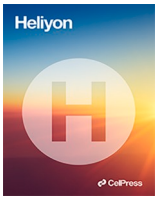
Heliyon 10 (2024) e33826

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Novel approach exploring the correlation between presepsin and routine laboratory parameters using explainable artificial intelligence

Jae-Seung Jeong [a], Tak Ho Kang [b], Hyunsu Ju [c,**], Chi-Hyun Cho [d,*]

[a] Division of Artificial Intelligence Convergence Engineering, Sahmyook University, South Korea
[b] Department of Laboratory Medicine, College of Medicine, Korea University Anam Hospital, South Korea
[c] Post-Silicon Semiconductor Institute, Korea Institute of Science and Technology, South Korea
[d] Department of Laboratory Medicine, College of Medicine, Korea University Ansan Hospital, South Korea

ARTICLE INFO

Keywords:
Presepsin
Routine laboratory parameters
Machine learning classifiers
Missing data management
Explainable artificial intelligence (XAI)

ABSTRACT

Although presepsin, a crucial biomarker for the diagnosis and management of sepsis, has gained prominence in contemporary medical research, its relationship with routine laboratory parameters, including demographic data and hospital blood test data, remains underexplored. This study integrates machine learning with explainable artificial intelligence (XAI) to provide insights into the relationship between presepsin and these parameters. Advanced machine learning classifiers provide a multilateral view of data and play an important role in highlighting the interrelationships between presepsin and other parameters. XAI enhances analysis by ensuring transparency in the model's decisions, especially in selecting key parameters that significantly enhance classification accuracy. Utilizing XAI, this study successfully identified critical parameters that increased the predictive accuracy for sepsis patients, achieving a remarkable ROC AUC of 0.97 and an accuracy of 0.94. This breakthrough is possibly attributed to the comprehensive utilization of XAI in refining parameter selection, thus leading to these significant predictive metrics. The presence of missing data in datasets is another concern; this study addresses it by employing Extreme Gradient Boosting (XGBoost) to manage missing data, effectively mitigating potential biases while preserving both the accuracy and relevance of the results. The perspective of examining data from higher dimensions using machine learning transcends traditional observation and analysis. The findings of this study hold the potential to enhance patient diagnoses and treatment, underscoring the value of merging traditional research methods with advanced analytical tools.

1. Introduction

The identification of biomarkers for specific human conditions or diseases has significant potential for predicting health outcomes and guiding treatment. Traditionally, such research has relied on statistical analyses of clinical data, using methods such as t-tests, analysis of variance (ANOVA), and regression models. With the rise of artificial intelligence (AI), tasks such as diagnostics and

* Corresponding author.
** Corresponding author.
E-mail addresses: hyunsuju@kist.re.kr (H. Ju), 9754091@korea.ac.kr (C.-H. Cho).

https://doi.org/10.1016/j.heliyon.2024.e33826
Received 18 June 2024; Received in revised form 27 June 2024; Accepted 27 June 2024
Available online 1 July 2024

therapeutic recommendations frequently employ machine learning techniques, including specialized areas such as presepsin biomarker analysis [1–8]. Analyses using AI allow data to be observed in high dimensions through a training process. Consequently, raw data and patterns that were previously considered meaningless can be transformed into higher-dimensional representations, allowing for the derivation of essential insights. Building upon these advancements, recent studies have demonstrated that integrating advanced machine learning algorithms with traditional statistical methods can uncover complex patterns and correlations that are not apparent through conventional analysis alone. For example, the combination of XGBoost and SHAP values has been proven effective in revealing significant features that influence prediction outcomes in medical datasets. Machine learning algorithms have been successfully applied to various medical datasets for tasks such as disease diagnosis, risk prediction, and treatment response evaluation. For instance, Ambale-Venkatesh et al. employed Random Forest and XGBoost algorithms to identify key risk factors for cardiovascular diseases using electronic health record data [9]. Similarly, Shouval et al. demonstrated the utility of machine learning algorithms for clinical predictive modeling in stem cell transplantation [10]. Furthermore, recent studies have employed advanced AI techniques to predict various health outcomes. Kwon et al. developed an explainable AI (XAI) model to predict in-hospital cardiac arrest, demonstrating high predictive performance and interpretability [11,12]. Similarly, Lee et al. utilized an XAI approach to predict acute kidney injury after cardiovascular surgery, showcasing the potential of these methods in clinical decision support [13,14].

Presepsin levels increase in response to bacterial infections and decrease after healing or efficient treatment, as evidenced by several studies [15–17]. Thus, presepsin is the most promising emerging biomarker for sepsis [18]. Several studies have demonstrated the clinical utility of presepsin in the early diagnosis and prognosis of sepsis. Compared to conventional biomarkers such as procalcitonin and C-reactive protein, presepsin has shown higher sensitivity and specificity in detecting sepsis, particularly in the early stages of the disease [19,20]. Although extensive research has been conducted on presepsin measurement methods [17,21,22], the relationship between presepsin levels and routine laboratory parameters, encompassing demographic data and hospital blood test data, has not yet been explored [23–25]. Understanding this relationship is clinically significant; it can help identify the conditions that influence presepsin levels and consequently provide a basis for setting condition-specific cutoff values. Moreover, clarity in the relationship between the presepsin levels and the routine laboratory parameters may result in more accurate diagnoses and prognostic predictions for sepsis and other infectious diseases [16,26]. To investigate these aspects, this study employs traditional statistical methods such as correlation and t-tests as well as advanced machine learning algorithms such as k-nearest neighbors (k-NN) [27], naive Bayes classifier [28], Random Forest [29], and Extreme Gradient Boosting (XGBoost) [30]. Moreover, missing data frequently occur in medical research as not all patients undergo the same set of tests; the specific tests administered vary depending on the patient's condition, situation, and the objectives of the medical visit. Therefore, data being missing in these datasets is almost inevitable [31]. To resolve this issue, the XGBoost algorithm is employed to mitigate potential biases while ensuring the accuracy and relevance of the results. Shapley Additive Explanations (SHAP) [32] are incorporated into advanced machine learning algorithms to clarify the significance and relevance of each of the routine laboratory parameters in relation to the presepsin level. By comparing traditional statistical methods with advanced machine learning algorithms, this study intends to reveal the unexplored relationship between presepsin levels and the routine laboratory parameters. Moreover, the identified correlations may provide novel insights into the pathophysiological mechanisms underlying sepsis and other infectious diseases [33,34], paving the way for more precise and status-specific diagnostic and predictive tools. Additionally, by using XGBoost to process missing data and SHAP to interpret machine

