



Original Article



Immune Cell Communication Networks and Machine Learning-based Diagnostic Signatures in Sepsis: Insights from Single-cell RNA Sequencing and Cross-dataset Validation

Yu-Long Wang^{1#}, Qing Su^{2#}, Ming-Gao Zhu¹, Man Li¹, Feng-Zhi Zhao¹, Hai-Yan Yin^{1*}  and Wan-Jie Gu^{1*} 

¹Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China; ²Department of Emergency Medicine, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China

Received: December 17, 2025 | Revised: February 24, 2026 | Accepted: March 19, 2026 | Published online: June 29, 2026

Abstract

Background and objectives: Sepsis is a life-threatening syndrome associated with high morbidity and mortality, underscoring the urgent need for early diagnostic biomarkers and therapeutic targets. However, current diagnostic strategies remain insufficiently precise because of the complex immune dysregulation and immune microenvironment heterogeneity that characterize sepsis. This study aimed to identify reliable diagnostic biomarkers for sepsis and explore their immune regulatory mechanisms together with potential therapeutic relevance using multidimensional bioinformatic analyses.

Methods: Single-cell transcriptomic and bulk RNA sequencing datasets were integrated to screen candidate diagnostic genes for sepsis. Immune infiltration, co-expression network and pathway enrichment analyses were performed to explore immune regulatory mechanisms. Machine-learning approaches were used to validate the diagnostic signature, and molecular docking was conducted to predict candidate targeted compounds.

Results: A total of 346 differentially expressed genes were identified and were mainly enriched in immune, coagulation, and metabolic pathways. CIBERSORT and single-cell analyses revealed increased neutrophils, monocytes, and $\gamma\delta$ T cells and reduced CD8⁺ T cells and resting natural killer cells. Four diagnostic genes (*S100A12*, *CD22*, *CSTA*, and *UPP1*) were prioritized. The four-gene model showed robust external performance (area under the receiver operating characteristic curve = 0.860; sensitivity = 0.781; specificity = 0.780), and interpretability analysis highlighted *UPP1* and *S100A12* as dominant predictors. Molecular docking suggested potential interactions between these targets and anti-inflammatory compounds.

Conclusions: This integrative framework identifies four immune-related diagnostic genes for sepsis and links them to immune-cell remodeling and candidate therapeutic interactions, providing a basis for future mechanistic and clinical validation.

Introduction

Sepsis, defined as life-threatening organ dysfunction caused by a

dysregulated host response to infection, is a leading cause of global mortality, with 49 million annual cases and 11 million deaths. Its progression to multiple organ dysfunction syndrome substantially increases mortality and imposes a heavy burden on healthcare systems.¹ Despite advances in critical care, early diagnosis and effective therapies remain limited because of the disease's complex pathogenesis, which is characterized by dysregulated immune responses, including initial hyperinflammation ("cytokine storm") followed by immunosuppression and increased susceptibility to secondary infections.^{2,3}

Recent advances in bioinformatics offer powerful tools for elucidating sepsis mechanisms. Transcriptome profiling enables biomarker identification, patient stratification, and therapeutic target discovery.^{4,5} Previous studies have mostly constructed sepsis diagnostic models based on bulk transcriptomic data, but they have

Keywords: Sepsis; Biomarkers; Immune infiltration; Single-cell sequencing; Machine learning; Diagnostic model; SHAP interpretability analysis; Molecular docking.

***Correspondence to:** Wan-Jie Gu and Hai-Yan Yin, Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University, 613 Huangpu Avenue West, Guangzhou, Guangdong 510630, China. ORCID: <https://orcid.org/0000-0003-4923-8282> (WJG); <https://orcid.org/0009-0008-0176-5204> (HYY). Tel: +86-15850546835 (WJG); +86-13318831222 (HYY), E-mail: guwanjie@jnu.edu.cn (WJG); haiyanyin1867@126.com (HYY)

[#]These authors contributed equally to this work.

How to cite this article: Wang Y-L, Su Q, Zhu M-G, Li M, Zhao F-Z, Yin H-Y, et al. Immune Cell Communication Networks and Machine Learning-based Diagnostic Signatures in Sepsis: Insights from Single-cell RNA Sequencing and Cross-dataset Validation. *J Transl Crit Care Med* 2026;8(2):e00027. doi: 10.14218/JTCCM.2025.00027.

often lacked analyses of immune microenvironment heterogeneity and the cellular sources of marker genes. In this study, immune infiltration analysis, weighted gene co-expression network analysis (WGCNA), and single-cell data analysis were used to characterize the immune environment of patients with sepsis and explore immune microenvironment heterogeneity, providing new insights into the diagnosis and treatment of this condition.

This study aimed to systematically identify sepsis-related biomarkers by integrating multiple datasets, combining multiple bioinformatics methods to construct an efficient sepsis diagnostic model, and enhancing the translational potential of diagnostic markers through single-cell validation and interpretable analysis. In addition, functional analysis of the selected diagnostic genes may contribute to a more comprehensive understanding of the mechanisms involved in sepsis development and progression and provide a scientific basis for identifying potential drug targets. These findings provide new insights into the early diagnosis and precision treatment of sepsis. This work is not only a diagnostic model study but also a comprehensive investigation of immune dysregulation in sepsis that may support precision diagnosis and treatment.

Materials and methods

Data acquisition and processing

The overall design of this study is shown in [Supplementary Figure 1](#). We retrieved whole-blood transcriptomic data from patients with sepsis and healthy controls from the Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>).⁶ Datasets were eligible if they included whole-blood transcriptomic data, sepsis patients and healthy controls, a sample size ≥ 20 , and an Affymetrix or Illumina platform. Datasets were excluded if they included pediatric patients, lacked RNA measurements or an original expression matrix, had severe missing clinical information, or used non-human samples. Five datasets from 2011 to 2021 met the criteria: GSE28750, GSE69063, GSE95233, and GSE154918 were used to construct the diagnostic model, and GSE65682 was used for validation. The GEOquery package in R (version 4.3.3) was used to convert probe data into gene expression matrices. The data were normalized, corrected, and merged using the limma package in R. The ComBat function in the sva package was used to remove potential batch effects. Principal component analysis (PCA) was used to evaluate batch effects.

Identification of differentially expressed genes (DEGs)

Differential expression analysis between the healthy control and sepsis groups was performed using the limma package,⁷ with selection criteria of $|\log_2 \text{fold change}| > 1$ and an adjusted P value < 0.05 . Results were visualized using the ggplot and pheatmap packages. This study followed the Transparent Reporting of a Multi-variable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement ([Supplementary Table 1](#)).

Pathway enrichment analysis of differentially expressed genes

Gene Ontology (GO) enrichment analysis is commonly used in bioinformatics to assess the enrichment of specific gene sets in biological processes, molecular functions, and cellular components.⁸ We used the clusterProfiler package in R to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and GO functional annotation of sepsis-related DEGs.⁹ To control false positives caused by multiple hypothesis testing, Benjamini-Hochberg false discovery rate (FDR) correction was applied. Significantly enriched gene sets were defined as those

with both an unadjusted $P < 0.05$ and an FDR-adjusted q value < 0.05 . To further explore the functional characteristics and potential biological significance of the DEGs, gene set enrichment analysis (GSEA) was conducted using the c5.go.v7.4.symbols and c2.kegg.v7.4.symbols gene sets.¹⁰

Immune cell infiltration analysis

CIBERSORT (LM22 signature) was used to estimate immune-cell abundance in sepsis patients and controls.¹¹ PCA differentiated the groups based on immune profiles. Pearson correlation analysis with Benjamini-Hochberg FDR correction (q value < 0.05) was conducted to quantify pairwise correlations of immune-cell subsets between the sepsis and control groups, with statistical significance defined as $P < 0.05$. Significant immune-cell correlation networks ($|r| > 0.3$, $P < 0.05$) were constructed using the igraph package, where nodes represented immune-cell types and edges denoted significant correlations (edge width proportional to correlation strength; edge color indicating positive or negative correlation). Hierarchical clustering (Ward.D2 method) and dynamic tree cutting (minimum cluster size = 2) were applied to identify functional immune-cell modules. Module consistency was evaluated by mean intramodule correlation, and functional associations were annotated based on established immune-cell functions. Differential correlation analysis was further performed to compare immune-cell interaction patterns between the sepsis and control groups, identify sepsis-specific and control-specific immune-cell pairs, and calculate differential correlation coefficients (sepsis r - control r) to characterize remodeling of immune-cell interaction networks.

Finally, immune-related functions were compared between the high- and low-expression groups of target genes, and the results were visualized using box plots. Correlation analyses were performed to identify immune cells associated with target genes, and bubble plots were generated to visualize these associations, facilitating characterization of the immune cell composition in patients with sepsis.

Markov cluster algorithm (MCL), protein-protein interaction (PPI) network construction, and Friends analysis

MCL was used to determine which pathways were enriched for differentially expressed genes. MCL relies on the STRING database (<https://string-db.org/>) for online analysis of protein interactions, and Cytoscape software (version 3.10.2) was used to visualize the protein-interaction network.^{12,13} Key genes were extracted from the PPI network and further analyzed through Friends analysis, which evaluates gene-gene functional similarity based on GO semantic metrics.

WGCNA

WGCNA was used to construct a weighted gene co-expression network and analyze correlations among gene expression patterns.¹⁴ We performed hierarchical clustering based on weighted correlations to identify gene modules associated with immune cell infiltration in sepsis and analyzed their potential roles in the immune landscape.

Single-cell RNA sequencing data processing and analysis

Single-cell data analysis (data from GSE217906) was performed in two stages. First, single-cell data from the sepsis group (GSM8217323, GSM8217324, and GSM8217325) were analyzed. Second, single-cell data from the sepsis group (GSM6729711 and GSM6729712) and healthy control group (GSM6729713, GSM6729714, and GSM6729715) were compared. Data were pre-

processed using Seurat: low-quality cells (genes <50 or mitochondrial content >15%) were filtered, followed by log-normalization and selection of 1,500 highly variable genes. The Harmony package was used to correct batch effects, and Louvain clustering (resolution = 0.6) with PCA was used to identify cell clusters, which were visualized using t-SNE. Cell types were annotated using the SingleR package against seven reference datasets (BlueprintEncode, HumanPrimaryCellAtlas, etc.). Marker genes ($|\log_2$ fold change| >1, adjusted $P < 0.05$, Wilcoxon test) were identified and visualized using heatmaps. Cell trajectories were inferred using the monocle package with dimension reduction by DDRTree.¹⁵

Cell-cell communication analysis

Cell-cell communication analysis was performed using the CellChat package. The human ligand-receptor database (CellChatDB.human) was filtered to retain biologically relevant interactions.¹⁶ Overexpressed genes and ligand-receptor pairs were identified, and communication probabilities were then computed based on ligand-receptor co-expression patterns; interactions involving fewer than 10 cells were excluded to minimize noise. Pathway-level networks were integrated with protein interaction data. The results were visualized using circular plots (interaction counts/weights) and ligand-receptor bubble charts. Key pathways and receptor-ligand pairs linked to sepsis progression were identified using heatmaps and cell-type analysis plots.

Supervised machine learning and diagnostic model construction

Five independent feature selection methods were used to screen candidate biomarkers. Machine learning methods, namely Lasso regression,¹⁷ Support Vector Machine–Recursive Feature Elimination (SVM-RFE),¹⁸ random forest,¹⁹ eXtreme Gradient Boosting (XGBoost),²⁰ and Gradient Boosting Machine (GBM), were employed to construct a diagnostic model for sepsis.²¹ The selection of these five algorithms was based on their prior applications in bioinformatics and sepsis biomarker screening. The algorithms are complementary: Lasso is suitable for feature selection in high-dimensional data; SVM-RFE is suitable for transcriptomic datasets; random forest is resistant to overfitting; XGBoost and GBM improve model accuracy through gradient boosting; and all have been used in previous studies of sepsis biomarkers.^{22–24} Model performance was assessed using receiver operating characteristic (ROC) curves, including the area under the ROC curve (AUC), sensitivity, specificity, positive predictive value, and negative predictive value. Combining these methods was intended to improve the efficiency and accuracy of the model.

Diagnostic model performance evaluation and validation

To assess the accuracy of the constructed diagnostic model, we validated it using the GSE65682 validation set. The diagnostic performance of the model across samples was further evaluated using violin plots and ROC curves.

