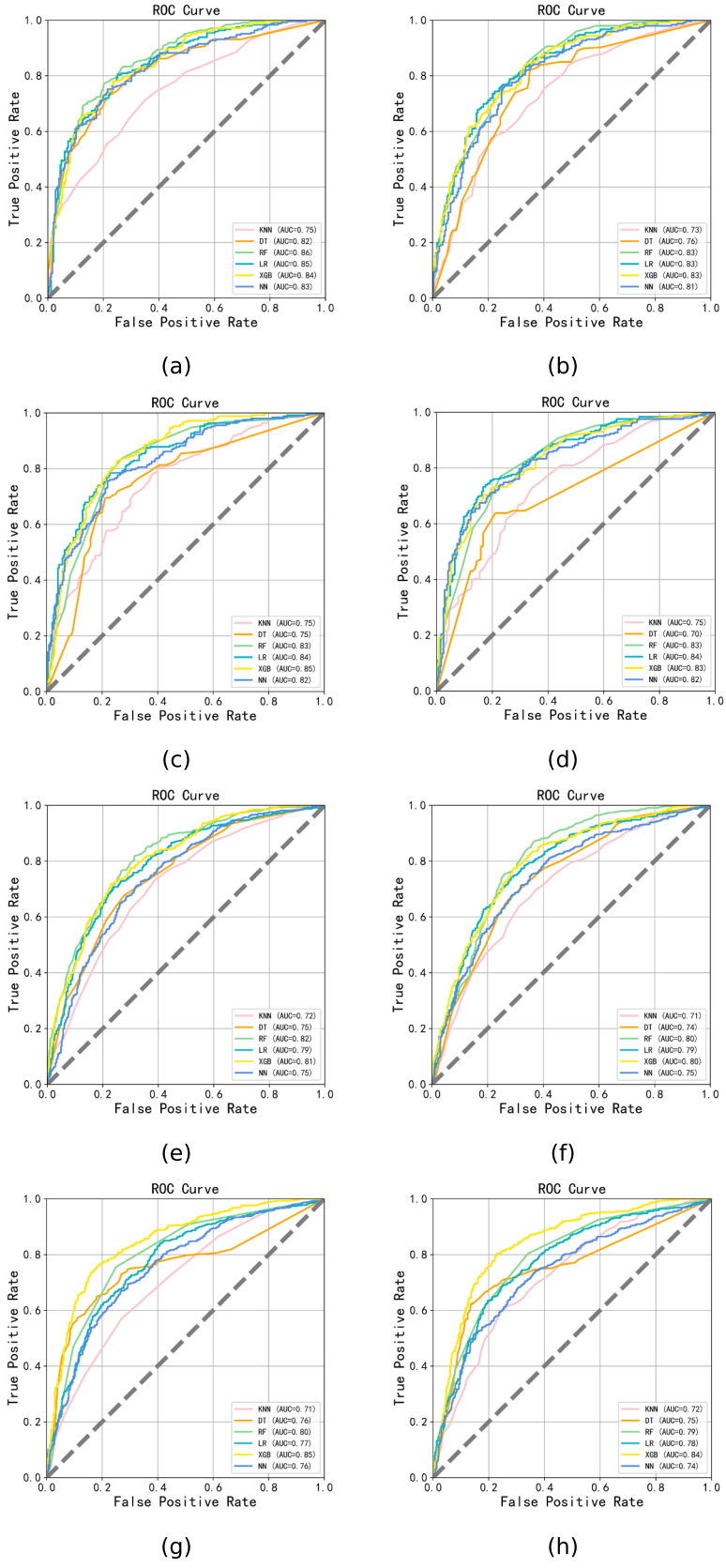

Fig. 2. The ROC curves for adult participants in MIMIC-III (2001–2012) and eICU (2014–2015) datasets. (a-d) correspond to the MIMIC-III dataset, while (e-h) correspond to the eICU dataset. For MIMIC-III: (a/b) present mean imputation (with SMOTE/without SMOTE); (c/d) present mean imputation (with SMOTE/without SMOTE); For eICU: (e/f) present mean imputation (with SMOTE/without SMOTE); (g/h) present mean imputation (with SMOTE/without SMOTE).