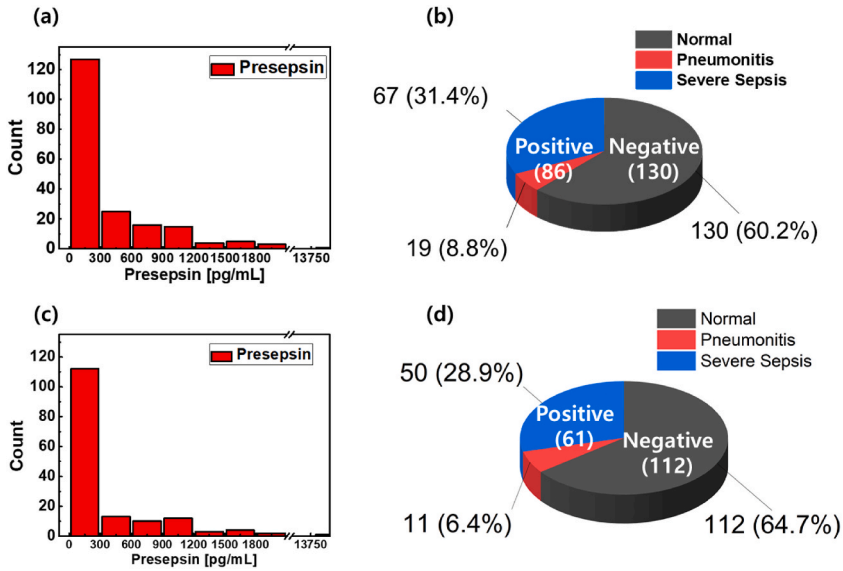


Fig. 1. Distribution and classification of presepsin levels in participants (with and w/o missing data). (a) Histogram illustrating the distribution of presepsin values among participants. (b) Categorizing participants into Positive and Negative groups based on their presepsin values. (c) Histogram illustrating the distribution of presepsin values among participants without missing data. (d) Categorizing participants into Positive, Negative groups based on their presepsin values without missing data.

learning models, this study seeks to provide a robust and transparent analytical framework that can be extrapolated to other biomarker studies, potentially accelerating the development of personalized diagnostic and treatment strategies.

2. Materials and methods

2.1. Data collection

This study was approved by the Institutional Review Board of Korea University Ansan Hospital (no. 2022AS0167; Ansan-si, Republic of Korea), and the requirement for written informed consent was waived owing to the retrospective nature of the study. Residual serum samples from 216 patients who visited the hospital between April and June 2022 were used. The distribution of the participants' presepsin levels is shown in Fig. 1. Presepsin levels were measured using HISCL presepsin analysis (Sysmex, Kobe, Japan) in a HISCL 5000 autoanalyzer (Sysmex) based on a delayed phase 1 sandwich chemiluminescence enzyme immunoassay. For sepsis, the manufacturer's suggested reference range for presepsin was up to 333.0 pg/mL [22,35]. Additionally, the distribution of presepsin levels and the number of patients diagnosed with Normal, Pneumonitis, and Severe Sepsis were analyzed for a subset of 53 samples (Supplementary Figure 1). These 53 samples were randomly selected from the original dataset and served as an independent test set for model evaluation. The presepsin levels of these samples were visualized using a histogram, and the patient composition was represented using a pie chart. The composition of this test set, in terms of presepsin levels and clinical diagnoses, was similar to that of the full dataset, ensuring its representativeness (Supplementary Figure 1.).

A retrospective review of the patients' medical records provided demographic, clinical, and laboratory data. Specifically, hepatic ultrasound, abdominal computed tomography (CT), magnetic resonance imaging (MRI), bacterial culture test results (from blood, respiratory, urine samples), hemoglobin (Hb), while blood cell (WBC), platelet count, neutrophil %, absolute neutrophil count, lymphocyte %, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), protein, albumin, bilirubin (total), bilirubin (direct), aspartate transaminase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), blood urea nitrogen (BUN), creatinine, height, gender, age, uric acid, lactate dehydrogenase (LD), creatine kinase (CK), amylase, high-density lipoprotein (HDL)-cholesterol, low-density lipoprotein (LDL)-cholesterol, and triglyceride results and records of underlying diseases were reviewed. The standard reference ranges, units, and measuring equipment for each laboratory parameter are detailed in Supplementary Table 1.

2.2. Simple statistical analysis: correlation, linear regression, and p-value correction

In the field of data analysis, correlation and regression are fundamental techniques commonly used to explore and model relationships between variables. Correlation analysis provides a measure of the strength and direction of the linear relationship between two continuous variables. This relationship is quantified using the Pearson correlation coefficient [36], r. Spanning from -1 to $+1$, this coefficient reflects the degree of linear association between the variables, with its magnitude indicating its strength and its sign denoting its direction. Statistical significance of these relationships was assessed by calculating p-values for each variable. A p-value represents the probability of observing the given result, or a more extreme one, assuming the null hypothesis is true. In this study, p-values were computed for each routine laboratory parameter in relation to presepsin levels. As shown in Supplementary Table 2, most parameters exhibited very small p-values (ranging from 3.046e-53 to 9.957e-12), suggesting a highly significant association with presepsin. However, when multiple hypotheses are tested simultaneously, as in this case with numerous laboratory parameters, the likelihood of obtaining false positives (Type I errors) increases. To mitigate this issue, the Benjamini-Hochberg procedure, a widely used method for controlling the False Discovery Rate (FDR) [37–39], was employed. This procedure adjusts the p-values to account for multiple comparisons, yielding corrected p-values that maintain the desired FDR. After applying the Benjamini-Hochberg correction, the adjusted p-values remained highly significant for most parameters (ranging from 5.178e-52 to 1.411e-11), confirming the robustness of the associations (Supplementary Table 2). These corrected p-values provide a more reliable basis for selecting features that are significantly related to presepsin levels. In this study, features with corrected p-values below the significance threshold of 0.05 were considered relevant and were included in subsequent analyses. This approach ensures that the selected features have a high probability of being truly associated with the target variable, rather than being false positives. By contrast, regression delves deeper by modeling the relationship between a dependent variable and one or more independent variables. Linear regression, which is the most straightforward form, captures potential linear relationships and expresses them using an equation. This equation, characterized by its slope and intercept, provides a linear approximation of how the dependent variable changes with respect to the predictors. Acquiring insights into these statistical techniques, including p-value calculation and correction for multiple comparisons, is essential for constructing the underlying knowledge of data analytics and setting the stage for more advanced methodologies. By incorporating these foundational concepts, this study aims to provide a robust and statistically sound analysis of the relationship between presepsin and routine laboratory parameters.

2.3. Data preprocessing

The data were analyzed by organizing 216 subsets, which were composed of all samples, and 173 subsets, excluding samples with missing parameters (ESR, LD, CK). The subset with 216 samples was labeled as "with missing data," whereas the subset with 173 samples was named "w/o missing data". These subsets were divided into training and test sets at a ratio of 7:3; learning and testing were conducted accordingly.