Gene set expression variation analysis

Using gene set variation analysis (GSVA), we evaluated functional differences and pathway changes between the high- and low-expression groups and further analyzed the biological impact of diagnostic genes at different expression levels.²⁵

Nomogram, decision curve analysis (DCA), and clinical impact curve (CIC)

We used the rms package in R to construct a nomogram,²⁶ which was combined with decision curve analysis (DCA) and a clinical

impact curve (CIC) to assess the clinical predictive value of the model and further validate its feasibility and effectiveness in clinical practice.

Shapley additive exPlanations (SHAP)-based interpretable machine learning analysis

This study employed the SHAP framework to interpret gene expression-based classification models.²⁷ A standardized gene expression matrix was processed to extract the expression profiles of four diagnostic genes (*S100A12*, *CD22*, *CSTA*, and *UPP1*), followed by matrix transposition and group labeling to construct a sample-feature dataset. The dataset was stratified into training and test sets (7:3 ratio) to ensure class balance. Using the caret package, multiple machine-learning algorithms (including random forest, support vector machine, XGBoost, and 10 additional algorithms) were trained and evaluated via 5-fold repeated cross-validation, with the optimal model selected based on the AUC. SHAP values were calculated using a permutation-based method (permshap) to quantify gene contributions, and visualizations (bar plot, bee plot, waterfall plot, and force plot) were generated using the shapviz package.

Molecular docking and targeted drug screening

Drug-selection criteria included established anti-inflammatory drugs that have been clinically used for sepsis or inflammation (e.g., dexamethasone and aspirin), drugs in the Comparative Toxicogenomics Database (<http://ctdbase.org/>) with known interactions with the four diagnostic genes, and expression characteristics specific to patients with sepsis. AutoDock software (version 1.5.7) was used for molecular docking analysis,²⁸ and results were visualized using PyMOL software.²⁹ This analysis aimed to prioritize potential candidate drugs for further experimental validation in sepsis.

Statistical analysis

All data processing and analyses were conducted in R version 4.3.3. For normally distributed continuous variables, independent-samples t-tests were used for group comparisons. For non-normally distributed variables, Mann-Whitney U tests (Wilcoxon rank-sum tests) were used. ROC curves for predicting binary classification variables were plotted using the pROC package. All statistical tests were two-sided, with $P < 0.05$ considered statistically significant.

Results

DEGs screening and biological function

Four datasets (122 sepsis cases and 116 controls) were analyzed. The basic information on these datasets is provided in [Supplementary Table 2](#). After PCA processing, samples from different experimental batches were randomly distributed. Box plots showed that the median and distribution ranges of each batch were more comparable after processing. The F values of genes generally decreased after processing, indicating that the variance explained by batch factors was reduced. Finally, differences in sample correlations within and between batches narrowed, suggesting that batch-specific associations were weakened. These results indicate that the processing method effectively eliminated batch effects ([Supplementary Fig. 2a–d](#)). A total of 346 DEGs were identified (230 upregulated and 116 downregulated; [Fig. 1a and b](#)). GO and KEGG analyses suggested that sepsis is associated with inflammatory responses to microbial pathogens, with key pathways including immune receptor activity, cytokine binding, T-cell differen-

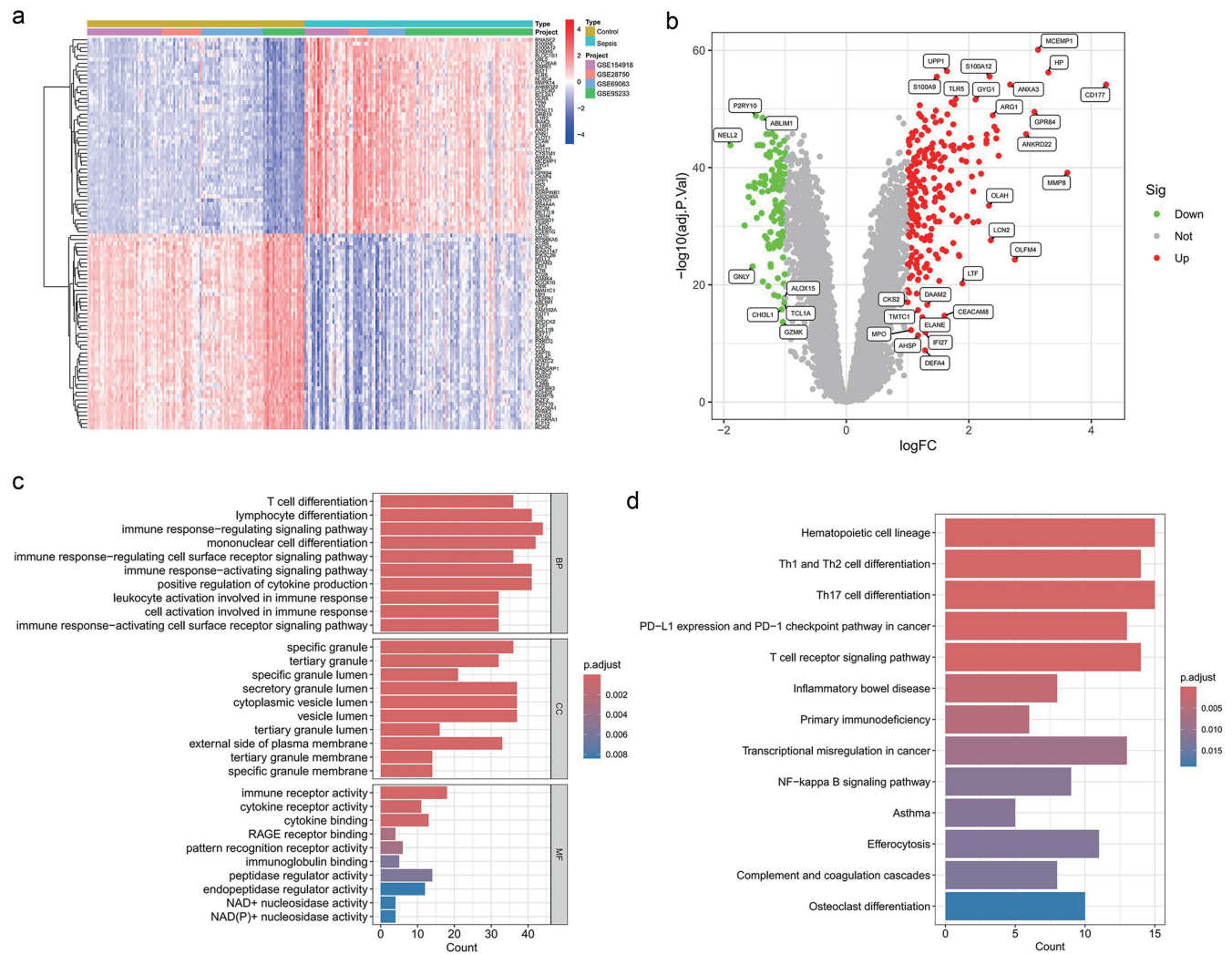


Fig. 1. Differential expression analysis and functional enrichment. (a, b) Heatmap and volcano plot showing differentially expressed genes (DEGs). (c, d) Gene Ontology (GO) enrichment analysis of the intersecting genes, showing the top 10 terms in biological process (BP), cellular component (CC), and molecular function (MF), together with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.

tiation, and immune-response regulation via cell-surface receptor signaling (Tables 1 and 2; Fig. 1c and d). These results suggest that DEGs are predominantly enriched in immune-related pathways. Notably, KEGG analysis highlighted the key role of T cells in sepsis, particularly in T-cell receptor signaling and T-helper cell differentiation. T cells are crucial immune cells involved in immune regulation, reflecting the association between immune dysregulation and sepsis.³⁰ Furthermore, GSEA revealed significant associations between sepsis and pathways related to coagulation, biochemical reactions, and autoimmune responses (Supplementary Fig. 2c–f).

Immune cell infiltration analysis

CIBERSORT analysis revealed immune-cell imbalances in sepsis: increased neutrophils, monocytes, M0 macrophages, and $\gamma\delta$ T cells, accompanied by reduced resting CD4⁺ T cells, CD8⁺ T cells, and natural killer (NK) cells (Fig. 2a and b). To explore immune-cell crosstalk, we analyzed correlation networks of immune subsets in sepsis patients and healthy controls. In the sepsis

group, strong positive correlations were observed among T-cell subsets (e.g., CD8⁺ T cells with follicular helper T cells, $r = 0.765$, $P < 0.001$; follicular helper T cells with $\gamma\delta$ T cells, $r = 0.688$, $P < 0.001$), whereas neutrophils exhibited significant negative correlations with multiple immune-cell subsets (e.g., CD8⁺ T cells, $r = -0.594$, $P < 0.001$; resting NK cells, $r = -0.430$, $P < 0.001$). In addition, pro-inflammatory M1 macrophages correlated positively with M2 macrophages ($r = 0.388$, $P < 0.001$) and activated dendritic cells ($r = 0.370$, $P < 0.001$). In healthy controls, distinct interaction patterns were observed, including a strong positive correlation between M1 macrophages and resting dendritic cells ($r = 0.976$, $P < 0.001$) and negative correlations of CD8⁺ T cells ($r = -0.650$, $P < 0.001$) and resting NK cells ($r = -0.615$, $P < 0.001$) with neutrophils (Fig. 2c, Supplementary Fig. 3a and b). These results indicate that sepsis induces marked remodeling of immune-cell networks, characterized by enhanced activation of T-cell subsets and disrupted crosstalk between neutrophils and other immune populations, which differs from the balanced immune interactions observed in healthy individuals.

Table 1. GO enrichment analysis of differentially expressed genes

Ontology	ID	Description	Adjusted <i>P</i>	q value
BP	GO:0030217	T cell differentiation	4.38E-16	3.69E-16
BP	GO:0030098	lymphocyte differentiation	8.42E-16	7.08E-16
BP	GO:0002764	immune response-regulating signaling pathway	8.42E-16	7.08E-16
BP	GO:1903131	mononuclear cell differentiation	4.57E-15	3.84E-15
BP	GO:0002768	immune response-regulating cell surface receptor signaling pathway	4.57E-15	3.84E-15
CC	GO:0042581	specific granule	7.47E-28	6.58E-28
CC	GO:0070820	tertiary granule	8.35E-23	7.35E-23
CC	GO:0035580	specific granule lumen	2.91E-20	2.56E-20
CC	GO:0034774	secretory granule lumen	1.62E-18	1.43E-18
CC	GO:0060205	cytoplasmic vesicle lumen	1.66E-18	1.46E-18
MF	GO:0140375	immune receptor activity	9.32E-08	8.07E-08
MF	GO:0004896	cytokine receptor activity	0.00035	0.00031
MF	GO:0019955	cytokine binding	0.00043	0.00037
MF	GO:0050786	RAGE receptor binding	0.00290	0.00251
MF	GO:0038187	pattern recognition receptor activity	0.00333	0.00288

BP, biological process; CC, cellular component; GO, Gene Ontology; MF, molecular function; adjusted *P*, *P* value adjusted using the Benjamini-Hochberg method; q value, false discovery rate-adjusted q value.

Module clustering analysis revealed distinct immune-cell functional modules between sepsis patients and healthy controls. In the sepsis group, key modules included the turquoise module (naive B cells, naive CD4⁺ T cells, regulatory T cells [Tregs], and M0 macrophages), the blue module (CD8⁺ T cells, follicular helper T cells, and $\gamma\delta$ T cells), the brown module (activated CD4⁺ memory T cells, resting NK cells, and neutrophils), and the yellow module (M1/M2 macrophages and activated dendritic cells). In healthy controls, module composition differed substantially. For example, the turquoise module contained activated CD4⁺ memory T cells,

Tregs, monocytes, and M2 macrophages, whereas neutrophils clustered with CD8⁺ T cells and resting NK cells in the yellow module (Supplementary Fig. 3c and d). These divergent module patterns indicate sepsis-induced remodeling of immune functional clusters and disrupted coordination of immune-cell subsets compared with the balanced modular organization in healthy individuals.

