DOI: 10.4274/forbes.galenos.2025.57855
Forbes J Med 2025;6(2):97-107

Original Article / Orijinal Araştırma

Artificial Intelligence-Based Prediction of Bloodstream Infections Using Standard Hematological and Biochemical Markers

Standart Hematolojik ve Biyokimyasal Belirteçler Kullanılarak Kan Dolaşımı Enfeksiyonlarının Yapay Zeka Tabanlı Tahmini

Ferhat DEMİRCİ[1], **Murat AKŞİT**[1], **Aylin DEMİRCİ**[2]

[1]University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Medical Biochemistry, İzmir, Türkiye
[2]University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Family Medicine, İzmir, Türkiye

Cite as: Demirci F, Akşit M, Demirci A. Artificial intelligence-based prediction of bloodstream infections using standard hematological and biochemical markers. Forbes J Med. [Epub Ahead of Print]

ABSTRACT

Objective: Bloodstream infections (BSIs) require rapid identification to initiate timely antimicrobial therapy, yet blood culture-the current diagnostic gold standard-suffers from delayed results and limited sensitivity. This study aimed to develop an interpretable machine learning (ML) model using routine laboratory parameters to predict blood culture positivity.

Methods: A total of 1,972 adult patients who underwent complete blood count, C-reactive protein, procalcitonin (PCT), and blood culture testing at a tertiary hospital were retrospectively included. Three models-random forest, H_2O automated ML, and an ensemble model-were developed and evaluated using standard classification metrics [area under the curve (AUC)- receiver operating characteristic (ROC), sensitivity, specificity, F1 score]. SHapley Additive exPlanations (SHAP) analysis was employed to enhance interpretability.

Results: The ensemble model yielded the best performance, achieving an AUC-ROC of 0.95, sensitivity of 0.78, specificity of 0.97, and F1 score of 0.84. External validation on an independent cohort confirmed the model's generalizability (AUC-ROC: 0.85). SHAP analysis revealed that age and PCT were the most influential features with both statistical and clinical relevance. Basophil count, while ranked highest by SHAP, showed low sensitivity, highlighting the difference between algorithmic weight and bedside utility.

Conclusion: These findings support the integration of routine, readily available laboratory data into an explainable AI framework to accurately predict culture positivity. The model's strong performance and interpretability suggest its potential application in clinical decision support systems to improve diagnostic stewardship, reduce unnecessary cultures, and optimize resource use in suspected BSI cases.

Keywords: Sepsis, blood culture, machine learning, procalcitonin, C-reactive protein

Received/Geliş: 12.04.2025
Accepted/Kabul: 06.05.2025
Epub: 17.07.2027

Corresponding Author/ Sorumlu Yazar:
Ferhat DEMİRCİ, MD,
University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Medical Biochemistry, İzmir, Türkiye
✉ drdemirci05@gmail.com
ORCID: 0000-0002-5999-3399

ÖZ

Amaç: Kan dolaşımı enfeksiyonlarında (KDE) erken tanı, zamanında antimikrobiyal tedavi başlatılması açısından kritik öneme sahiptir. Ancak, mevcut altın standart tanı yöntemi olan kan kültürü, gecikmeli sonuç vermesi ve düşük pozitiflik oranı nedeniyle sınırlıdır. Bu çalışmanın amacı, rutin laboratuvar verilerini kullanarak kan kültürü pozitifliğini öngörebilecek yorumlanabilir bir makine öğrenimi (ML) modeli geliştirmektir.

Yöntem: Üçüncü basamak bir hastanede tam kan sayımı, C-reaktif protein, prokalsitonin (PCT) ve kan kültürü testi yapılan toplam 1.972 yetişkin hasta retrospektif olarak çalışmaya dahil edilmiştir. Rastgele orman, H_2O otomatik ML ve bir ensemble (birleşik) model olmak üzere üç farklı model geliştirilmiş ve

AUC-ROC, duyarlılık, özgüllük ve F1 skoru gibi sınıflandırma ölçütleriyle değerlendirilmiştir. Modelin yorumlanabilirliğini artırmak amacıyla SHapley Additive exPlanations (SHAP) analizi uygulanmıştır.

Bulgular: Ensemble model en iyi performansı göstermiş; [alıcı işletim karakteristiği eğrisi (AUC)-eğri altındaki alan (ROC)]: 0,95, duyarlılık: 0,78, özgüllük: 0,97 ve F1 skoru: 0,84 olarak bulunmuştur. Bağımsız bir doğrulama veri seti üzerinde yapılan analiz, modelin genellenebilirliğini doğrulamıştır (AUC-ROC: 0,85). SHAP analizine göre yaş ve PCT hem istatistiksel hem de klinik açıdan en etkili değişkenler olarak öne çıkmıştır. Basofil sayısı ise algoritmik olarak yüksek önem taşımasına rağmen düşük duyarlılığı nedeniyle klinik faydası sınırlı bulunmuştur.