2.4. Machine learning algorithms

Various machine learning classifiers have been used to identify patterns within a dataset. The k-nearest neighbor (k-NN) function is a nonparametric, instance-based learning algorithm. In this approach, a sample's classification is determined by the majority class of its k closest training samples, with distance metrics such as the Euclidean distance commonly determining this closeness. Logistic regression is a statistical approach that models the relationship between a binary outcome and its independent predictors, outputting a probability score for an instance belonging to a particular category. The naive Bayes classifier operates on the Bayes theorem and assumes independence among the predictors. This indicates that each feature contributes independently to the outcome, a principle that often translates into efficacy in high-dimensional datasets. Random Forest is an ensemble learning method that combines multiple decision trees to produce a singular, more accurate, and stable prediction. The use of bootstrapping generates various datasets and subsequently builds a tree for each dataset, inherently reducing the variance. Among these classifiers, the best performer is eXtreme Gradient Boosting (XGBoost). This optimized gradient boosting algorithm is designed for efficiency and flexibility with notable adaptability to various prediction problems, including regression, classification, and ranking. The distinctive characteristics of XGBoost include its intrinsic ability to handle missing values, built-in cross-validation, and resilience to overfitting. The integration of a gradient descent algorithm with a regularized boosting technique ensures the derivation of optimized results, making XGBoost a central figure in the analytical process. The five machine learning models were selected based on their popularity in similar research [30,40–45], their ability to handle the specific characteristics of our dataset, and their diverse algorithmic approaches. Additionally, these models represent some of the most straightforward and widely used techniques in the field of machine learning, making them accessible for implementation and interpretation.

In this study, several machine learning models were employed with their hyperparameters tuned to optimize performance on the dataset. For the k-NN model, $k = 5$ was used to strike a balance between overfitting and underfitting, and the Euclidean distance metric was chosen as it is a common default. The effectiveness of k-NN in similar contexts is well-documented [46,47]. The logistic regression model was implemented with L2 regularization and $C = 1$, which are standard default settings that often yield good results. This model is chosen for its interpretability and efficiency in binary classification problems [48,49]. For the naive Bayes classifier, Laplace smoothing (alpha = 1) was used to handle potential zero probabilities in the dataset, which operates effectively in high-dimensional spaces due to its simplicity [50,51]. In the case of Random Forest, n_estimators were set to 100 to create a sufficiently large number of trees, max_depth to None to allow the trees to grow to their maximum depth, and min_samples_split to 2, which is the default value. The utility of Random Forest in similar applications is supported by existing literature [29,52]. For the XGBoost model, a learning rate (eta) of 0.3, a maximum depth of trees (max_depth) of 6, a minimum loss reduction (gamma) of 0, and a subsample ratio of columns (colsample_bytree) of 1 were used. These are the default values for XGBoost and often provide a good starting point for the model. The advantages of XGBoost in similar research are well-documented [30,53]. While primarily default hyperparameter settings were used, these values were chosen based on their proven effectiveness across a wide range of datasets and problem types. Future work could involve more extensive hyperparameter tuning using techniques such as grid search or random search to potentially improve model performance further. However, the current settings provided satisfactory results for the purposes of this study.

2.5. Combining explainable artificial intelligence with XGBoost

The need for clearer explanations of machine learning decisions has led to the use of Explainable Artificial Intelligence (XAI) methods, particularly complex models such as XGBoost. XAI methods help clarify how a model makes decisions, which is important for both researchers and healthcare providers. XGBoost offers the advantage of facilitating a clearer understanding of how the model operates through its decision tree structure; additionally, it effectively handles missing data during the learning process. This attribute is particularly valuable in medical settings where data can be incomplete.

SHAP uses game theory principles to provide a unified measure of feature importance. The SHAP values break down a prediction to demonstrate how each feature affects it. SHAP provides the parameter's importance for prediction and reveals whether these increase or decrease the likelihood of the prediction outcomes. This incorporation of SHAP with XGBoost explains the contributions of individual parameters to the predictions, facilitating an understanding of their effects on the predictive process. In conclusion, the

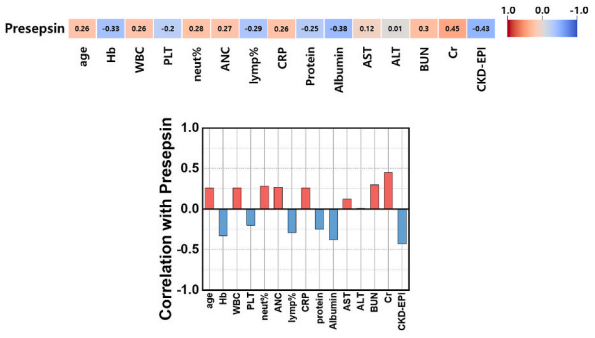


Fig. 2. Correlation between presepsin levels and the routine laboratory parameters in participants.

integration of XAI methods such as SHAP enables XGBoost to serve as a robust predictive tool while transforming it into a transparent and easily understandable resource.

3. Results

3.1. Simple analysis: correlations and linear relationships

Fig. 2 shows the relationship between presepsin levels and the routine laboratory parameters. Creatinine (Cr) has a correlation coefficient of 0.45, suggesting a moderately positive relationship [54]. The CKD-EPI shows a −0.43 correlation, indicating a moderately negative relationship. Other factors like age, WBC, and CRP had coefficients between 0.25 and 0.3, suggesting moderately positive relationships. Hb and Albumin had weaker correlations, ranging from −0.38 to 0.12. Typically, coefficients greater than 0.7 are considered to indicate strong relationships, and those less than 0.7 are judged as insignificant [54]. Considering that the highest correlation is 0.45, analyzing interpretable relationships requires advanced machine-learning techniques to discover patterns in high dimensions.

A comparison between the negative and positive groups based on presepsin levels is shown in Table 1. This comparison involves analyzing the average, minimum, and maximum values of each parameter. The p-values presented in Table 1 were obtained from independent t-tests comparing the means of each laboratory parameter between the presepsin negative and positive groups. A p-value less than 0.05 indicates a statistically significant difference between the two groups for the corresponding parameter. The resulting p-values, ranging from approximately 1E-28 to 0.000499, clearly highlight the statistical differences between the two groups for these parameters. A closer look at the range of values for each parameter suggests that they cannot be used as the sole criteria for cleanly separating the groups. Overlapping ranges and subtle variations indicate the limitations of relying solely on simple statistical analyses to draw definitive conclusions. Therefore, more advanced methods are required for a deeper understanding.