Differential correlation analysis uncovered distinct immune cell interaction patterns between sepsis patients and healthy controls. Sepsis-specific interactions included positive correlations among T-cell subsets (e.g., CD8⁺ T cells with activated CD4⁺ memory

Table 2. KEGG enrichment analysis of differentially expressed genes

Term	ID	Description	Adjusted <i>P</i>	q value
KEGG	hsa04640	Hematopoietic cell lineage	6.33E-07	3.37E-07
KEGG	hsa04658	Th1 and Th2 cell differentiation	7.24E-07	3.85E-07
KEGG	hsa04659	Th17 cell differentiation	7.24E-07	3.85E-07
KEGG	hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	3.44E-06	1.74E-06
KEGG	hsa04660	T cell receptor signaling pathway	1.60E-05	7.94E-06
KEGG	hsa05321	Inflammatory bowel disease	0.00299	0.00214
KEGG	hsa05340	Primary immunodeficiency	0.00522	0.00362
KEGG	hsa05202	Transcriptional misregulation in cancer	0.00831	0.00693
KEGG	hsa04064	NF-kappa B signaling pathway	0.01178	0.00744
KEGG	hsa05310	Asthma	0.01209	0.00851
KEGG	hsa04148	Efferocytosis	0.01227	0.00854
KEGG	hsa04610	Complement and coagulation cascades	0.01249	0.00854
KEGG	hsa04380	Osteoclast differentiation	0.01875	0.01229

KEGG, Kyoto Encyclopedia of Genes and Genomes; PD-1, programmed cell death protein 1; PD-L1, programmed death-ligand 1; Th1, T helper 1 cell; adjusted *P*, *P* value adjusted using the Benjamini-Hochberg method; q value, false discovery rate-adjusted q value.

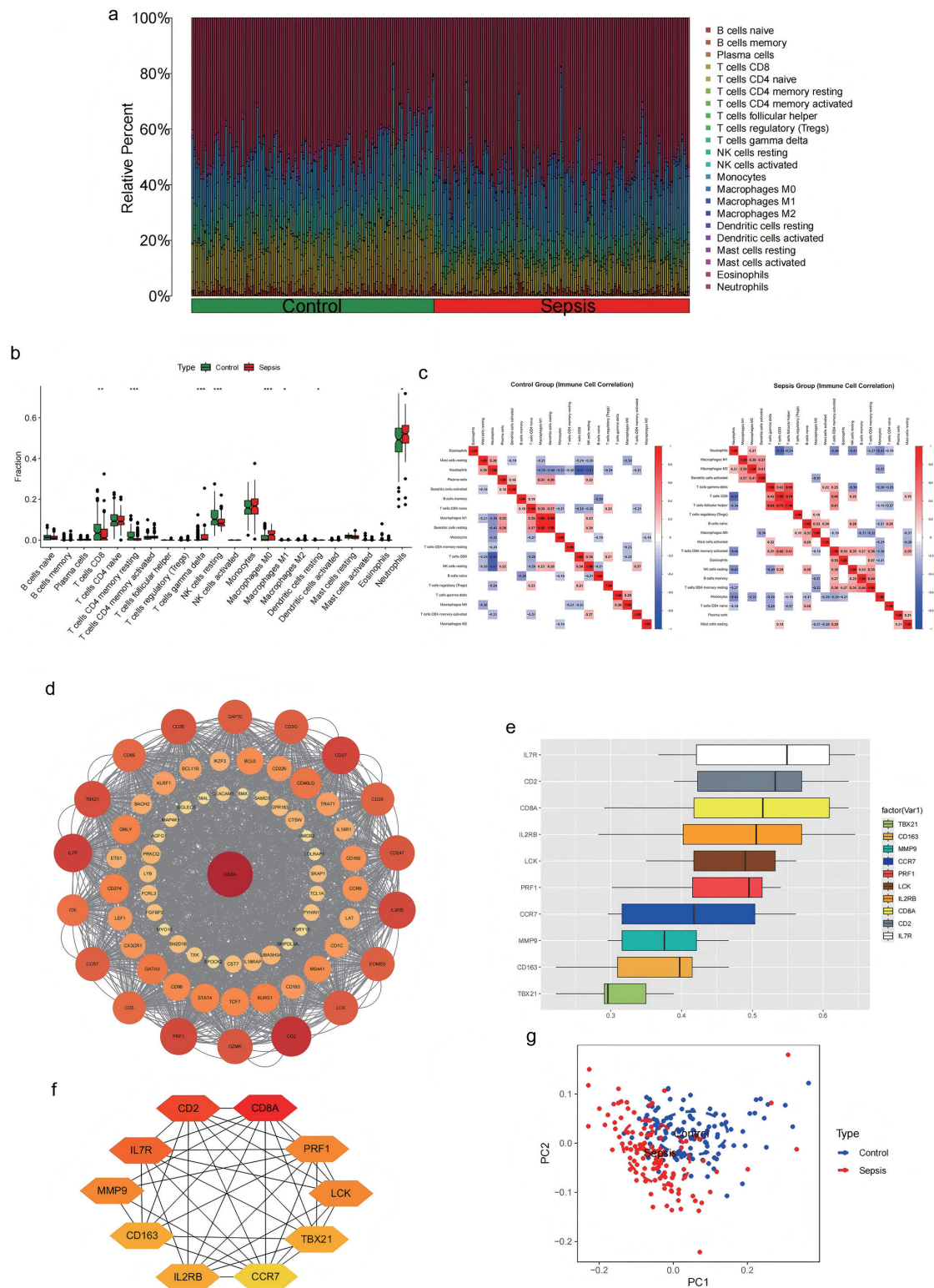


Fig. 2. Immune infiltration analysis in sepsis patients. (a) Bar plot showing the proportions of immune-cell infiltration in sepsis and healthy control samples. (b) Differential analysis of immune cells between sepsis and healthy control groups. (c) Immune-cell correlation plot comparing healthy control and sepsis groups. (d) Concentric circle plot illustrating interactions among 72 proteins enriched in the T-cell differentiation/adaptive immune pathway. (e) Top 10 hub genes identified through Friends analysis. (f) Protein-protein interaction (PPI) network showing interactions among the top 10 genes. (g) Principal component analysis (PCA) of immune-cell composition in the sepsis and healthy control groups.

T cells/ $\gamma\delta$ T cells) and between M1/M2 macrophages, whereas healthy controls exhibited unique correlations, such as negative correlations between naive B cells and memory B cells and between M1 macrophages and neutrophils. Additionally, several immune cell pairs showed significant interactions only in sepsis, including activated CD4⁺ memory T cells with resting natural killer (NK) cells/eosinophils and M0 macrophages with activated dendritic cells (negative) (Supplementary Fig. 3e; Supplementary Table 3). These divergent interaction profiles, together with altered modular clustering of immune cells, highlight sepsis-induced remodeling of the immune regulatory network, characterized by emergent pro-inflammatory cell crosstalk and loss of homeostatic immune interactions.

MCL clustering analysis, PPI network, and Friends analysis

MCL clustering identified 72 protein clusters in T-cell differentiation pathways (Supplementary Table 4). STRING/Cytoscape-based protein-protein interaction (PPI) network analysis highlighted CD8A as the top hub gene (Fig. 2d and f), linking sepsis to adaptive immunity. Friends analysis was then performed (Fig. 2e). Principal component analysis (PCA) further confirmed immune microenvironment heterogeneity (Fig. 2g).

WGCNA co-expression network and gene module identification

A sepsis-related co-expression network was constructed using WGCNA. In total, 4,024 sepsis-associated genes were analyzed in WGCNA, resulting in a scale-free network with a soft-thresholding power of $R^2 = 0.9$. The key soft-thresholding parameter was set to 6 (Fig. 3a), and 13 modules were identified using dynamic tree cutting (Fig. 3b). Sepsis-related genes were then mapped onto these modules (Fig. 3c), with significant enrichment in the brown, blue, turquoise, and yellow modules. We prioritized these four modules based on two core lines of evidence: first, our prior analysis identified eight differentially expressed immune cells between sepsis patients and healthy controls (CD8⁺ T cells, CD4⁺ memory resting T cells, $\gamma\delta$ T cells, resting NK cells, M0/M1 macrophages, resting dendritic cells, and neutrophils; Fig. 2b); second, module-trait correlation analysis demonstrated that these modules were significantly correlated with the above sepsis-related immune cells (e.g., the brown module was positively correlated with resting NK cells [correlation coefficient = 0.42] and negatively correlated with neutrophils [correlation coefficient = -0.59]; the turquoise module was positively correlated with M0 macrophages [correlation coefficient = 0.32] and neutrophils [correlation coefficient = 0.5]; Fig. 3c). In contrast, the remaining nine modules showed no significant associations with sepsis-related immune dysregulation and lacked enrichment of sepsis-associated DEGs; therefore, they were not included in downstream analysis. We further analyzed the interactions between these four modules and differential immune cells (Fig. 3d–k), with module-gene interaction details provided in Supplementary Table 5. Collectively, these findings highlight the role of immune dysregulation in sepsis pathogenesis.

Machine learning for key diagnostic gene identification

Supervised machine learning methods, including Lasso, SVM-RFE, random forest, XGBoost, and GBM, were applied to identify key diagnostic genes for sepsis and construct diagnostic models (Fig. 4a–f). Based on feature importance, Lasso selected 22 genes, SVM-RFE selected 37 genes, random forest selected 26 genes, XGBoost selected 31 genes, and GBM selected 62 genes. Detailed information on the basic parameter settings, hyperparameter tuning strategies, optimal parameters, feature selection criteria, and

cross-validation strategies for the five machine learning methods is presented in Supplementary Table 6. Diagnostic models were then constructed using the genes selected by all five methods (S100A12, UPP1, CD22, and CSTA). The expression levels of the selected features are shown in Supplementary Figure 4.

Diagnostic model performance and predictive ability of selected genes

To compare the performance of each feature-selection method, classifier performance was evaluated for each model using the validation dataset (Supplementary Table 7). The XGBoost and GBM models achieved high AUC, sensitivity, and specificity, whereas the SVM-RFE model showed the lowest AUC (0.813) and low specificity (Fig. 4g–k). Because multivariable methods can select features with varying accuracy, we employed an ensemble learning algorithm using the DEGs selected by each method. The ensemble model had an AUC of 0.835, sensitivity of 0.988, and specificity of 0.685, outperforming the SVM-RFE model (Fig. 4l). We also focused on overlapping genes selected by all five feature-selection methods and evaluated their performance in sepsis diagnosis. Among the four genes in the training set, *UPP1* performed best, with the highest AUC (0.990), whereas in the validation set, *S100A12* showed the best performance, with the highest AUC (0.841). ROC curves for the genes selected by machine learning in the training and control groups are shown in Supplementary Figure 5a–h. In addition, integration with the WGCNA results showed that *S100A12* and *UPP1* were selected from the yellow module, *CD22* from the brown module, and *CSTA* from the blue module. These results confirmed the strong diagnostic performance of the model based on *S100A12*, *UPP1*, *CD22*, and *CSTA* (Supplementary Table 8). Thus, the selected features are clearly associated with sepsis diagnosis and warrant further investigation as therapeutic targets.

Independent dataset validation of diagnostic model

To evaluate the predictive performance of the diagnostic model, we obtained GSE65682 from the GEO database for external validation. ROC curve analysis was performed for the selected overlapping genes (Supplementary Table 9). In the GSE65682 external validation set, the AUC of the four-gene diagnostic model was 0.860, with sensitivity of 0.781 and specificity of 0.780 (Supplementary Fig. 5i, Supplementary Table 9). External validation confirmed that the four-gene diagnostic model performed well in sepsis.

Nomogram, DCA, and CIC visualization of the diagnostic model

To visualize the diagnostic model, the four diagnostic genes were integrated into a risk nomogram to predict sepsis occurrence (Fig. 5a). The calibration curve for sepsis occurrence showed that the actual occurrence rate closely matched the rate predicted by the nomogram (Fig. 5b), indicating good predictive value. Figure 5c presents the DCA for the diagnostic genes (*S100A12*, *UPP1*, *CD22*, and *CSTA*) and the integrated model. Based on the DCA results, we further plotted the CIC to assess the clinical utility of the nomogram. CIC visualization showed superior overall net benefits across a wide range of threshold probabilities, indicating that the diagnostic model had excellent predictive value (Fig. 5d). The same analyses, including the risk nomogram, calibration curve, DCA, and CIC, were also performed in the validation group for the four selected genes (Fig. 5e–h).