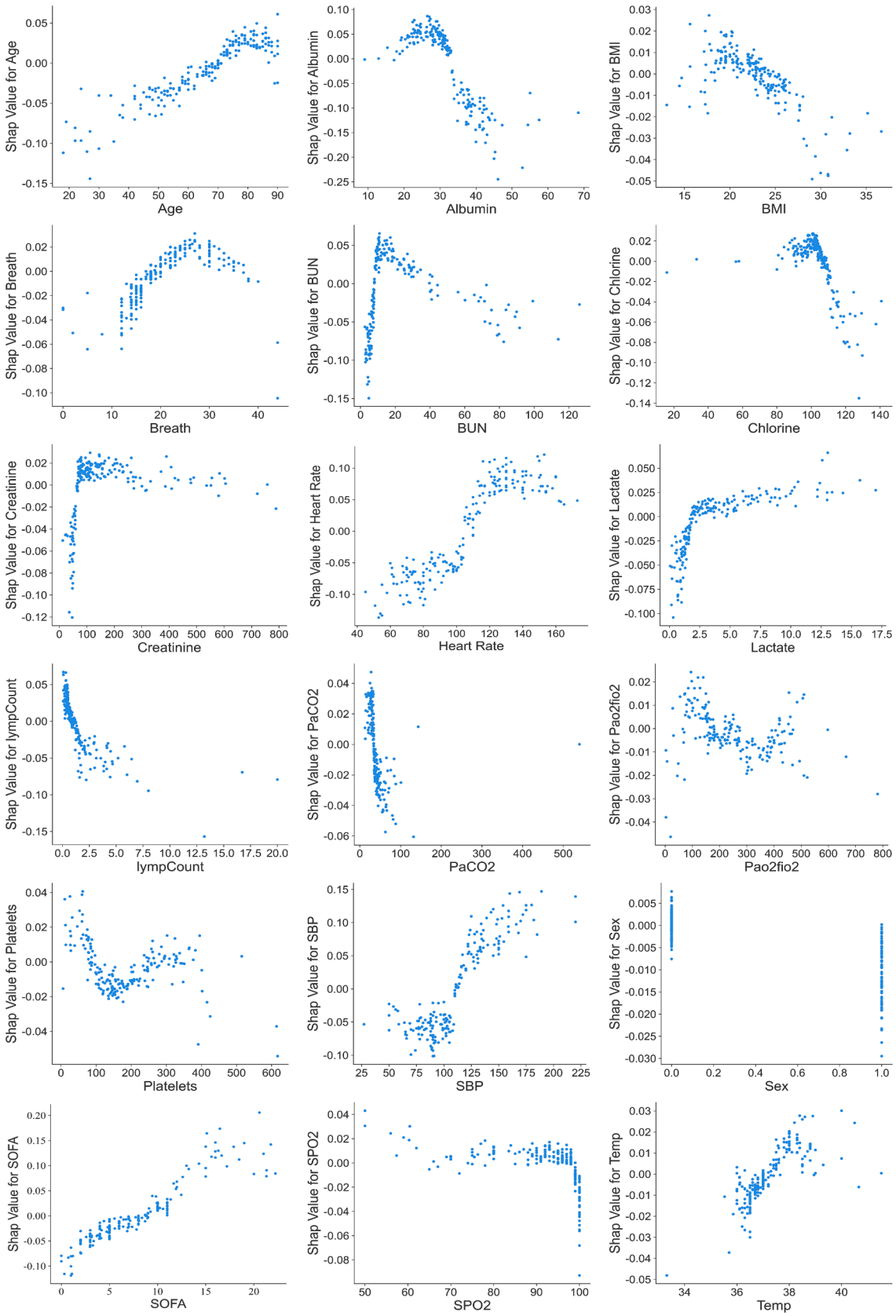

Fig. 3. Partial dependency plots of clinical features among adult participants in Anhui, China (April-July 2022). Clinical values for each feature are plotted on the X-axis, while the Y-axis shows corresponding SHAP values. The threshold is selected as 0, meaning that the risk of sepsis increases when this value is exceeded.

Albumin, while Sex, Temp, and etc., are relatively unimportant for the model. Although clinical parameters are constantly changing, the risk associated with the same parameter is not likely to vary significantly. Therefore, it is important to have "threshold"s for changes in patient risk. In order to quantify the relationship between changes in risks and features, the partial dependency graph for the SHAP analysis of the features is drawn as Fig. 3. For instance, patients with clinical values of the age in the range of more than 70, the Heart Rate greater than 100, the Lactate more than 2.5, and the SOFA score greater than 5 are more likely to develop the sepsis.

4. Discussion and Conclusion

The research aims to reduce the mortality rate of sepsis by improving the accuracy of early diagnosis. Currently, the diagnosis of sepsis relies primarily on physiological indicators and imaging methods, supplemented by the clinical judgment. However, the diagnostic approach is prone to be biased and time-consuming. With the advent of the big data era, machine learning technology can learn potential patterns from a large amount of data and apply them to the development of new predictive models. These models will undergo multiple rounds of training and learning, with the aim of assisting physicians in early diagnosis and improving the accuracy of sepsis diagnosis.