3.2. Explainable artificial intelligence and classification algorithms

In the analysis of 173 complete datasets, various classifiers were used to sort the data into presepsin groups. Using a threshold of 333, the samples were marked as Positive (greater than 333) or Negative (333 or lower). Table 2 summarizes the performance of each algorithm using metrics such as precision, recall, F1 score, Receiver Operating Characteristic Area Under Curve (ROC AUC), and accuracy (based on 5-fold cross-validation). The k-NN classifier had stable scores for precision, recall, and F1 score, all at 0.77. The ROC AUC of 0.81 suggests moderate group separation with an overall accuracy of 0.80. Logistic regression stood out, with a precision of 0.93, a recall of 0.92, and an F1 score of 0.93. Importantly, its ROC AUC of 0.98 indicated excellent ability to distinguish between Positive and Negative groups and it achieved an overall accuracy of 0.91. The Naive Bayes Classifier had similar scores as the Logistic Regression for precision, recall, and F1 score; however, its ROC AUC was slightly lower (0.95) although its accuracy was slightly higher (0.92). Random Forest, another tree-based method, produced strong results; its precision, recall, and F1 score ranged from 0.94 to 0.95. It also showed a high ROC AUC of 0.97, comparable to that of Logistic Regression, with an overall accuracy of 0.91. Finally, XGBoost showed the best overall performance, matching Random Forest in terms of precision, recall, and F1 score. Its ROC AUC was also 0.97, but its accuracy of 0.94 was notable, being the highest among all classifiers. This suggests that XGBoost is not only effective in differentiating between the groups but is also the most reliable in making predictions across the dataset. In summary, each algorithm has its strengths, but XGBoost excels in both group differentiation and prediction accuracy.

Table 1
Mean, minimum, and maximum values, and *t*-test p-values of laboratory parameters based on presepsin classification (Negative, Positive).

routine laboratory parameters	presepsin Negative mean (min ~ max)	presepsin Positive mean (min ~ max)	p-value
age	46.20 (1–83)	68.49 (25–97)	3.40E-19
Hb	13.52 (20–298)	9.88 (5–16)	9.30E-28
WBC	6.05 (3.7–18.6)	10.50 (0.76–44.06)	2.19E-07
PLT	264.73 (72–802)	182.18 (8–588)	1.21E-08
neut%	56.51 (31.2–93.5)	77.76 (25–99.2)	9.42E-23
ANC	3610.63 (1377–17363)	8861.79 (0.3–42694)	4.11E-10
lymp%	33.17 (1.3–79.5)	12.34 (0.3–53.6)	4.92E-28
CRP	0.78 (0.02–19.32)	8.66 (0.03–46.54)	1.56E-19
Protein	7.20 (4.5–8.6)	5.9 (3.7–8.1)	1.15E-22
Albumin	4.58 (2.7–5.3)	3.13 (1.7–4.8)	2.03E-45
AST	21.80 (11–47)	72.71 (5–922)	3.36E-06
ALT	18.70 (5–60)	45.01 (5–472)	0.000499
BUN	14.24 (4–46.2)	29.95 (4.6–243.4)	2.70E-07
Cr	0.78 (0.3–2.24)	1.83 (0.34–11)	9.02E-08
CKD-EPI	99.09 (25.46–132.89)	63.31 (0.34–11)	7.07E-18

Table 2
Prediction (classification) results for presepsin groups (positive, negative) based on different artificial intelligence classifier algorithms without missing data.

Dataset (n = 173)	Algorithm	Precision	Recall	F1 Score	ROC AUC	Accuracy (Cross Validation, k = 5)
routine laboratory parameters	k-NN	0.77	0.77	0.77	0.81	0.80
~ presepsin	Logistic Regression	0.93	0.92	0.93	0.98	0.91
Classification	Naive Bayes Classifier	0.93	0.92	0.93	0.95	0.92
	Random Forest	0.95	0.94	0.94	0.97	0.91
	XGBoost	**0.95**	**0.94**	**0.94**	**0.97**	**0.94**

3.3. *Algorithm performance through confusion matrices and ROC curves*

To provide a complete overview of the algorithmic effectiveness, confusion matrices are compared in Fig. 3. The confusion matrices were generated using the independent test set of 53 samples (Supplementary Figure 1.), which had a similar composition to the full dataset in terms of presepsin levels and clinical diagnoses. This figure is organized as follows: Fig. 3. (a) shows the k-NN classifier, Fig. 3. (b) shows logistic regression and naïve Bayes, and Fig. 3. (c) shows the tree-based Random Forest and XGBoost models. These matrices show the number of true positives, true negatives, false positives, and false negatives using an independent test set of 53 samples, which were randomly selected from the original dataset of 173 samples (i.e., the dataset without missing values). This test set was not used during the model training process and served as a fair evaluation of each model's performance on unseen data. In the case of the k-NN classifier, 12 samples were correctly labeled as True Positives, 29 as True Negatives, 8 as False Negatives, and 4 as False Positives. Logistic regression and naïve Bayes, as depicted in the second matrix, achieved remarkable precision with only three False Positives and one False Negative among the 53 test samples. These findings reinforce the previously mentioned high precision and recall values of 0.93 and 0.92, respectively, along with an F1 score of 0.93. The third matrix details the performance of Random Forest and XGBoost. Each algorithm detects only three False Positives, accurately classifying the rest. Such consistent outcomes resonate with the earlier metrics of precision, recall, and F1 score, which fall between 0.94 and 0.95; notably, XGBoost achieved the highest accuracy (0.94). To supplement these matrices, Fig. 3. (d) displays the ROC curves for each algorithm. The k-NN classifier had an AUC-ROC of 0.81, indicating a reasonable performance. Logistic regression yielded an AUC-ROC of 0.98; naïve Bayes yielded a score of 0.95; and both Random Forest and XGBoost achieved a high AUC-ROC of 0.97. In conclusion, the confusion matrices and ROC curves shown in Fig. 3 effectively corroborates the performance metrics. In summary, among the algorithms tested, XGBoost demonstrated the highest capability for accurately differentiating between Positive and Negative presepsin groups.