Interpretability analysis and model optimization of SHAP

Among the tested algorithms, XGBoost achieved the highest AUC

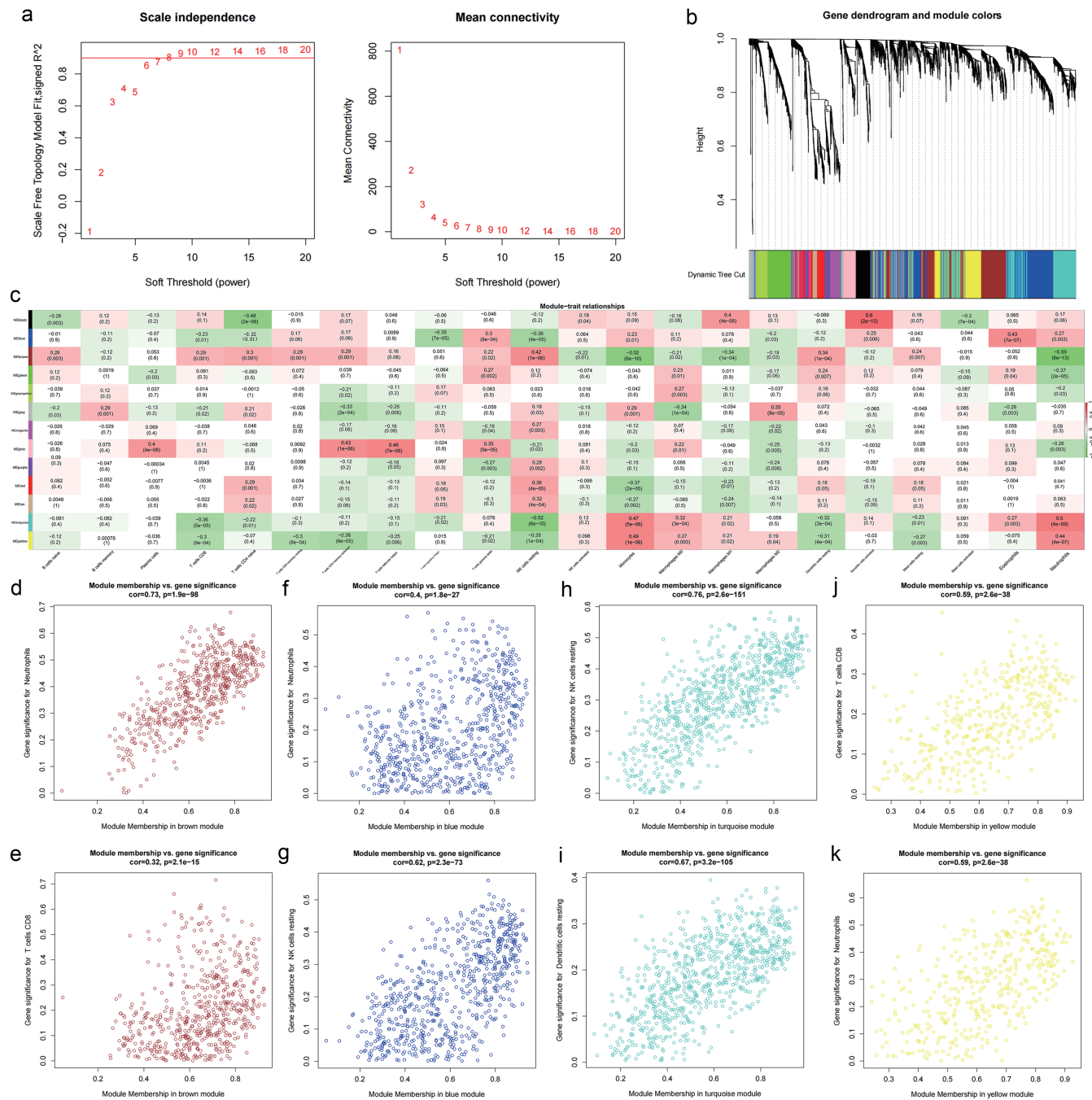


Fig. 3. Weighted gene co-expression network analysis (WGCNA). (a) Scale-free topology fitting index plot used to select the optimal soft threshold (power). (b) Hierarchical clustering tree diagram for module identification. (c) Correlation plot between sepsis-related genes and immune cells. (d) Correlation between the brown module and neutrophils. (e) Correlation between the brown module and CD8+ T cells. (f) Correlation between the blue module and neutrophils. (g) Correlation between the blue module and resting natural killer (NK) cells. (h) Correlation between the turquoise module and resting NK cells. (i) Correlation between the turquoise module and resting dendritic cells. (j) Correlation between the yellow module and CD8+ T cells. (k) Correlation between the yellow module and neutrophils.

(0.991; Fig. 6a). The SHAP bar plot showed that the mean SHAP value of *UPP1* was the highest among the four diagnostic genes (Fig. 6b), indicating that it contributed most to the model. This finding was consistent with ROC curve analysis in the training group using the neural network model. The SHAP bee plot also

showed that *UPP1* had the largest mean SHAP value (Fig. 6c). For *UPP1*, *CSTA*, and *S100A12*, higher expression was associated with classification as sepsis, whereas for *CD22*, higher expression was associated with classification as healthy control. The waterfall plot showed that *UPP1* and *S100A12* had relatively large effects

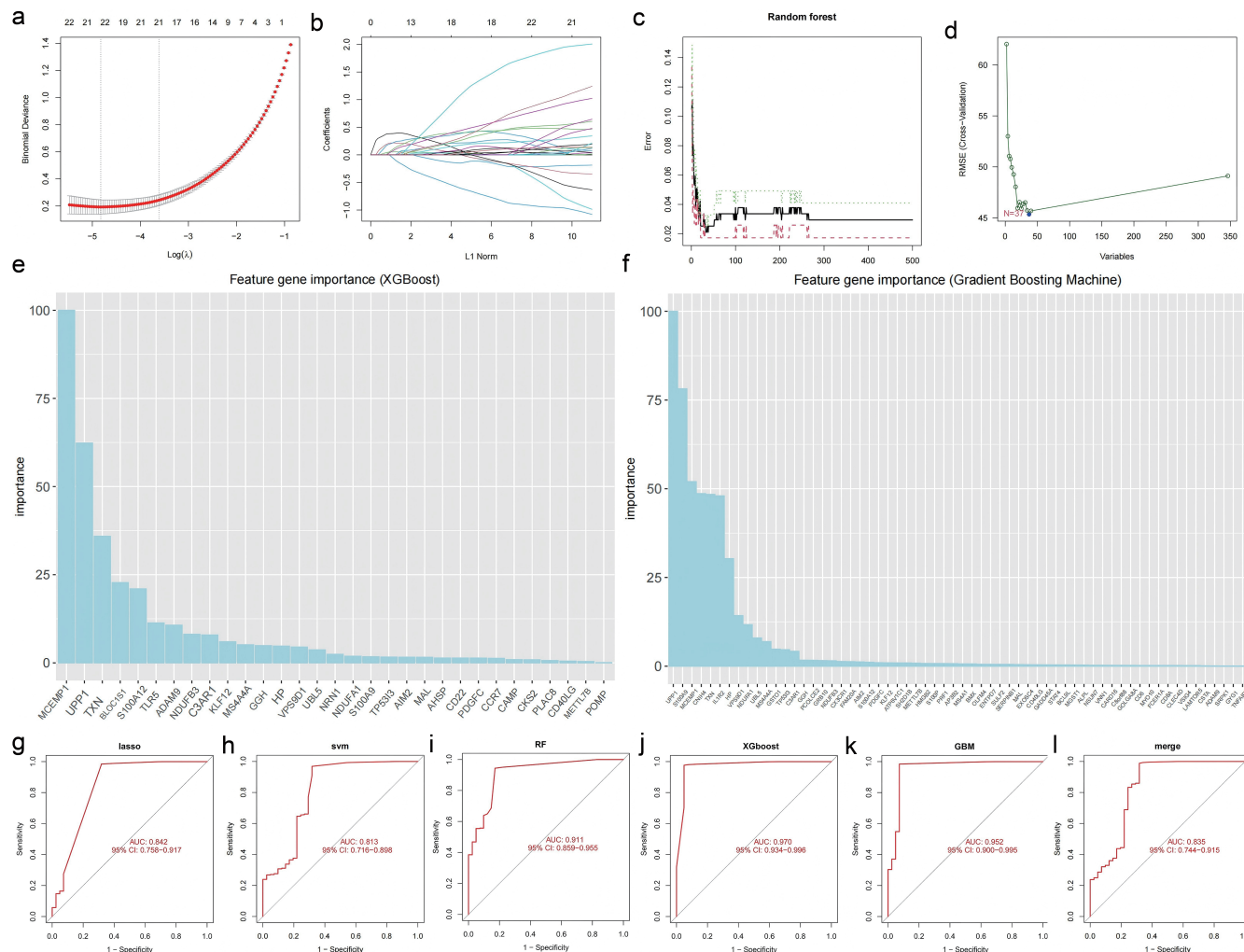


Fig. 4. Screening of sepsis-related diagnostic genes using machine learning and evaluation of feature-selection methods. (a) Mean squared error versus $\log(\lambda)$ in Lasso regression. (b) Regression coefficient versus $\log(\lambda)$ curve. (c) Random forest analysis plot; the horizontal axis represents the number of trees, and the vertical axis represents the cross-validation error. (d) Screening of 37 key genes using the support vector machine-recursive feature elimination (SVM-RFE) algorithm. (e) Genes ranked by feature importance using the eXtreme Gradient Boosting (XGBoost) algorithm. (f) Genes ranked by feature importance using the Gradient Boosting Machine (GBM) algorithm. (g) Receiver operating characteristic (ROC) curve of Lasso. (h) ROC curve of SVM-RFE. (i) ROC curve of random forest. (j) ROC curve of XGBoost. (k) ROC curve of GBM. (l) ROC curve of the integrated machine-learning model. CI, confidence interval.

on prediction results (Fig. 6d). The combined predictive score of the four-gene diagnostic model for the representative sample was -0.00025 , which was lower than the predefined cutoff of 0.512 , indicating that the sample was classified as negative (healthy control). The force plot was consistent with the waterfall plot (Fig. 6e).

Immune function and immune correlation analysis of diagnostic genes

Significant differences in immune-related functions were observed between the high- and low-expression groups for *S100A12*, *CD22*, *CSTA*, and *UPP1* (Supplementary Tables 10–13). Visualization results for immune function analysis are presented in Supplementary Figure 6a–d. We performed immune-correlation analyses of *S100A12*, *CD22*, *CSTA*, and *UPP1* to explore their associations with key immune-cell subsets in sepsis (Supplementary Fig. 6e–h). These results suggest that *S100A12* and *UPP1* participate in sepsis progression by enhancing pro-inflammatory responses and

neutrophil activation, whereas *CD22* and *CSTA* may affect disease progression by regulating B-cell function and immunosuppressive pathways. Gene-immune-cell correlations further revealed specific regulatory networks of diagnostic genes in the sepsis immune microenvironment.

Two-phase single-cell RNA-seq analysis: initial profiling of sepsis patients and subsequent group comparison between sepsis and healthy control groups

The percentage of mitochondrial genes in both groups was relatively low (mostly $<20\%$) (Supplementary Fig. 7a and d). Cells with mitochondrial gene content $>15\%$ or gene count <50 were filtered. Sequencing depth was strongly positively correlated with the number of genes (correlation coefficients = 0.77 and 0.87) and weakly correlated with mitochondrial content (correlation coefficients = 0.24 and -0.03) (Supplementary Fig. 7b and e). Feature-gene variance plots showed that the top 10 genes were mainly

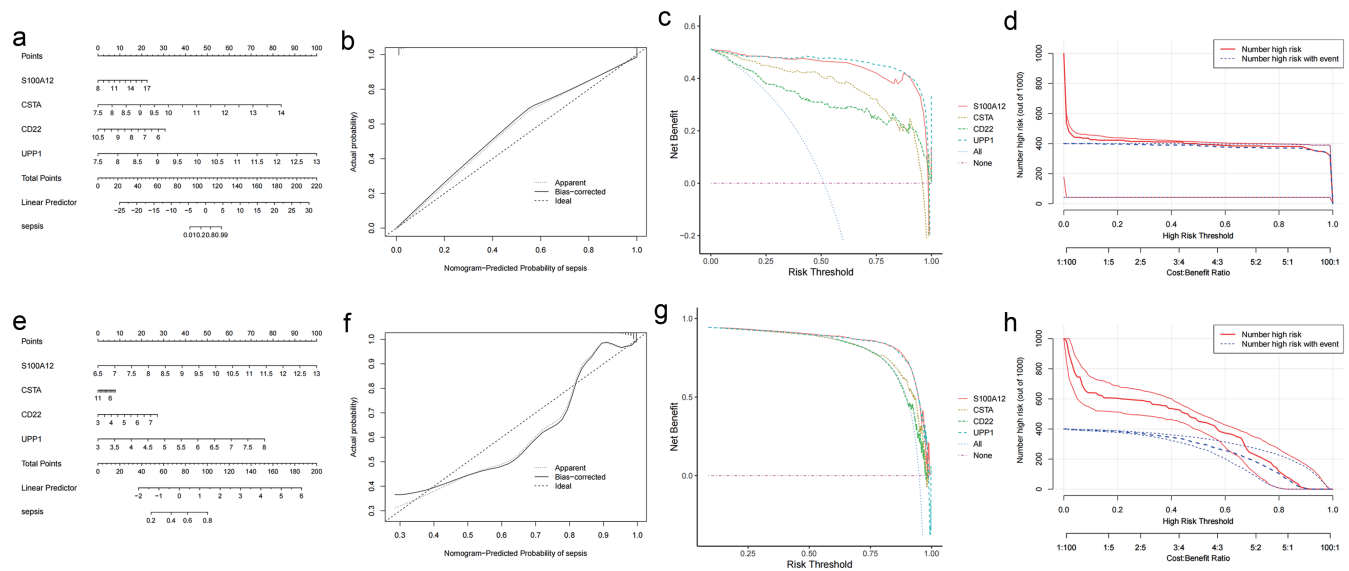


Fig. 5. Nomogram, decision curve analysis (DCA), and clinical impact curve (CIC) of the diagnostic model. (a) Nomogram for evaluating sepsis risk. (b) Calibration curve of nomogram prediction. (c) DCA curve of nomogram prediction. (d) CIC of nomogram prediction. (e-h) Nomogram, calibration curve, DCA, and CIC in the validation group.