While previous studies predominantly focused on factors influencing mortality in sepsis, there was a significant lack of research specifically addressing early prediction of sepsis. Existing literature often relied on large open-center datasets like MIMIC-III. However, the generalizability of models based on single-center data was questioned. Although the eICU dataset was derived from multiple centers, few studies utilized it. These studies highlighted limitations, such as the lack of necessary prospective clinical validation, casting doubts on the model's applicability (Ocampo-Quintero et al., 2022; Schinkel et al., 2019; O'Reilly et al., 2023; Komorowski et al., 2022). Considering such limitations, this study collected patient data from the First Affiliated Hospital of Anhui Medical University and partner hospitals. The dataset was then processed using statistical analysis and machine learning methods, an early prediction model was constructed, and the results were clinically validated. This application of polycentric data greatly improves the generalization performance of the model, and ensure that the model is more suitable for the clinical practice. In terms of model effectiveness, the model demonstrated a good performance on public datasets indicating that the model was applicable to local datasets as well as the public datasets MIMIC-III and eICU, enhancing its generalizability and reliability.

Typically, the number of sepsis patients is much lower compared with those non-sepsis, while the study needs to use similar proportions of data for the analysis. Therefore, the unbalanced data is processed by the SMOTE resampling method. We also demonstrate the necessary of this procedure. In the result, our model demonstrates an 87 % AUC value which proves its reliability. We also verify the feasibility of the model containing 18 key features in practice.

The study is still on its primary stage and further improvement is required. The interpolation method is employed to fill in some of the missing values, which could affect the accuracy of the model to a certain extent, though we have already considered the approach to minimized the impact. The data is collected from hospitals in a single province, and more data from other areas is required to be covered in the future to further improve the accuracy.

5. Summary

The study develops the model for the early prediction of the sepsis in ICU patients. Totally eighteen features are identified and six kinds of algorithms are utilized for the analysis. The RF has the best performance with the highest F1, recall, and AUC. From the result, special attention is needed for ICU patients who have 18 feature parameters of the included model exceeding the thresholds as they are highly probable to be sepsis. The development of such a type of model could help ICU physicians to recognize sepsis patients and intervene earlier to reduce mortality.

Funding

This study was supported by the Research Fund of Anhui Institute of translational medicine (2022zhyxC26).

Author contributions

Zhou Luyao was responsible for the paper writing and experimental design, Shao Min was responsible for the data collection and validation, Wang Cui was responsible for the questionnaire design, and Wang Yu was responsible for the questionnaire system construction.

Informed consent

The study protocol was approved by the research ethics board of our institution, and all patients provided informed consent before their medical records was reviewed. The performance of this study conformed with the Declaration of Helsinki.

CRediT authorship contribution statement

Luyao Zhou: Writing – original draft. **Min Shao:** Data curation. **Cui Wang:** Investigation. **Yu Wang:** Software.

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.

Data availability

The original contributions presented in this study are included in the article, further inquiries can be directed to the corresponding author/s.

References

Aceña, V., Martín de Diego, I., Fernández, R., Moguerza, J.M., 2022. Minimally overfitted learners: A general framework for ensemble learning. Knowl.-Based Syst. 254, 109669 https://doi.org/10.1016/j.knosys.2022.109669.

Bao, C., Deng, F., Zhao, S., 2023. Machine-learning models for prediction of sepsis patients mortality. Med. Intensiva 47 (6), 315–325. https://doi.org/10.1016/j.medin.2022.06.004.

Cabot John, H., Gyang, R.E., 2023. Evaluating prediction model performance. Surgery. https://doi.org/10.1016/J.SURG.2023.05.023.

Calvert, J.S., Price, D.A., Chettipally, U.K., Barton, C.W., Feldman, M.D., Hoffman, J.L., et al., 2016. A computational approach to early sepsis detection. Comput. Biol. Med. 74, 69–73. https://doi.org/10.1016/j.compbiomed.2016.05.003.

Elfeky, S., Golabi, P., Otgonsuren, M., Djurkovic, S., Schmidt, M.E., Younossi, Z.M., 2017. The epidemiologic features, temporal trends, predictors of death, and discharge disposition in patients with a diagnosis of sepsis: A cross-sectional retrospective cohort study. J. Crit. Care 39, 48–55. https://doi.org/10.1016/j.jcrc.2017.01.006.