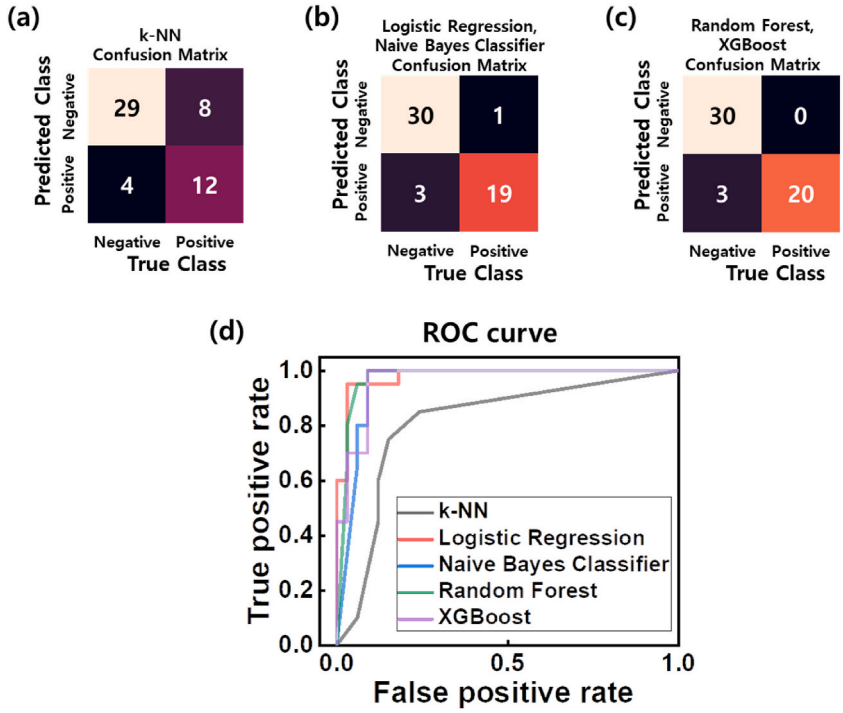


Fig. 3. Performance evaluation of classification algorithms on presepsin dataset. (a) Confusion matrix for the k-NN classifier, (b) Confusion matrices for Logistic Regression and Naive Bayes classifiers, (c) Confusion matrices for tree-based classifiers, Random Forest and XGBoost, and (d) Receiver Operating Characteristic (ROC) curves showcasing each algorithm's diagnostic ability at varying discrimination thresholds.

3.4. Expanded XGBoost analysis with incomplete data

In the extended analysis, a larger dataset of 216 samples was investigated, incorporating instances of missing data. All the performance metrics for this broader analysis are listed in Table 3. Despite the challenges of missing data, XGBoost demonstrates impressive performance stability. The algorithm scored consistently across multiple metrics, achieving precision, recall, and F1 score of 0.91. The ability of the algorithm to distinguish between Positive and Negative presepsin groups remained exceptional, with an ROC AUC score of 0.98. Although the analysis showed that the overall accuracy was slightly reduced to 0.86, it is important to maintain high precision, recall, and F1 scores given the general difficulty in processing missing data from medical datasets. The results confirmed the robustness and suitability of XGBoost for irregularities in data collection.

3.5. SHAP analysis for feature importance

To investigate the role of each variable in the model predictions, an analysis was conducted on the 173-sample dataset. As shown in Fig. 4. (a), the SHAP dot plot revealed the magnitude and direction of the impact of each parameter on the model output; particularly, Albumin, lymp%, Hb, and Cr have been confirmed to be the main influencing factors. Higher SHAP values correlated with lower Albumin, lymp%, and Hb values, indicating these features hold substantial influence on the model. In contrast, Cr exhibited an inverse relationship with higher values corresponding to higher SHAP values. As depicted in Fig. 4. (b), the feature importance was calculated based on the absolute SHAP values. Albumin and lymp% were identified as the most influential features, followed by Hb and Cr. This plot also includes additional information on 15 other variables, reinforcing the essential roles of Albumin and lymp% in the classification task. The SHAP analysis highlights significant variables and establishes a foundation for future investigations into the factors influencing presepsin levels.

4. Discussion

In this study, we employed XGBoost to handle missing data, as it can effectively manage missing values during the learning process. While other methods, such as data imputation, could have been used, XGBoost offers several advantages. First, it eliminates the need for a separate imputation step, which can introduce additional bias or noise into the data. Second, XGBoost's tree-based structure allows it to naturally handle missing values by considering them as a separate category during the split point selection. This enables the model to learn the best way to handle missing data based on the patterns in the available data. However, it is important to note that XGBoost's effectiveness in handling missing data may depend on the missingness pattern and the proportion of missing values in the dataset. Traditional statistical methods, such as correlation analyses, often provide limited insights into the complex relationships between presepsin and other the routine laboratory parameters. For example, Fig. 2 reveals that the highest correlation value for Cr is only 0.45, which is insufficient to draw clear clinical implications. Generally, a strong correlation is considered to have a coefficient value greater than 0.7, which makes it possible to employ more sophisticated analytical tools. More specifically, when examining the scatter plot of Cr and presepsin values, as depicted in Fig. 5 (a), a noticeable relationship is evident in only approximately 10 of the 216 data points. However, most of the data did not demonstrate a clear association. As observed in Fig. 5. (b), even upon closer examination of the presepsin values expanded to a detailed view up to approximately 2 k, no noticeable correlation could be found. In particular, the blue dashed line, representing a presepsin value of 333, revealed no clear correlation with the Cr values. Furthermore, as shown in Fig. 6, the histogram of the Cr values for the classification of sepsis as positive or negative, based on a presepsin value of 333, appears to be highly arbitrary, with a considerable amount of overlapping data. Consequently, even when examining Cr, which shows the relatively highest correlation in the routine laboratory parameters, no distinct relationship was identified through a basic correlation analysis of the data. Nevertheless, it was observed to function as an important feature of XGBoost with XAI analysis. Considering this, machine learning techniques, such as XGBoost, enhanced with SHAP for model explainability, offer a more detailed understanding of data. As observed in Fig. 4. (a), the distinct contributions of individual features to the target classification become apparent. Albumin, lymp%, and Hb are particularly influential, with lower values yielding positive SHAP values. This suggests a higher likelihood of participants being classified into the positive group with high presepsin levels. Conversely, Cr behaves differently; lower values were associated with negative SHAP values, indicating a preference for classification into a negative group with lower presepsin levels. The significance of these features is further confirmed in Fig. 4. (b), which presents the absolute values of the SHAP scores. Albumin, lymp%, and Hb emerge as the most crucial features, setting them apart as the primary drivers in this dataset. Another important aspect of the analysis is the robustness of XGBoost in handling missing data. This is a frequent issue in medical research data, which are often impossible to collect comprehensively for each patient. Despite the expansion of the dataset to 216 samples, including those with missing data, the model's accuracy showed only a minor reduction, and other key performance metrics such as the ROC AUC and F1 score remained stable, validating the robustness of the model. The findings from both traditional and machine learning-based analyses

Table 3
Prediction (classification) results for presepsin groups (positive, negative) using the XGBoost algorithm, with missing data.