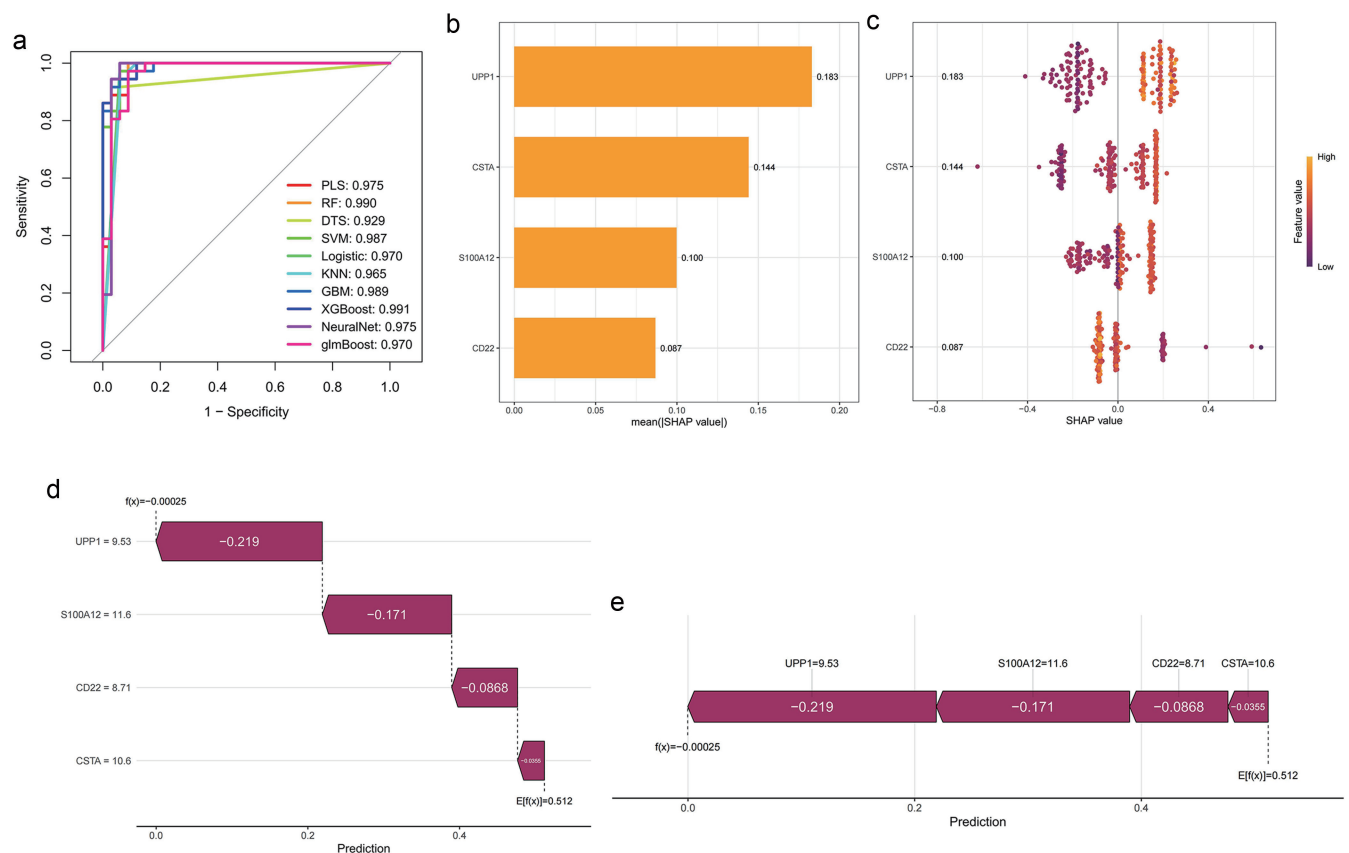


Fig. 6. Shapley additive exPlanations (SHAP)-based interpretable machine-learning analysis. (a) Multiple machine-learning algorithms, including random forest, support vector machine, eXtreme Gradient Boosting (XGBoost), and 10 additional machine-learning methods, were trained and evaluated using 5-fold repeated cross-validation, followed by receiver operating characteristic (ROC) curve analysis. (b) SHAP bar plot. (c) SHAP bee plot. (d) SHAP waterfall plot. (e) SHAP force plot.

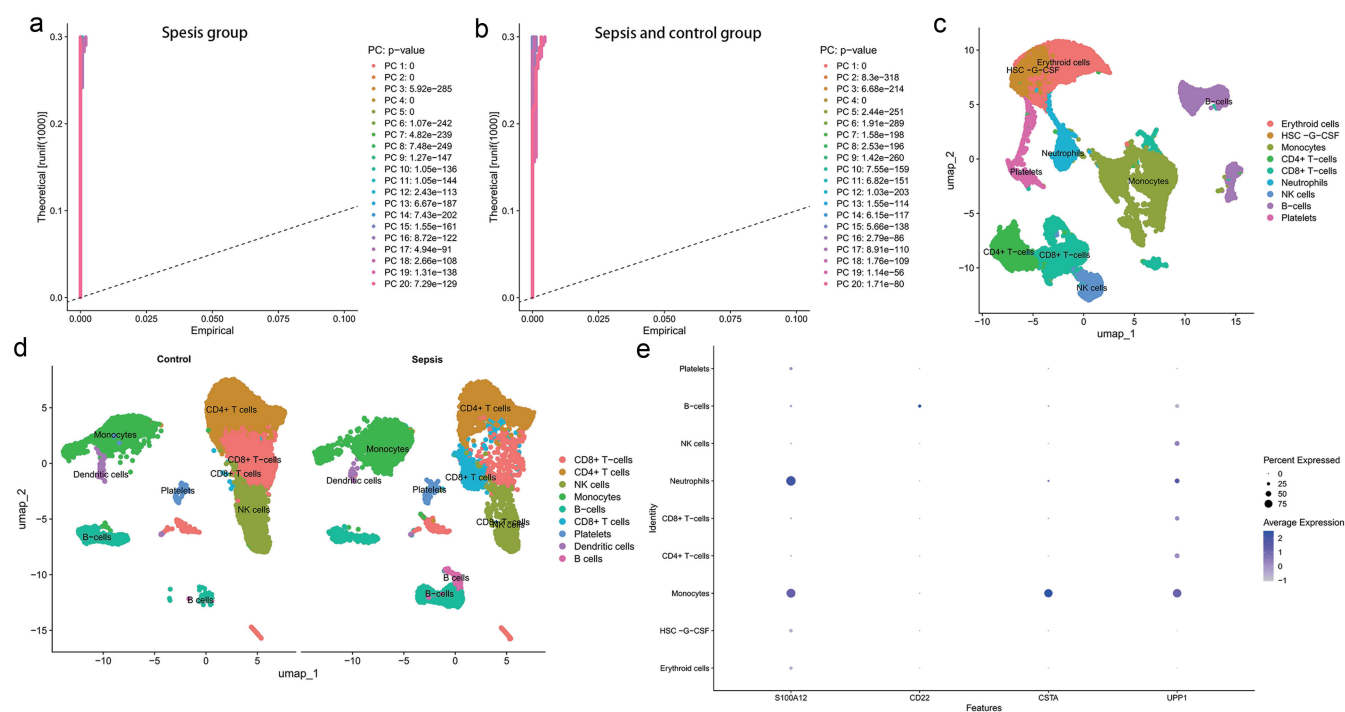


Fig. 7. Single-cell data analysis. (a) Distribution of *P* values for each principal component (PC) in the sepsis group. (b) Distribution of *P* values for each PC in the combined sepsis and healthy control population. (c) Cell-cluster analysis in the sepsis group. (d) Cell-cluster analysis in the combined sepsis and healthy control population. (e) Bubble plots of the four diagnostic genes in each cluster of the sepsis group.

IGKV-series genes (Supplementary Fig. 7c and f). PCA identified 20 significant components ($P < 0.05$) with distinct gene expression patterns in cell clusters (Fig. 7a and b). The sepsis single-cell sequencing data were then annotated (Fig. 7c). By combining cell annotation results with the immune infiltration analysis, we found that monocytes, T cells, NK cells, and neutrophils in sepsis patients were closely related to immune dysregulation in sepsis. In addition, we annotated single-cell sequencing data from the sepsis and healthy control groups (Fig. 7d). Through systematic annotation of distinct cell subsets, we identified differential gene expression patterns between sepsis and healthy control groups across various cell types. Specifically, *S100A12* expression in B cells, CD4+ T cells, CD8+ T cells, monocytes, and NK cells differed between the two groups. *CD22* expression in B cells differed between the two groups, and *CSTA* expression in CD8+ T cells differed between the two groups (Supplementary Table 14). Cell-type differential analysis and visualization of diagnostic genes in sepsis patients showed that *S100A12* and *UPP1* were upregulated in monocytes and neutrophils, *CSTA* was upregulated in monocytes, and *CD22* was downregulated in B cells (Fig. 7e, Supplementary Fig. 8a and b, and Supplementary Table 15). The cell-type difference analysis of the four diagnostic genes was consistent with our immune-correlation analysis, indicating that these genes regulate different immune cells and participate in sepsis progression. Finally, cell trajectory analysis showed that B cells were the most differentiated, followed by CD4+ T cells, CD8+ T cells, and dendritic cells, with subsequent differentiation into monocytes, CD8+ T cells, erythroid cells, NK cells, and platelets (Supplementary Fig. 9).

Cell communication analysis at the pathway level

The ligand-receptor pair analysis showed that the interactions mainly involved secreted signaling, extracellular matrix (ECM)-

receptor interactions, and cell-cell contact, with most annotations derived from the KEGG database (Fig. 8a). In the graph showing the number of intercellular interactions, the monocyte node was the largest, indicating that monocytes were the most abundant interacting cells. The connection between monocytes and B cells was the thickest, suggesting that monocytes may act as ligand-sending cells and B cells as receptor cells (Fig. 8b). Figure 8c shows the intensity of intercellular interactions, in which line thickness represents interaction strength and weight. The line connecting monocytes and CD8+ T cells was the thickest, indicating the strongest interaction between these cell types. According to the bubble plot, interactions between B cells and monocytes and between CD8+ T cells and monocytes were most likely mediated through the macrophage migration inhibitory factor receptor-ligand pair (CD74 + CD44) (Fig. 8d). We then analyzed cell communication at the pathway level. Among the pathways examined, we focused on the protease-activated receptor (PAR) pathway because of its role in pro- and anti-inflammatory mechanisms. The cell-communication heatmap suggested that, in the PAR pathway, CD8+ T cells can act as ligand cells that send signals to NK cells (Fig. 8e). Cell-type analysis further indicated that CD8+ T cells can act as senders and NK cells as receivers (Fig. 8f). These pathway-level cell-communication analyses indicate that pro-inflammatory pathways are important in sepsis development and that CD8+ T cells and NK cells play central roles, consistent with the immune-infiltration and enrichment analyses.