Sonuç: Bu sonuçlar, rutin laboratuvar verilerinin açıklanabilir yapay zeka çerçevesinde kullanılarak kan kültürü pozitifliğinin yüksek doğrulukla öngörülebileceğini göstermektedir. Modelin güçlü performansı ve yorumlanabilirliği, tanı yönetimini iyileştirmek, gereksiz kültürleri azaltmak ve şüpheli KDE vakalarında kaynak kullanımını optimize etmek için klinik karar destek sistemlerinde potansiyel bir uygulama olduğunu göstermektedir.

Anahtar Kelimeler: Sepsis, kan kültürü, makine öğrenimi, prokalsitonin, C-reaktif protein

INTRODUCTION

Bloodstream infections and sepsis remain leading causes of morbidity and mortality, especially in critically ill and immunocompromised patients. Early identification of bacteremia is essential to initiate timely antimicrobial therapy, which can significantly reduce adverse outcomes. However, blood culture-the current gold standard diagnostic method-is limited by low positivity rates and delayed results, often requiring 24-72 hours.[1,2]

To bridge this diagnostic delay, clinicians frequently rely on nonspecific biomarkers such as complete blood count (CBC), C-reactive protein (CRP), and procalcitonin (PCT). While these markers offer some insight, their standalone predictive value remains suboptimal. Studies have shown that PCT outperforms CRP in specificity for bacterial infections, but both lack adequate sensitivity to reliably predict positive cultures.[3,4] Moreover, traditional clinical assessment is often inaccurate and inconsistent in estimating bacteremia risk.[5]

Recent developments in artificial intelligence (AI) and machine learning (ML) have opened new avenues for early bacteremia prediction using routine clinical data. Several studies have demonstrated that ML algorithms can improve predictive accuracy by combining laboratory values, vital signs, and demographic information.[6,7] These models have yielded area under the curve (AUC) values of up to 0.84, indicating high potential in differentiating between true infections and false alarms.[8]

However, many existing models are limited by dataset specificity, exclusion of key demographic factors (such as age and sex), and lack of explainability, which hinders clinical adoption.[9] Additionally, hematological markers like neutrophil-to-lymphocyte ratio, band count, and platelet levels (which are routinely available and cost-effective) are often underutilized in current models, despite evidence supporting their role in predicting bacteremia.[10]

The present study aims to address these limitations by developing a ML-based model that integrates hemogram parameters, CRP, PCT, age, and gender to predict blood culture positivity. This approach not only leverages data readily available at the point of care but also contributes to diagnostic stewardship by reducing unnecessary testing and improving the timing of antimicrobial interventions.

METHODS

Study Population/Subjects

This study was conducted at University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital. Patients who presented to this center and its affiliated hospital (AH) between January 1, 2024, and March 31, 2025, and underwent first-time blood culture, CBC, PCT, and CRP tests were included. The baseline characteristics of the study population are shown in Table 1. Patients with incomplete test results, sub-parameters missing, or contaminating agents detected in blood cultures were excluded.

Hemogram samples were analyzed in both hospitals using Sysmex XN-1000 (Kobe, Japan) hematology analyzers; CRP tests were analyzed using Beckman Coulter AU-5800 in the main hospital and Beckman Coulter AU680 (California, USA) in the AH; and PCT tests were analyzed using Siemens Advia Centaur XPT (chemiluminescence immune assay, Erlangen, Germany) at the main hospital and Beckman Coulter DXI-800 (chemiluminescence immune assay, California, USA) at the AH.

Venous blood samples were collected under aseptic conditions into automated blood culture bottles from the Biomerieux BacT/Alert 3D (France) brand. Translated with DeepL.com (free version). The bottles were placed in the corresponding brand-specific incubator system for continuous monitoring. Bottles that flagged positive for microbial growth were subcultured onto appropriate culture media. Following incubation, colony morphology was assessed, and species identification was performed by a clinical microbiologist using Gram staining, biochemical assays, and/or automated identification systems.

The reagents and calibrators were provided by Sysmex for hemogram analyses and by Beckman Coulter for CRP and PCT measurements. Using the respective automated analyzers, all analyses were carried out in accordance with