Dataset (n = 216)	Algorithm	Precision	Recall	F1 Score	ROC AUC	Accuracy (Cross Validation, k = 5)
Routine laboratory parameters ~ presepsin Classification	XGBoost (with missing data)	0.91	0.91	0.91	0.98	0.86

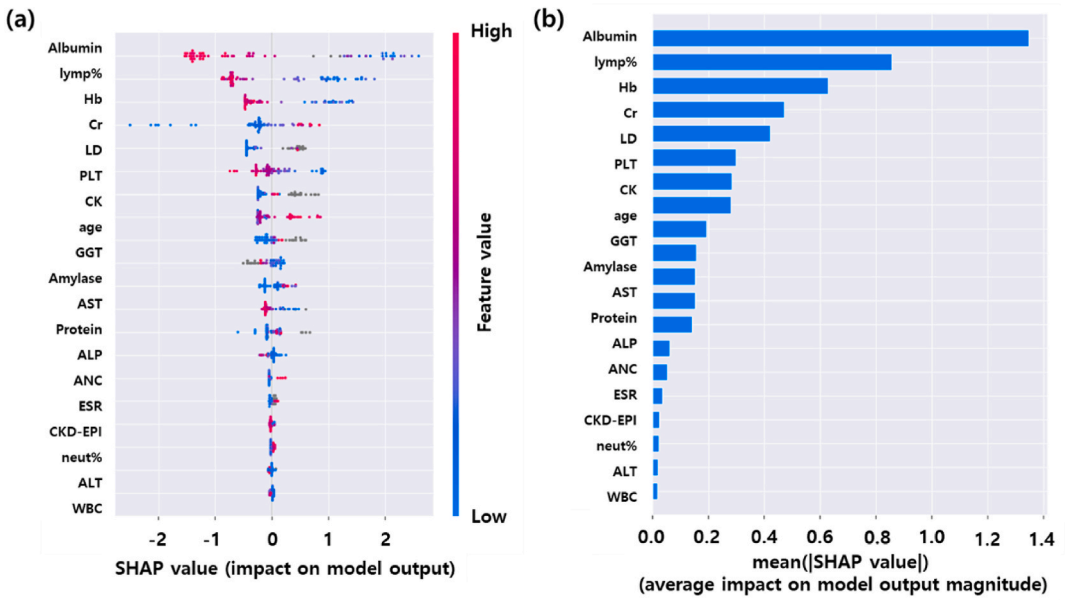


Fig. 4. SHAP analysis results for XGBoost classifier. (a) SHAP summary plot illustrating the magnitude and direction of each feature's impact on the model's output. (b) Feature importance plot, based on the absolute SHAP values, highlighting the most influential features.

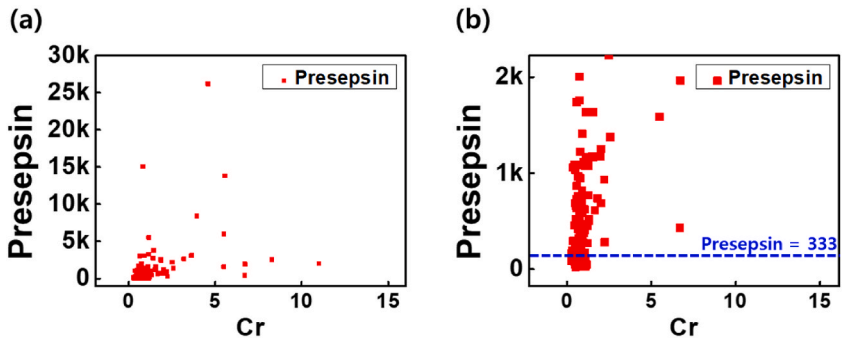


Fig. 5. Relationship between presepsin and Cr values. (a) Scatter plot between presepsin and Cr values, with a noticeable relationship in a small subset of data points. (b) An enlarged view of (a), focusing on up to approximately 2 k and featuring a blue dashed line to indicate the boundary at a presepsin value of 333. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

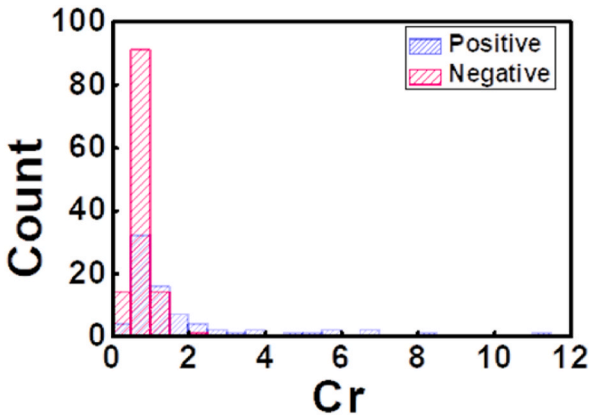


Fig. 6. Histogram of Cr values in relation to sepsis classification based on presepsin value.

indicate that while simple statistical methods can reveal some insights, the use of advanced techniques such as XGBoost and SHAP provides more detailed understanding. Low albumin levels linked to high presepsin levels are consistent with previous studies that have found significantly lower serum albumin levels in the sepsis group with high presepsin levels compared to the non-sepsis group [20,26]. This observation is further supported by two other studies [26,55]. Given that decreased serum albumin serves as a marker of malnutrition and is related to systemic inflammatory responses [55], its lower levels may be correlated with elevated presepsin levels. Although this study does not establish the albumin level as a predictor of presepsin levels, it suggests that the albumin level is a primary feature influencing the XGBoost model. Future research should focus on clearly defining the relationship between presepsin and albumin levels using a larger sample size. Regarding Cr, this study showed that low creatinine levels tended to be associated with low presepsin levels, which aligns with the literature demonstrating that higher presepsin concentrations are related to high creatinine levels [23,56,57]. Another study based on patients with pneumonia reported a moderate positive correlation between plasma presepsin and serum creatinine levels ($rs = 0.524$, $p < 0.001$) [58]. Applying advanced machine learning techniques to the analytical framework yielded a more comprehensive understanding of the relationship between presepsin levels and other parameters. Whereas, there are some limitations in this study. First, the data used in this analysis were collected from a single institution, which may limit the generalizability of the findings to other healthcare settings. To ensure robustness and reproducibility, future research should validate these results using data from multiple centers. Second, the study focused on a specific time period, and further investigation is needed to assess the temporal stability of the identified relationships between presepsin and routine laboratory parameters. Third, the retrospective nature of the study may introduce potential biases, such as selection bias, which could influence the results. Fourth, the feature importance ranking based on SHAP values (Fig. 4) evaluated the relative importance of variables in improving the model's performance. However, whether they form the optimal combination and if all variables are indispensable was not determined in this study. Nevertheless, the results of this study have significant implications for clinical practice. The identified relationships between presepsin and routine laboratory parameters could guide the development of personalized diagnostic and treatment strategies for sepsis. By integrating presepsin with readily available clinical data, clinicians may be able to make more informed decisions regarding patient management. Future prospective, large-scale and multi-institutional studies are necessary to validate the relationships between presepsin and routine laboratory parameters and to investigate the longitudinal dynamics of presepsin and its associated factors. Addtionally, those studies will have to include comparing model performance changes based on variable combinations and identifying essential variables by reducing the number of variables, assessing the generalizability of the machine learning models and developing user-friendly clinical decision support tools incorporating these algorithms.