Screening of sepsis-related drugs and molecular docking with diagnostic genes

Functional associations of the four diagnostic genes were further explored via GSVA using two databases: KEGG and GO. Specifically, Supplementary Figures 10a, c, e, and g correspond to KEGG-

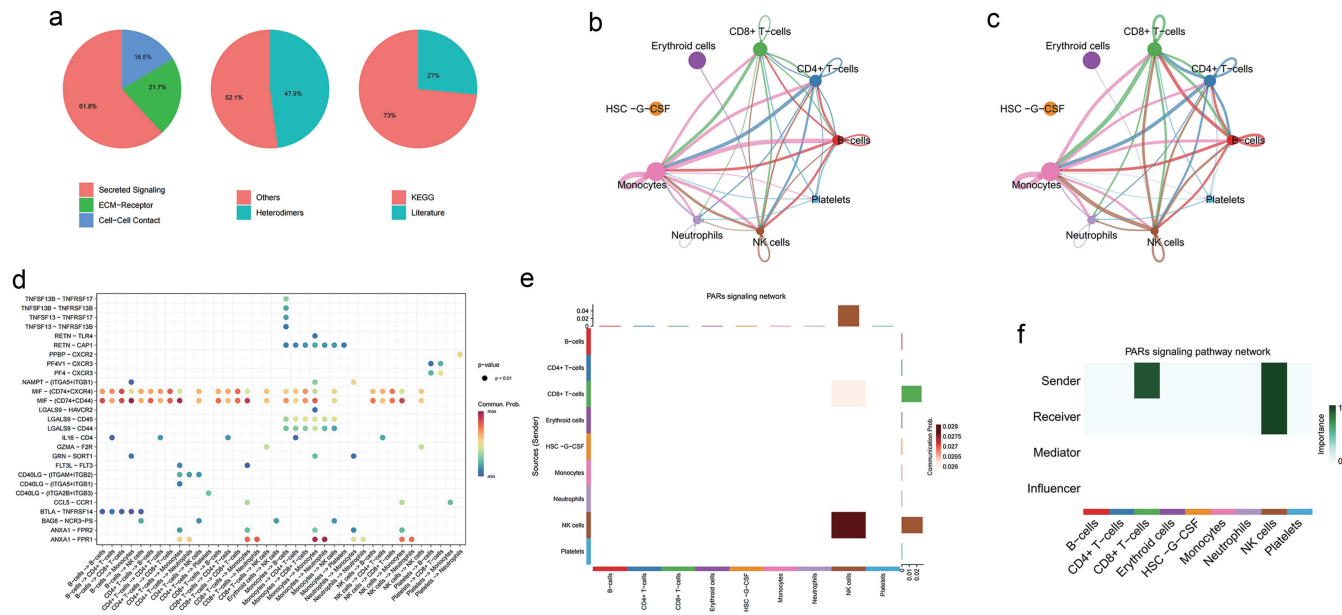


Fig. 8. Cell communication analysis. (a) Distribution of ligand-receptor pair types. (b) Cell-communication network showing the number of interactions. (c) Cell-communication network showing interaction strength. (d) Bubble plot of receptor-ligand pairs. (e) Heatmap of cell communication. (f) Cell-type analysis diagram. ECM, extracellular matrix; KEGG, Kyoto Encyclopedia of Genes and Genomes.

based GSEA results, whereas [Supplementary Figures 10b, d, f, and h](#) represent GO-based GSEA findings for *S100A12*, *CD22*, *CSTA*, and *UPP1*, respectively.

For *S100A12* (significantly upregulated in sepsis), KEGG-based GSEA revealed higher enrichment scores for the primary immunodeficiency and T-cell receptor signaling pathways ([Supplementary Fig. 10a](#)). Additionally, GO-based GSEA for *S100A12* showed significant upregulation of pathways related to the regulation of lymphocyte apoptotic processes, which was associated with the negative regulation of pantothenic acid and coenzyme A biosynthesis pathways, as well as autophagy regulation ([Supplementary Fig. 10b](#)).

Regarding *CD22* (downregulated in sepsis), KEGG-based GSEA indicated upregulation of the complement and coagulation cascades and cytoplasmic DNA-sensing pathways, alongside marked reductions in B-cell receptor signaling, primary immunodeficiency, and T-cell receptor signaling pathways ([Supplementary Fig. 10c](#)).³¹ GO-based GSEA for *CD22* further revealed significant upregulation of pathways involved in the classical pathway of complement activation ([Supplementary Fig. 10d](#)).

For *CSTA* (upregulated in sepsis), KEGG-based GSEA demonstrated its effects on signaling pathways (e.g., phosphatidylinositol signaling) and immune responses ([Supplementary Fig. 10e](#)). GO-based GSEA for *CSTA* showed upregulation of mitophagy-related pathways ([Supplementary Fig. 10f](#)).

For *UPP1* (upregulated in sepsis), KEGG-based GSEA revealed higher enrichment scores for the primary immunodeficiency and T-cell receptor signaling pathways ([Supplementary Fig. 10g](#)). GO-based GSEA for *UPP1* indicated significant upregulation of pathways related to the negative regulation of smooth muscle cell differentiation, vascular-associated pathways, and the positive regulation of Rho protein signaling ([Supplementary Fig. 10h](#)).

Molecular docking showed favorable binding properties. Acetaminophen bound to *S100A12*, with an optimal docking energy of -5.4 kcal/mol ([Fig. 9a](#)). Estradiol bound to *CD22*, with an optimal

docking energy of -6.6 kcal/mol ([Fig. 9b](#)). Dexamethasone bound to *UPP1*, with a docking energy of -8.4 kcal/mol ([Fig. 9c](#)), and aspirin bound to *CSTA*, with a docking energy of -6.6 kcal/mol ([Fig. 9d](#)).

Discussion

Sepsis is a highly lethal syndrome and a serious global public health issue. The Sepsis-3 definition emphasizes life-threatening organ dysfunction caused by a dysregulated host response to infection, marked by excessive inflammation and immunosuppression. Among the various cell types and mediators involved in sepsis-related excessive inflammation, prominent features include leukocytes (such as neutrophils, macrophages, and NK cells), endothelial cells, cytokines, complement products, and activation of the coagulation system.^{32–35} In recent years, the critical role of immune cell apoptosis in sepsis-related immune dysfunction has been elucidated.³⁶ Sepsis-induced immune cell apoptosis not only leads to depletion of key immune effector cells but also contributes to immunosuppression. High-throughput sequencing is a major advance in genomics research and has been widely applied in the search for disease candidate genes.³⁷ Machine learning, a subset of artificial intelligence, uses data and algorithms to identify patterns and can contribute to the diagnosis, prediction, and treatment of sepsis.³⁸ However, relying on a single machine learning method for feature screening may lead to method-specific bias. Therefore, this study integrated multiple machine learning methods to develop a diagnostic model for sepsis. Combining the advantages of each machine learning approach reduced method-specific biases that can arise during feature selection. External validation and SHAP interpretability analysis were subsequently performed to assess the feasibility of the diagnostic model, and the results indicated that the selected genes had high predictive value. The diagnostic genes identified through these methods may improve overall predictive accuracy.³⁹

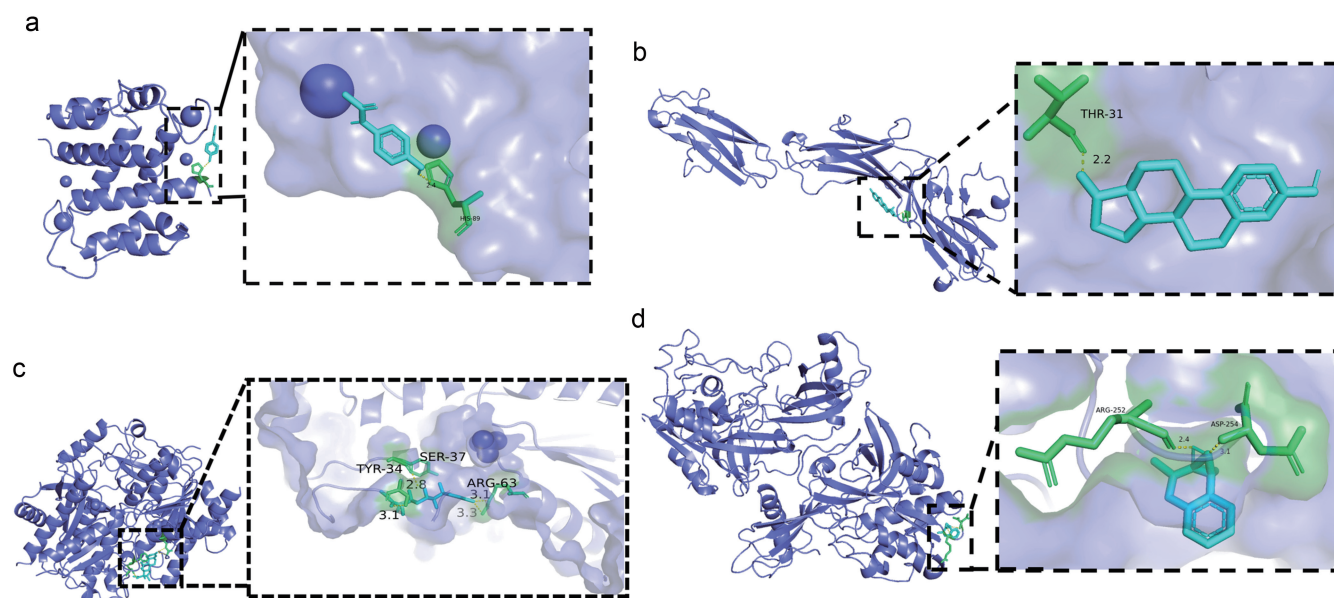


Fig. 9. Molecular docking of diagnostic genes and candidate compounds. (a) Docking result of *S100A12* with acetaminophen. (b) Docking result of *CD22* with estradiol. (c) Docking result of *UPP1* with dexamethasone. (d) Docking result of *CSTA* with aspirin.

In this study, we identified 346 DEGs between sepsis patients and healthy controls, with 230 genes upregulated and 116 down-regulated in sepsis samples. Through GO and KEGG enrichment analyses, we found that the DEGs between sepsis patients and healthy controls were mainly enriched in pathways related to immune receptor activity, cytokine binding, T-cell differentiation, immune response regulation, cell surface receptor signaling, and T helper cell differentiation. These findings suggest that immunosuppression in sepsis involves various cell types and features, such as enhanced immune cell apoptosis, T-cell dysfunction, and impaired T-cell receptor signaling. Gene alterations are associated with cellular reprogramming and reduced expression of activated cell-surface molecules. Immunosuppression is linked to increased susceptibility to secondary infections in sepsis patients, often caused by opportunistic pathogens and viral reactivation. In sepsis, apoptosis occurs primarily in T cells, B cells, NK cells, and dendritic cells and may play a crucial role in shaping the immune microenvironment.

Significant differences were observed in neutrophils, monocytes, $\gamma\delta$ T cells, resting CD4⁺ T cells, CD8⁺ T cells, NK cells, and M0 macrophages between sepsis patients and healthy controls. Activation of these immune cells is a hallmark of excessive inflammation in sepsis. Our data suggest that sepsis development is also associated with significant lymphocyte depletion, characterized by decreased CD8⁺ and CD4⁺ T cells and NK cells. Previous studies have shown that neutrophils can promote excessive inflammation in sepsis by releasing proteases and reactive oxygen species (ROS).⁴⁰ Neutrophils can release neutrophil extracellular traps (NETs), which consist of chromatin fibers containing antimicrobial peptides and proteases, such as myeloperoxidase, elastase, and proteinase G. NETs capture and kill bacteria and promote antimicrobial defense, whereas deoxyribonuclease I (DNase) inhibits NET formation, increases bacterial load in the blood, and reduces survival in septic animals. However, like many innate immune components, NETs have dual roles during infection. In sepsis, excessive NETosis may be harmful through multiple mechanisms,

including intravascular thrombosis and multiple organ failure. In our immune differential analysis, neutrophils were more abundant in sepsis patients than in healthy controls, suggesting that excess neutrophil activation may contribute to thrombosis and multiple organ failure in sepsis.⁴¹ *CD8A* had the highest degree score among proteins in the adaptive immune pathway related to T-cell differentiation, highlighting its importance in the sepsis immune microenvironment and underscoring the association between sepsis pathogenesis and adaptive immunity. GSEA highlighted complement and coagulation cascades as central to sepsis pathogenesis.⁴² These two evolutionarily linked systems drive pro-inflammatory responses: complement activation releases C3a/C5a, recruiting leukocytes and endothelial cells, whereas uncontrolled activation causes tissue damage. Conversely, coagulation activation initiates immune thrombosis, aiding pathogen defense but exacerbating microvascular thrombosis in sepsis. Dysregulation of these systems can culminate in disseminated intravascular coagulation, reflecting their dual roles in immune protection and pathological injury.⁴³

S100A12 was significantly upregulated in sepsis. According to the GSVA results, *S100A12* was significantly upregulated in immune-related pathways, the T-cell receptor signaling pathway, and pathways related to the regulation of lymphocyte apoptosis in sepsis patients, indicating its involvement in immune responses and pathogen clearance. *S100A12* is an EF-hand calcium-binding protein of the S100 family that is primarily expressed and secreted by neutrophils. According to our integrated omics results, monocytes and neutrophils were increased in sepsis patients, and *S100A12* expression in these two cell types was increased during sepsis. Combined with the ROC curve, nomogram, and SHAP interpretability analyses of *S100A12* expression in the training and validation groups, *S100A12* may serve as a promising marker for the diagnosis of sepsis. Clinical evidence suggests that *S100A12* may be a sensitive and specific diagnostic biomarker for local inflammatory processes.⁴⁴ Recent research indicates that acetaminophen may prevent and treat organ dysfunction in critically ill patients with sepsis.⁴⁵ Molecular docking showed that acetaminophen

could bind closely to *S100A12*, with an optimal docking binding energy of -5.4 kcal/mol.