Faix, J.D., 2013. Biomarkers of sepsis. Crit. Rev. Clin. Lab. Sci. 50 (1), 23–36. https://doi.org/10.3109/10408363.2013.764490.

Hassan, M.M., Karim, A., Mollick, S., Azam, S., Ignatious, E., Haque, A.S.M.F.A., 2023. An Apriori Algorithm-Based Association Rule Analysis to detect Human Suicidal Behaviour. Procedia Comput. Sci. 219, 1279–1288. https://doi.org/10.1016/j.procs.2023.01.412.

He, Y., Liu, Y., Liu, Y., He, H., Liu, W., Huang, D., et al., 2023. A machine-learning approach for prediction of hospital mortality in cancer-related sepsis. Clinical eHealth 6, 17–23. https://doi.org/10.1016/j.ceh.2023.06.003.

Hernandez, G., Bellomo, R., Bakker, J., 2018. The ten pitfalls of lactate clearance in sepsis. Intensive Care Med. 45 (1), 82–85. https://doi.org/10.1007/s00134-018-5213-x.

Hu, W., Chen, H., Wang, H., Peng, Q., Wang, J., Huang, W., et al., 2023. Identifying high-risk phenotypes and associated harms of delayed time-to-antibiotics in patients with ICU onset sepsis: A retrospective cohort study. J. Crit. Care 74. https://doi.org/10.1016/j.jcrc.2022.154221.

Jiang, Z., Bo, L., Xu, Z., Song, Y., Wang, J., Wen, P., et al., 2021. An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission. Comput. Methods Programs Biomed. 204 https://doi.org/10.1016/j.cmpb.2021.106040.

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. Sci. Data 3(1). doi: 10.1038/sdata.2016.35.

Johnson, A., Pollard, T., Shen, L., et al., 2016. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035. https://doi.org/10.1038/sdata.2016.35.

Kanyongo, W., Ezugwu, A.E., 2023. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. Inf. Med. Unlocked 38. https://doi.org/10.1016/j.imu.2023.101232.

Kijpaisalratana, N., Sanglertsinlapachai, D., Techaratsami, S., Musikatavorn, K., Saoraya, J., 2022. Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study. Int. J. Med. Inf. 160 https://doi.org/10.1016/j.ijmedinf.2022.104689.

Komorowski, M., Green, A., Tatham, K.C., Seymour, C., Antcliffe, D., 2022. Sepsis biomarkers and diagnostic tools with a focus on machine learning. EBioMedicine 86, 104394. https://doi.org/10.1016/j.ebiom.2022.104394.

Kucheryavskiy, S., Rodionova, O., Pomerantsev, A., 2023. Procrustes cross-validation of multivariate regression models. Anal. Chim. Acta 1255, 341096. https://doi.org/10.1016/j.aca.2023.341096.

Le, S., Hoffman, J., Barton, C., Fitzgerald, J.C., Allen, A., Pellegrini, E., et al. (2019). Pediatric Severe Sepsis Prediction Using Machine Learning. 7. doi: 10.3389/fped.2019.00413.

Leśnik, P., Janc, J., Mierzchala-Pasierb, M., Tański, W., Wierciński, J., Łysenko, L., 2023. Interleukin-7 and interleukin-15 as prognostic biomarkers in sepsis and septic shock: Correlation with inflammatory markers and mortality. Cytokine 169. https://doi.org/10.1016/j.cyto.2023.156277.

Li, J.-L., Li, G., Jing, X.-Z., Li, Y.-F., Ye, Q.-Y., Jia, H.-H., et al., 2018. Assessment of clinical sepsis-associated biomarkers in a septic mouse model. J. Int. Med. Res. 46 (6), 2410–2422. https://doi.org/10.1177/0300060518764717.

Madushani, R.W.M.A., Patel, V., Loftus, T., Ren, Y., Li, H.J., Velez, L., et al., 2022. Early Biomarker Signatures in Surgical Sepsis. J. Surg. Res. 277, 372–383. https://doi.org/10.1016/j.jss.2022.04.052.

Margherita, R., Vincent, F., 2021. MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis. PLoS One 16 (5).

Mutasa, S., Sun, S., Ha, R., 2020. Understanding artificial intelligence based radiology studies: What is overfitting? Clin. Imaging 65, 96–99. https://doi.org/10.1016/j.clinimag.2020.04.025.

Nemati, S., Holder, A., Razmi, F., Stanley, M.D., Clifford, G.D., Buchman, T.G., 2018. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit. Care Med. 46 (4), 547–553. https://doi.org/10.1097/ccm.0000000000002936.