5. Conclusion

This study explores the relationship between presepsin, a critical biomarker for sepsis diagnosis, and the routine laboratory parameters. Using advanced machine learning classifiers, this study offers detailed insights into the intricate connections between presepsin and other biomarkers. Furthermore, the analysis demonstrated the ability to analyze incomplete data owing to missing values within the dataset. The utilization of robust machine learning algorithms, particularly XGBoost, enables the acquisition of effective results from data analysis. In particular, the introduction of interpretable artificial intelligence (XAI) methods through SHAP values enhances the transparency and comprehensiveness of the research findings. Achieving a remarkable ROC AUC of 0.97 and an accuracy of 0.94, this study demonstrates the effectiveness of combining advanced machine learning techniques with XAI. This significant improvement in predictive metrics highlights the practical relevance of these approaches in enhancing sepsis diagnosis and treatment. This allows the extraction of high-dimensional feature importance that cannot be adequately captured through statistical analysis alone. The discrepancies between the key features identified in the basic statistical analysis and those highlighted by XAI emphasize the value of employing more advanced methods for medical data analysis. The application of AI enhances understanding beyond what basic statistical analysis can confirm, revealing hidden significance and introducing a novel approach to data interpretation and utilization. These findings offer valuable insights for enhancing the diagnosis and treatment of sepsis, emphasize the importance of integrating modern computational methods with traditional medical research, and steer the field toward new explorations. Future research should focus on validating these findings in larger, multi-center cohorts and developing user-friendly clinical decision support tools incorporating these algorithms.

Ethical statement

This study was approved by the Institutional Review Board of Korea University Ansan Hospital. (IRB No. 2022AS0167). Patient informed consent was waived for this study by IRB of Korea University Ansan Hospital. We have uploaded the Proof of Consent Waiver with our submission (Refer to the file "Proof of Ethics WAIER").

Funding statement

This study was supported and funded by the Korea University grant (grant number K2208441), the Korea University Ansan Hospital grant (grant number K2409121), the Korean National Police Agency (KNPA) (PR08-04-000-23-C5), the Ministry of Culture, Sports and Tourism (MCST), Korea Creative Content Agency (KOCCA) (CR202104002), and the Commercialization Promotion Agency for R&D Outcome (COMPA) (2022SCPO_B_0200).

Data availability statement

Data associated with the study has not been deposited into a publicly available repository. Data will be made available on request.

CRediT authorship contribution statement

Jae-Seung Jeong: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation. **Tak Ho Kang:** Writing – original draft, Validation, Formal analysis, Data curation. **Hyunsu Ju:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Chi-Hyun Cho:** Writing – original draft, Validation, Supervision, Project administration, Investigation, Data curation.

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.

Acknowledgement

Hyunsu Ju and Chi-Hyun Cho are co-corresponding authors. Jae-Seung Jeong and Tak Ho Kang contributed equally to this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e33826.

References

[1] J. Zhou, M. Du, S. Chang, Z. Chen, Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis, Cardiovasc. Ultrasound 19 (2021) 1–11.
[2] O. Elemento, C. Leslie, J. Lundin, G. Tourassi, Artificial intelligence in cancer research, diagnosis and therapy, Nat. Rev. Cancer 21 (2021) 747–752.
[3] C. Ao, S. Jin, H. Ding, Q. Zou, L. Yu, Application and development of artificial intelligence and intelligent disease diagnosis, Curr. Pharmaceut. Des. 26 (2020) 3069–3075.
[4] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. La Spada, M. Mirmozafari, M. Dehghani, Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment, IEEE Access 8 (2020) 109581–109595.
[5] A. Mitsala, C. Tsalikidis, M. Pitiakoudis, C. Simopoulos, A.K. Tsaroucha, Artificial intelligence in colorectal cancer screening, diagnosis and treatment. A new era, Curr. Oncol. 28 (2021) 1581–1607.
[6] J.C. Ahn, A. Connell, D.A. Simonetto, C. Hughes, V.H. Shah, Application of artificial intelligence for the diagnosis and treatment of liver diseases, Hepatology 73 (2021) 2546–2563.
[7] M. Rezaei, E. Rahmani, S.J. Khouzani, M. Rahmannia, E. Ghadirzadeh, P. Bashghareh, F. Chichagi, S.S. Fard, S. Esmaeili, R. Tavakoli, Role of artificial intelligence in the diagnosis and treatment of diseases, Kindle 3 (2023) 1–160.
[8] S.E. Dilsizian, E.L. Siegel, Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment, Curr. Cardiol. Rep. 16 (2014) 1–8.
[9] B. Ambale-Venkatesh, X. Yang, C.O. Wu, K. Liu, W.G. Hundley, R. McClelland, A.S. Gomes, A.R. Folsom, S. Shea, E. Guallar, Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis, Circ. Res. 121 (2017) 1092–1101.
[10] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, A. Nagler, Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT, Bone Marrow Transplant. 49 (2014) 332–337.
[11] Y.K. Kim, J.H. Koo, S.J. Lee, H.S. Song, M. Lee, Explainable artificial intelligence warning model using an ensemble approach for in-hospital cardiac arrest prediction: retrospective cohort study, J. Med. Internet Res. 25 (2023) e48244.
[12] F.H. Yagin, S. Yasar, Y. Gormez, B. Yagin, A. Pinar, A. Alkhateeb, L.P. Ardigò, Explainable artificial intelligence paves the way in precision diagnostics and biomarker discovery for the subclass of diabetic retinopathy in type 2 diabetics, Metabolites 13 (2023).
[13] F.H. Yagin, A. Alkhateeb, A. Raza, N.A. Samee, N.F. Mahmoud, C. Colak, B. Yagin, An explainable artificial intelligence model proposed for the prediction of myalgic encephalomyelitis/chronic fatigue syndrome and the identification of distinctive metabolites, Diagnostics 13 (2023).
[14] H.C. Lee, S.B. Yoon, S.M. Yang, W.H. Kim, H.G. Ryu, C.W. Jung, K.S. Suh, K.H. Lee, Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. Logistic regression model, J. Clin. Med. 7 (2018).
[15] M.Y. Memar, N. Alizadeh, M. Varshochi, H.S. Kafil, Immunologic biomarkers for diagnostic of early-onset neonatal sepsis, J. Matern. Fetal Neonatal Med. 32 (2019) 143–153.
[16] M.Y. Memar, H.B. Baghi, Presepsin: a promising biomarker for the detection of bacterial infections, Biomed. Pharmacother. 111 (2019) 649–656.
[17] A. Piccioni, M.C. Santoro, T. de Cunzo, G. Tullo, S. Cicchinelli, A. Saviano, F. Valletta, M.M. Pascale, M. Candelli, M. Covino, Presepsin as early marker of sepsis in emergency department: a narrative review, Medicina 57 (2021) 770.
[18] A. Piccioni, M.C. Santoro, T. de Cunzo, G. Tullo, S. Cicchinelli, A. Saviano, F. Valletta, M.M. Pascale, M. Candelli, M. Covino, F. Franceschi, Presepsin as early marker of sepsis in emergency department: a narrative review, Medicina (Kaunas) 57 (2021).
[19] Q. Zou, W. Wen, X.-c. Zhang, Presepsin as a novel sepsis biomarker, World J. Emergency Med. 5 (2014) 16.
[20] J. Wu, L. Hu, G. Zhang, F. Wu, T. He, Accuracy of presepsin in sepsis diagnosis: a systematic review and meta-analysis, PLoS One 10 (2015) e0133057.
[21] E. Galliera, L. Massaccesi, E. de Vecchi, G. Banfi, M.M.C. Romanelli, Clinical application of presepsin as diagnostic biomarker of infection: overview and updates, Clin. Chem. Lab. Med. 58 (2019) 11–17.
[22] M. Park, M. Hur, H. Kim, C.H. Lee, J.H. Lee, H.W. Kim, M. Nam, Prognostic utility of procalcitonin, presepsin, and the VACO index for predicting 30-day mortality in hospitalized COVID-19 patients, Annals of Laboratory Medicine 42 (2022) 406–414.
[23] E. Galliera, L. Massaccesi, E. de Vecchi, G. Banfi, M.M.C. Romanelli, Clinical application of presepsin as diagnostic biomarker of infection: overview and updates, Clin. Chem. Lab. Med. 58 (2019) 11–17.