CD22 (Siglec-2) is a member of the sialoglycan-binding immunoglobulin-like lectin family (Siglecs). As a core inhibitory receptor of B cells, *CD22* exerts its regulatory effects primarily by recruiting the tyrosine phosphatase SHP-1 via its intracellular immunoreceptor tyrosine-based inhibitory motif and dephosphorylating adjacent substrates to dampen excessive B-cell receptor (BCR) signaling. This “brake” mechanism is essential for preventing aberrant B-cell activation and maintaining tonic signaling thresholds during B-cell development, thereby ensuring quality control of functional B cells and humoral immune homeostasis.^{46,47} The significant downregulation of *CD22* in sepsis, together with the reduced activity of BCR signaling and the key immune effector pathways identified by GSVA, may be biologically important because it directly links *CD22* to sepsis-induced immune dysfunction. In sepsis, the loss of *CD22* expression disrupts this inhibitory cascade: diminished *CD22* levels impair SHP-1-mediated dephosphorylation, leading to unchecked BCR signaling that drives hyperactivation of mature B cells. Furthermore, *CD22* enhances its self-regulatory capacity through homotypic clustering, a process that amplifies local *CD22* concentration and strengthens BCR signal suppression.⁴⁸ The downregulation of *CD22* in sepsis abrogates this synergistic inhibitory effect, further exacerbating B-cell overactivation. Notably, *CD22*'s role in maintaining immune tolerance (as implicated in autoimmune diseases, such as systemic lupus erythematosus) suggests that its downregulation in sepsis may also disrupt B-cell tolerance, promoting the activation of autoreactive B cells and further fueling immunopathological damage.^{49,50} Collectively, the downregulation of *CD22* in sepsis removes a key inhibitory checkpoint of B-cell activation, triggering a cascade of humoral immune dysregulation characterized by excessive inflammation, impaired immune homeostasis, and failed pathogen clearance, all of which are central to the progression of sepsis-related immune dysfunction. This finding highlights *CD22* as a critical regulatory node in sepsis-associated immune derangement, underscoring its potential as a target for restoring the immune balance in sepsis. Estradiol has been shown to improve inflammatory responses.⁵¹ Molecular docking suggested a possible interaction between estradiol and *CD22*, with an optimal docking binding energy of -6.6 kcal/mol.

UPP1 encodes uridine phosphorylase 1, an enzyme involved in pyrimidine metabolism.⁵² According to the GSVA results from the GO database, *UPP1* was significantly upregulated in the Rho protein signaling pathway. In severe infection and sepsis, *UPP1*-related metabolic changes may be associated with vascular injury, systemic inflammatory response syndrome, and hypercoagulability,^{53,54} whereas Rho proteins, a family of GTPases in the Ras superfamily, play important roles in eukaryotic cells, particularly in cytoskeletal assembly.⁵⁵ In *Escherichia coli*, Rho proteins terminate transcription by removing RNA polymerase from the DNA template via RNA-dependent ATPase activity. The upregulation of *UPP1* promotes transcription termination in *Escherichia coli*, which is closely related to metabolic disorders and immune regulation induced by sepsis.⁵⁶ These findings suggest that targeting uridine metabolism could support the development of new therapies for cancer and metabolic diseases and may also help regulate immune responses. Taken together, our omics results showed that monocytes and neutrophils were increased in sepsis patients and that *UPP1* expression was increased in both cell types during sepsis. Combined with the ROC curve, nomogram, and SHAP interpretability analyses of *UPP1* expression in the training and validation groups, *UPP1* may also be a promising marker for the

diagnosis of sepsis. Studies have shown that dexamethasone may improve endothelial injury and inflammation.⁵⁷ Molecular docking suggested tight binding between dexamethasone and *UPP1*, with a docking binding energy of -8.4 kcal/mol.

CSTA is involved in clathrin-mediated endocytosis, vesicle transport, membrane dynamics, autophagy, cell division/cytokinesis, and cell migration.⁵⁸ GSVA results from the GO database also showed that upregulated *CSTA* in sepsis affected signaling pathways such as phosphoinositide signaling, which plays an important role in clathrin-mediated endocytosis, vesicle transport, membrane dynamics, autophagy, cell division/cytokinesis, and cell migration by inducing changes in the cytoskeleton and actin remodeling. *CSTA* is upregulated in mitophagy and plays an important role in apoptosis. Apoptosis and necrosis are two forms of cell death, with apoptosis playing an important role in maintaining tissue homeostasis. Apoptosis occurs through two distinct pathways: the receptor-activated caspase 8-mediated pathway and the mitochondrial caspase 9-mediated pathway, with caspase 3 activation being the final common pathway for both. Previous studies suggest that extensive apoptosis in cells from other organs occurs during the later stages of sepsis, leading to multiple organ dysfunction.⁵⁹ Thus, *CSTA* may play a key role in sepsis. Aspirin is one of the most widely used antipyretic, analgesic, and anti-inflammatory drugs globally and also has anti-thrombotic effects.⁶⁰ Molecular docking suggested tight binding between aspirin and *CSTA*, with a binding energy of -6.6 kcal/mol.

Although treatments for sepsis have advanced, mortality remains high. Therefore, our research team aims to explore whether therapeutic agents targeting the four key genes (*S100A12*, *CD22*, *CSTA*, and *UPP1*) could mitigate sepsis progression.

While molecular docking offers valuable preliminary insights into the binding modes between target proteins and candidate ligands, it has inherent limitations that preclude direct clinical extrapolation. Notably, docking predicts binding affinity trends but cannot quantify true dissociation constants, a key parameter requiring validation by surface plasmon resonance (SPR) or isothermal titration calorimetry. It also does not account for *in vivo* pharmacokinetics (e.g., bioavailability and metabolism), drug toxicity, or off-target effects, which are critical for clinical applicability. Thus, our docking results should be interpreted as a preliminary screening tool to prioritize candidates, not as evidence of clinical efficacy. Our research team will validate binding affinity via SPR and assess therapeutic potential through *in vitro* and *in vivo* experiments to address these gaps.

Single-cell sequencing data showed that *S100A12* was specifically upregulated in neutrophils, suggesting that it may be involved in sepsis progression by regulating neutrophil-mediated inflammatory responses. *CD22* was specifically downregulated in B cells from sepsis patients, suggesting that it may participate in sepsis progression by disrupting B cell-mediated antigen presentation or antibody secretion. *CSTA* was specifically upregulated in monocytes from sepsis patients, suggesting that it may participate in sepsis-related immune imbalance by inhibiting protease activity and regulating the release of inflammatory mediators in monocytes. *UPP1* was specifically upregulated in monocytes and neutrophils from sepsis patients, suggesting that it may participate in excessive inflammatory responses by enhancing nucleoside metabolism and promoting the synthesis and release of inflammatory mediators, such as interleukin-1 β .

The strengths of this study include the use of multiple machine-learning methods, which increased confidence in the diagnostic model, and external validation, which supported the feasibility of the model. The SHAP framework was used to interpret the gene

expression-based classification model. Multiple complementary approaches were used to characterize the immune environment of sepsis, including immune infiltration, WGCNA, and single-cell analysis. Single-cell data analysis focused on cell heterogeneity and subpopulation identification, cell development and differentiation trajectories, and intercellular communication networks. In addition, WGCNA integrated gene modules with immune cells, aiding the identification of sepsis-related genes associated with immune-cell function. Finally, molecular docking provided a preliminary strategy for prioritizing candidate pharmacological interventions.

However, this study also has some limitations. First, all transcriptomic and single-cell sequencing data analyzed in this study were obtained from public databases rather than from independent prospective clinical cohorts at our center. Incomplete clinical metadata in these public datasets limited further exploration of the associations between the four diagnostic genes and detailed clinical phenotypes of patients with sepsis. Second, this study mainly relied on multi-omics bioinformatics analyses and lacked corresponding *in vitro* cell experiments or *in vivo* animal model validation. Therefore, the immune regulatory functions and diagnostic performance of the identified signature genes require experimental validation in future studies. Third, although molecular docking provides preliminary insights into potential binding patterns between target proteins and candidate ligands, it has inherent limitations and cannot support direct clinical extrapolation. Docking can predict trends in binding affinity but cannot quantitatively determine dissociation constants (K_D), a key parameter that needs to be verified by SPR or isothermal titration calorimetry. In addition, docking does not account for *in vivo* pharmacokinetics, such as bioavailability and metabolism, or drug toxicity and off-target effects, which are essential for clinical applicability.

Conclusions

In summary, by integrating bioinformatics and multiple machine-learning algorithms, we identified four diagnostic genes for sepsis in patient whole-blood samples. These signature genes suggest that immune dysregulation, metabolic derangement, and uncontrolled apoptosis leading to organ failure may be key mechanisms in sepsis development and progression. Combined WGCNA, CIBERSORT, and single-cell sequencing analyses indicated that the regulatory effects of these diagnostic genes on immune cells may contribute to sepsis progression and that their expression levels can distinguish patients with sepsis from healthy individuals. In addition, we developed a diagnostic model that may help support individualized risk assessment and treatment optimization after further clinical validation. These findings provide new insights into the diagnosis and management of sepsis.

Supporting information

Supplementary material for this article is available at <https://doi.org/10.14218/JTCCM.2025.00027>.

Acknowledgments

The authors sincerely thank the investigators who generated and deposited the publicly available sepsis transcriptomic datasets in the GEO database. During the preparation of this work the authors used Lasso regression (v4.1-8), SVM-RFE (v1.7-14), random forest (v4.7-1.1), XGBoost (v6.0-94) and GBM (v6.0-94) in order to screen key diagnostic genes and construct diagnostic models. After

using these tools, the authors carefully reviewed, edited, and verified all content related to machine learning parameters and other parts of the manuscript as needed. The authors take full responsibility for the accuracy, integrity, and scientific rigor of the published content, including all parameters presented in [Supplementary Table 6](#) and other research-related content.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 82072232), the Science and Technology Program of Guangzhou, China (grant number 202201020028), the Special Projects in Key Areas of General Colleges and Universities in Guangdong Province (grant number 2022ZDZX2003), the 2021 Annual Medical Teaching and Education Management Reform Research Project of Jinan University (grant number 2021YXJG029).

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization (YLW, QS, MGZ); Methodology, data curation, visualization (ML, FZZ); Writing – original draft (YLW); Writing – review & editing (YLW, QS, WJG, HYY); Supervision (MGZ, WJG, HYY); Project administration, Funding acquisition (HYY). All authors have approved the final version and agreed to the publication of the manuscript.

Ethical statement

Ethics approval and informed consent were not required for this study because only publicly available and de-identified datasets were analyzed. The study was conducted in accordance with the principles of the Declaration of Helsinki (as revised in 2024).

Data sharing statement

All data used to support the findings of this study are included within the article and supplementary materials. The datasets used and analyzed in the current study are available from GEO (<http://www.ncbi.nlm.nih.gov/geo>). The analyzed datasets generated during the study are available from the corresponding author upon reasonable request.