Ni, J., Li, L., Wang, Y., Ji, C., and Zheng, C. (2022). MDSCMF: Matrix Decomposition and Similarity-Constrained Matrix Factorization for miRNA– Disease Association Prediction. 13(6), 1021. doi:10.3390/genes13061021.

Nordin, N., Zainol, Z., Mohd Noor, M.H., Chan, L.F., 2023. An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach. Asian J. Psychiatr. 79 https://doi.org/10.1016/j.ajp.2022.103316.

Ocampo-Quintero, N., Vidal-Cortés, P., del Río Carbajo, L., Fdez-Riverola, F., Reboiro-Jato, M., Glez-Peña, D., 2022. Enhancing sepsis management through machine learning techniques: A review. Med. Intensiva 46 (3), 140–156. https://doi.org/10.1016/j.medin.2020.04.003.

O'Reilly, D., McGrath, J., Martin-Loeches, I., 2023. Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future. J. Intensive Med. https://doi.org/10.1016/j.jointm.2023.10.001.

Ounpraseuth, S., Lensing, S.Y., Spencer, H.J., Kodell, R.L., 2012. Estimating misclassification error: a closer look at cross-validation based methods. BMC. Res. Notes 5 (1), 656. https://doi.org/10.1186/1756-0500-5-656.

Ouyang, Y., Cheng, M., He, B., Zhang, F., Ouyang, W., Zhao, J., et al., 2023. Interpretable machine learning models for predicting in-hospital death in patients in the intensive care unit with cerebral infarction. Comput. Methods Programs Biomed. 231 https://doi.org/10.1016/j.cmpb.2023.107431.

Perez-Melo, S., Kibria, B.M.G., 2020. On Some Test Statistics for Testing the Regression Coefficients in Presence of Multicollinearity: A Simulation Study. Stats 3 (1), 40–55. https://doi.org/10.3390/stats3010005.

Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O., 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci. Data. https://doi.org/10.1038/sdata.2018.178.

Scherpf, M., Gräßer, F., Malberg, H., Zaunseder, S., 2019. Predicting sepsis with a recurrent neural network using the MIMIC III database. Comput. Biol. Med. 113 https://doi.org/10.1016/j.compbiomed.2019.103395.

Schinkel, M., Paranjape, K., Nannan Panday, R.S., Skyttberg, N., Nanayakkara, P.W.B., 2019. Clinical applications of artificial intelligence in sepsis: A narrative review. Comput. Biol. Med. 115 https://doi.org/10.1016/j.compbiomed.2019.103488.

Shankar-Hari, M., Phillips, G.S., Levy, M.L., Seymour, C.W., Liu, V.X., Deutschman, C.S., et al., 2016. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock. JAMA 315 (8). https://doi.org/10.1001/jama.2016.0289.

Sharma, D., Kumar, R., Jain, A., 2022. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. Measurement: Sensors 24. https://doi.org/10.1016/j.measen.2022.100560.

Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., et al., 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315 (8). https://doi.org/10.1001/jama.2016.0287.

Singh, L.K., Khanna, M., Singh, R., 2023. Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. Adv. Eng. Softw. 175, 103338 https://doi.org/10.1016/j.advengsoft.2022.103338.

Ullrich, E., Heidinger, P., Soh, J., Villanova, L., Grabuschnig, S., Bachler, T., et al., 2020. Evaluation of host-based molecular markers for the early detection of human sepsis. J. Biotechnol. 310, 80–88. https://doi.org/10.1016/j.jbiotec.2020.01.013.

Verdonk, F., Blet, A., Mebazaa, A., 2017. The new sepsis definition. Curr. Opin. Anaesthesiol. 30 (2), 200–204. https://doi.org/10.1097/aco.0000000000000446.

Wang, D., Li, J., Sun, Y., Ding, X., Zhang, X., Liu, S., et al., 2021. A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. Front. Public Health 9. https://doi.org/10.3389/fpubh.2021.754348.

Wang, X., Ren, H., Ren, J., Song, W., Qiao, Y., Ren, Z., et al., 2023. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput. Methods Programs Biomed. 230 https://doi.org/10.1016/j.cmpb.2023.107340.

Yagin, F.H., Gülü, M., Gormez, Y., Castañeda-Babarro, A., Colak, C., Greco, G., et al., 2023. Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique. Appl. Sci. 13 (6) https://doi.org/10.3390/app13063875.

Ziyang, W., Yushan,L., Zidu, X., Yaowen, G., Jiao, L. (2022). Comparison of Mortality Predictive Models of Sepsis Patients Based on Machine Learning. 37(3), 201-209. doi: 10.24920/004102.