[24] S. Lee, J. Song, D.W. Park, H. Seok, S. Ahn, J. Kim, J. Park, H.-j. Cho, S. Moon, Diagnostic and prognostic value of presepsin and procalcitonin in non-infectious organ failure, sepsis, and septic shock: a prospective observational study according to the Sepsis-3 definitions, BMC Infect. Dis. 22 (2022) 8.
[25] J. Wu, X. Zhan, S. Wang, X. Liao, L. Li, J. Luo, The value of plasma presepsin as a diagnostic and prognostic biomarker for sepsis in Southern China, Inflamm. Res. 72 (2023) 1829–1837.
[26] M. Kaplan, T. Duzenli, A. Tanoglu, B. Cakir Guney, Y. Onal Tastan, H.S. Bicer, Presepsin: albumin ratio and C-reactive protein: albumin ratio as novel sepsis-based prognostic scores: a retrospective study, Wien Klin. Wochenschr. 132 (2020) 182–187.
[27] E. Fix, J.L. Hodges, Discriminatory analysis. Nonparametric discrimination: consistency properties, International Statistical Review/Revue Internationale de Statistique 57 (1989) 238–247.
[28] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
[29] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[30] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (2016) 785–794.
[31] J.M. Brick, G. Kalton, Handling missing data in survey research, Stat. Methods Med. Res. 5 (1996) 215–238.
[32] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[33] M.S. Thiese, Z.C. Arnold, S.D. Walker, The misuse and abuse of statistics in biomedical research, Biochem. Med. 25 (2015) 5–11.
[34] T. Zhu, X. Liao, T. Feng, Q. Wu, J. Zhang, X. Cao, H. Li, Plasma monocyte chemoattractant protein 1 as a predictive marker for sepsis prognosis: a prospective cohort study, Tohoku J. Exp. Med. 241 (2017) 139–147.
[35] Y. Okamura, H. Yokoi, Development of a point-of-care assay system for measurement of presepsin (sCD14-ST), Clin. Chim. Acta 412 (2011) 2157–2161.
[36] K. Pearson, Note on regression and inheritance in the case of two parents, Proc. Roy. Soc. Lond. 58 (1895) 240–242.
[37] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Stat. Soc. B 57 (1995) 289–300.
[38] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. USA 100 (2003) 9440–9445.
[39] A. Reiner, D. Yekutieli, Y. Benjamini, Identifying differentially expressed genes using false discovery rate controlling procedures, Bioinformatics 19 (2003) 368–375.
[40] B. Deekshatulu, P. Chandra, Classification of heart disease using k-nearest neighbor and genetic algorithm, Procedia technology 10 (2013) 85–94.
[41] S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, Efficient kNN classification with different numbers of nearest neighbors, IEEE Transact. Neural Networks Learn. Syst. 29 (2017) 1774–1785.
[42] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inf. 35 (2002) 352–359.
[43] L. Jiang, H. Zhang, Z. Cai, J. Su, Learning tree augmented naive bayes for ranking. Database Systems for Advanced Applications: 10th International Conference, DASFAA 2005, Beijing, China, April 17-20, 2005. Proceedings 10, Springer, 2005, pp. 688–698.
[44] H. Zhang, The optimality of naive Bayes, Aa 1 (2004) 3.
[45] A. Lebedev, E. Westman, G. Van Westen, M. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, Neuroimage: Clinical 6 (2014) 115–125.
[46] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Statistician 46 (1992) 175–185.
[47] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1967) 21–27.
[48] D.W. Hosmer, Jr, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley, 2013. Sons.
[49] S. Menard, Applied Logistic Regression Analysis, 2002. Sage.
[50] D.J. Hand, K. Yu, Idiot's Bayes—not so stupid after all? Int. Stat. Rev. 69 (2001) 385–398.
[51] I. Rish, An empirical study of the naive Bayes classifier, in: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Citeseer, 2001, pp. 41–46.
[52] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (2002) 18–22.
[53] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, Xgboost: extreme gradient boosting, R package version 0 1 (4–2) (2015) 1–4.
[54] H. Akoglu, User's guide to correlation coefficients, Turkish journal of emergency medicine 18 (2018) 91–93.
[55] M. Tambo, S. Taguchi, Y. Nakamura, T. Okegawa, H. Fukuhara, Presepsin and procalcitonin as predictors of sepsis based on the new Sepsis-3 definitions in obstructive acute pyelonephritis, BMC Urol. 20 (2020) 1–7.
[56] M. Tambo, S. Taguchi, Y. Nakamura, T. Okegawa, H. Fukuhara, Presepsin and procalcitonin as predictors of sepsis based on the new Sepsis-3 definitions in obstructive acute pyelonephritis, BMC Urol. 20 (2020) 23.
[57] T. Nagata, Y. Yasuda, M. Ando, T. Abe, T. Katsuno, S. Kato, N. Tsuboi, S. Matsuo, S. Maruyama, Clinical impact of kidney function on presepsin levels, PLoS One 10 (2015) e0129159.
[58] M. Ugajin, Y. Matsuura, K. Matsuura, H. Matsuura, Impact of initial plasma presepsin level for clinical outcome in hospitalized patients with pneumonia, J. Thorac. Dis. 11 (2019) 1387–1396.