References

- [1] Angus DC, van der Poll T. Severe sepsis and septic shock. *N Engl J Med* 2013;369(9):840–851. doi:10.1056/NEJMr1208623, PMID:23984731.
- [2] Hotchkiss RS, Monneret G, Payen D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. *Nat Rev Immunol* 2013;13(12):862–874. doi:10.1038/nri3552, PMID:24232462.
- [3] van der Poll T, Shankar-Hari M, Wiersinga WJ. The immunology of sepsis. *Immunity* 2021;54(11):2450–2464. doi:10.1016/j.immuni.2021.10.012, PMID:34758337.
- [4] Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, *et al*. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci* 2015;72(18):3425–3439. doi:10.1007/s00018-015-1934-y, PMID:26018601.

- [5] Figueiredo RQ, Raschka T, Kodamullil AT, Hofmann-Apitius M, Mubeen S, Domingo-Fernández D. Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. *Nucleic Acids Res* 2021;49(14):7939–7953. doi:10.1093/nar/gkab556, PMID:34197603.
- [6] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, *et al*. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;35(Database issue):D760–D765. doi:10.1093/nar/gkl887, PMID:17099226.
- [7] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al*. limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res* 2015;43(7):e47. doi:10.1093/nar/gkv007, PMID:25605792.
- [8] Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;43(Database issue):D1049–D1056. doi:10.1093/nar/gku1179, PMID:25428369.
- [9] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 2023;51(D1):D587–D592. doi:10.1093/nar/gkac963, PMID:36300620.
- [10] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102, PMID:16199517.
- [11] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;12(5):453–457. doi:10.1038/nmeth.3337, PMID:25822800.
- [12] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498–2504. doi:10.1101/gr.1239303, PMID:14597658.
- [13] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47(D1):D607–D613. doi:10.1093/nar/gky1131, PMID:30476243.
- [14] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559. doi:10.1186/1471-2105-9-559, PMID:19114008.
- [15] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;15(6):e8746. doi:10.15252/msb.20188746, PMID:31217225.
- [16] Armingol E, Baghdassarian HM, Lewis NE. The diversification of methods for studying cell-cell interactions and communication. *Nat Rev Genet* 2024;25(6):381–400. doi:10.1038/s41576-023-00685-8, PMID:38238518.
- [17] Freijeiro-González L, Febrero-Bande M, González-Manteiga W. A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *Int Stat Rev* 2022;90(1):118–145. doi:10.1111/insr.12469.
- [18] Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans Nanobioscience* 2005;4(3):228–234. doi:10.1109/tnb.2005.853657, PMID:16220686.
- [19] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. *Brief Bioinform* 2023;24(2):bbad002. doi:10.1093/bib/bbad002, PMID:36653905.
- [20] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13–17; San Francisco, CA, USA. New York (NY): Association for Computing Machinery; 2016:785–794. doi:10.1145/2939672.2939785.
- [21] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R (eds). *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook (NY): Curran Associates Inc; 2017:3149–3157.
- [22] Zheng Y, Wang J, Ling Z, Zhang J, Zeng Y, Wang K, *et al*. A diagnostic model for sepsis-induced acute lung injury using a consensus machine learning approach and its therapeutic implications. *J Transl Med* 2023;21(1):620. doi:10.1186/s12967-023-04499-4, PMID:37700323.
- [23] Zhang WY, Chen ZH, An XX, Li H, Zhang HL, Wu SJ, *et al*. Analysis and validation of diagnostic biomarkers and immune cell infiltration characteristics in pediatric sepsis by integrating bioinformatics and machine learning. *World J Pediatr* 2023;19(11):1094–1103. doi:10.1007/s12519-023-00717-7, PMID:37115484.
- [24] Xia F, Chen H, Liu Y, Huang L, Meng S, Xu J, *et al*. Development of genomic phenotype and immunophenotype of acute respiratory distress syndrome using autophagy and metabolism-related genes. *Front Immunol* 2023;14:1209959. doi:10.3389/fimmu.2023.1209959, PMID:37936685.
- [25] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;14:7. doi:10.1186/1471-2105-14-7.
- [26] Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol* 2015;16(4):e173–e180. doi:10.1016/S1470-2045(14)71116-7, PMID:25846097.
- [27] Lundberg SM, Lee SI. Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4–9; Long Beach, CA, USA 2017; Red Hook (NY): Curran Associates Inc; 4768–4777.
- [28] Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* 2021;61(8):3891–3898. doi:10.1021/acs.jcim.1c00203, PMID:34278794.
- [29] Seeliger D, de Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J Comput Aided Mol Des* 2010;24(5):417–422. doi:10.1007/s10822-010-9352-6, PMID:20401516.
- [30] Chapman NM, Boothby MR, Chi H. Metabolic coordination of T cell quiescence and activation. *Nat Rev Immunol* 2020;20(1):55–70. doi:10.1038/s41577-019-0203-y, PMID:31406325.
- [31] Rimmelé T, Payen D, Cantaluppi V, Marshall J, Gomez H, Gomez A, *et al*. Immune cell phenotype and function in sepsis. *Shock* 2016;45(3):282–291. doi:10.1097/SHK.0000000000000495, PMID:26529661.
- [32] Nedeva C. Inflammation and Cell Death of the Innate and Adaptive Immune System during Sepsis. *Biomolecules* 2021;11(7):1011. doi:10.3390/biom11071011, PMID:34356636.
- [33] Brown KA, Treacher DF. Neutrophils as potential therapeutic targets in sepsis. *Discov Med* 2006;6(33):118–122. PMID:17234146.
- [34] Ramoni D, Tirandi A, Montecucco F, Liberale L. Sepsis in elderly patients: the role of neutrophils in pathophysiology and therapy. *Intern Emerg Med* 2024;19(4):901–917. doi:10.1007/s11739-023-03515-1, PMID:38294676.
- [35] Boomer JS, Green JM, Hotchkiss RS. The changing immune system in sepsis: is individualized immuno-modulatory therapy the answer? *Virulence* 2014;5(1):45–56. doi:10.4161/viru.26516, PMID:24067565.
- [36] Cao C, Yu M, Chai Y. Pathological alteration and therapeutic implications of sepsis-induced immune cell apoptosis. *Cell Death Dis* 2019;10(10):782. doi:10.1038/s41419-019-2015-1, PMID:31611560.
- [37] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17(6):333–351. doi:10.1038/nrg.2016.49.
- [38] Deo RC. Machine Learning in Medicine. *Circulation* 2015;132(20):1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593, PMID:26572668.
- [39] Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv* 2021;49:107739. doi:10.1016/j.biotechadv.2021.107739, PMID:33794304.
- [40] Zou S, Jie H, Han X, Wang J. The role of neutrophil extracellular traps in sepsis and sepsis-related acute lung injury. *Int Immunopharmacol* 2023;124(PtA):110436. doi:10.1016/j.intimp.2023.110436, PMID:37688916.
- [41] Denning NL, Aziz M, Gurien SD, Wang P. DAMPs and NETs in Sep-

- sis. *Front Immunol* 2019;10:2536. doi:10.3389/fimmu.2019.02536, PMID:31736963.
- [42] Garred P, Tenner AJ, Mollnes TE. Therapeutic Targeting of the Complement System: From Rare Diseases to Pandemics. *Pharmacol Rev* 2021;73(2):792–827. doi:10.1124/pharmrev.120.000072, PMID:33687995.
- [43] Jacobi J. Pathophysiology of sepsis. *Am J Health Syst Pharm* 2002;59(Suppl 1):S3–S8. doi:10.1093/ajhp/59.suppl_1.S3, PMID:11885412.
- [44] Xia P, Ji X, Yan L, Lian S, Chen Z, Luo Y. Roles of S100A8, S100A9 and S100A12 in infection, inflammation and immunity. *Immunology* 2024;171(3):365–376. doi:10.1111/imm.13722, PMID:38013255.
- [45] Ware LB, Files DC, Fowler A, Aboodi MS, Aggarwal NR, Brower RG, *et al*. Acetaminophen for Prevention and Treatment of Organ Dysfunction in Critically Ill Patients With Sepsis: The ASTER Randomized Clinical Trial. *JAMA* 2024;332(5):390–400. doi:10.1001/jama.2024.8772, PMID:38762798.
- [46] Pezzutto A, Rabinovitch PS, Dörken B, Moldenhauer G, Clark EA. Role of the CD22 human B cell antigen in B cell triggering by anti-immunoglobulin. *J Immunol* 1988;140(6):1791–1795. PMID:3257985.
- [47] Alborzian Deh Sheikh A, Akatsu C, Abdu-Allah HHM, Sukanuma Y, Imamura A, Ando H, *et al*. The Protein Tyrosine Phosphatase SHP-1 (PTPN6) but Not CD45 (PTPRC) Is Essential for the Ligand-Mediated Regulation of CD22 in BCR-Ligated B Cells. *J Immunol* 2021;206(11):2544–2551. doi:10.4049/jimmunol.2100109, PMID:33990399.
- [48] Tsubata T. The ligand interactions of B cell Siglecs are involved in the prevention of autoimmunity to sialylated self-antigens and in the quality control of signaling-competent B cells. *Int Immunol* 2023;35(10):461–473. doi:10.1093/intimm/dxad030, PMID:37504378.
- [49] Clark EA, Giltiay NV. CD22: A Regulator of Innate and Adaptive B Cell Responses and Autoimmunity. *Front Immunol* 2018;9:2235. doi:10.3389/fimmu.2018.02235, PMID:30323814.
- [50] Abdu-Allah HHM, Wu SC, Lin CH, Tseng YY. Design, synthesis and molecular docking study of α -triazolysialosides as non-hydrolyzable and potent CD22 ligands. *Eur J Med Chem* 2020;208:112707. doi:10.1016/j.ejmech.2020.112707, PMID:32942185.
- [51] Straub RH. The complex role of estrogens in inflammation. *Endocr Rev* 2007;28(5):521–574. doi:10.1210/er.2007-0001, PMID:17640948.
- [52] Du W, Tu S, Zhang W, Zhang Y, Liu W, Xiong K, *et al*. UPP1 enhances bladder cancer progression and gemcitabine resistance through AKT. *Int J Biol Sci* 2024;20(4):1389–1409. doi:10.7150/ijbs.83774, PMID:38385072.
- [53] Ismail J, Sankar J. Systemic Inflammatory Response Syndrome (SIRS) and Sepsis - An Ever-evolving Paradigm. *Indian J Pediatr* 2015;82(8):675–676. doi:10.1007/s12098-015-1810-7, PMID:26084550.
- [54] Giustozzi M, Ehrlinder H, Bongiovanni D, Borovac JA, Guerreiro RA, Gąsecka A, *et al*. Coagulopathy and sepsis: Pathophysiology, clinical manifestations and treatment. *Blood Rev* 2021;50:100864. doi:10.1016/j.blre.2021.100864, PMID:34217531.
- [55] Bement WM, Goryachev AB, Miller AL, von Dassow G. Patterning of the cell cortex by Rho GTPases. *Nat Rev Mol Cell Biol* 2024;25(4):290–308. doi:10.1038/s41580-023-00682-z, PMID:38172611.
- [56] Nwosu ZC, Ward MH, Sajjakulnukit P, Poudel P, Ragulan C, Kasperek S, *et al*. Uridine-derived ribose fuels glucose-restricted pancreatic cancer. *Nature* 2023;618(7963):151–158. doi:10.1038/s41586-023-06073-w, PMID:37198494.
- [57] Kim WY, Kweon OJ, Cha MJ, Baek MS, Choi SH. Dexamethasone may improve severe COVID-19 via ameliorating endothelial injury and inflammation: A preliminary pilot study. *PLoS One* 2021;16(7):e0254167. doi:10.1371/journal.pone.0254167, PMID:34214123.
- [58] Hammond GR, Balla T. Polyphosphoinositide binding domains: Key to inositol lipid biology. *Biochim Biophys Acta* 2015;1851(6):746–758. doi:10.1016/j.bbalip.2015.02.013, PMID:25732852.
- [59] Lelubre C, Vincent JL. Mechanisms and treatment of organ failure in sepsis. *Nat Rev Nephrol* 2018;14(7):417–427. doi:10.1038/s41581-018-0005-7, PMID:29691495.
- [60] Hybiak J, Broniarek I, Kiryczński G, Los LD, Rosik J, Machaj F, *et al*. Aspirin and its pleiotropic application. *Eur J Pharmacol* 2020;866:172762. doi:10.1016/j.ejphar.2019.172762, PMID:31669590.