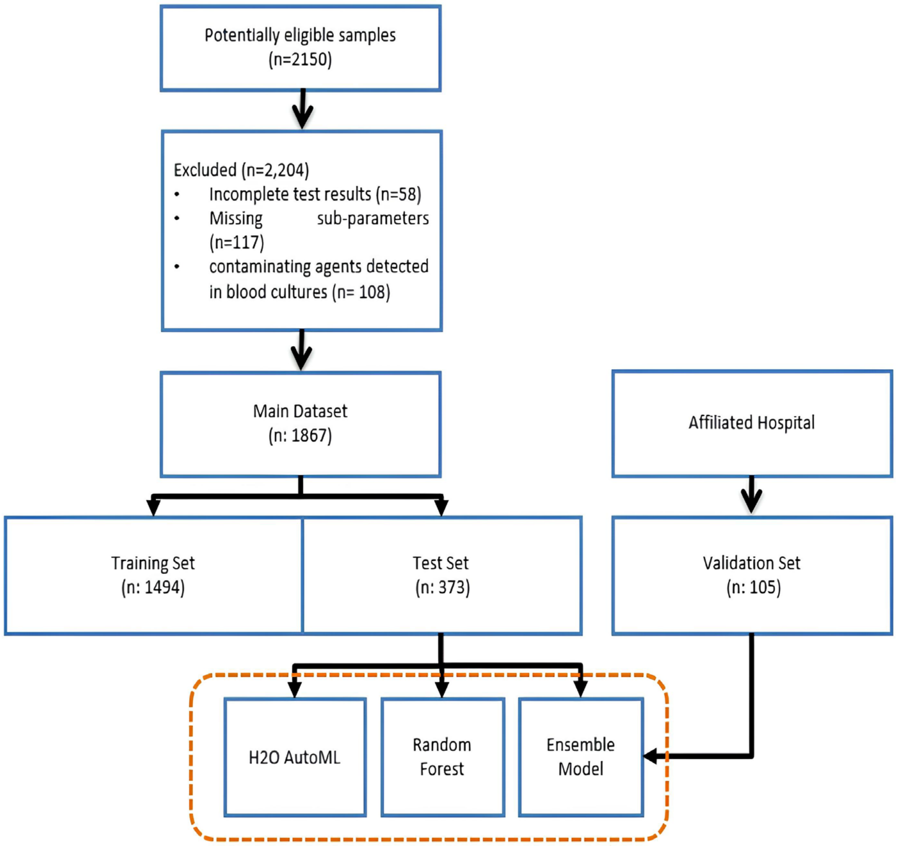

Figure 1. The Standards for Reporting Diagnostic Accuracy diagram

AutoML: Automated machine learning

the manufacturers' instructions. Routine maintenance and calibration procedures were performed regularly to ensure analytical accuracy and reliability.

Internal quality control for the hemogram was ensured using materials supplied by the manufacturer (Sysmex), whereas quality control for CRP and PCT assays was performed using materials obtained from Bio-Rad (California, USA).

Study Design

Before starting this retrospective study, ethical approval was obtained from the Ethics Committee of University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital (decision number: 2025/03-27, date: 10.04.2025). Patient identity information was anonymized, and a dataset containing age, sex, CRP, PCT, CBC, and blood culture results from 2345 patients (2150 from the main building and 195 from the AH) was transferred to Microsoft Excel 2021 (USA).

After applying exclusion criteria, the final dataset included 1972 patients (1867 from the main hospital and 105 from the AH). This dataset was then transferred to Python software (version 3.11, USA) for ML analysis.

Following data cleaning, the dataset was randomly partitioned into training and testing sets in an 80:20 ratio using stratified sampling based on the binary outcome variable, ensuring preservation of the original class distribution. The Standards for Reporting Diagnostic Accuracy diagram illustrating the patient flow throughout the study is presented in Figure 1.

Data Preprocessing and Training of Machine Learning Algorithms

For data preprocessing, patient results were transferred to Microsoft Excel. Cases with missing values were excluded from the dataset. Bacterial culture results were evaluated and converted into a binary classification. Patient samples in which bacterial species were identified by a clinical microbiologist were classified as "growth present = 1," whereas samples with no growth were classified as "no growth = 0." Samples reported as contamination were excluded from the study. Additionally, sex was encoded as a binary variable, with male = 0 and female = 1. The cleaned dataset was transferred to Python for ML analysis.

The models were trained using the 15 most important predictive parameters, which included:

- **Demographic variables:** Age, sex
- **Biochemical variables:** PCT and CRP
- **Hematologic variables:** Hemogram leucocyte variables (total white blood cell, neutrophil, lymphocyte, monocyte, eosinophile, basophile count) and hemoglobin.

Following model training, performance evaluation was conducted using the test and validation datasets.

The cleaned dataset was imported into the Python programming environment for ML analysis. Model development was conducted using Python within the PyCharm integrated development environment (IDE). PyCharm is a widely adopted and robust IDE for Python, offering advanced functionalities such as intelligent code completion, comprehensive debugging tools, and integrated testing frameworks. These features facilitate efficient data preprocessing, model training, and algorithm optimization, while providing seamless integration with widely used ML libraries, including scikit-learn, thereby supporting streamlined and scalable project workflows.[11]

A total of three AI ML algorithms were evaluated in this study: random forest (RF), H_2O automated ML (AutoML) (version 3.46), and an ensemble ML method. Model development was carried out in a Python 3.11 environment using H_2O AutoML.[12] To overcome the limitations inherent in manual model development-particularly when the primary expertise of the user is not in data science-AutoML tools have emerged as a practical solution. AutoML tools automate key steps such as feature engineering, model building, and hyperparameter optimization, which traditionally require extensive domain expertise. Despite the clear advantages and the growing interest in ML applications, few studies have applied AutoML tools within the clinical laboratory context.[13] The best-performing model within the H_2O AutoML framework was selected based on AUC-receiver operating characteristic (ROC) and logloss values.

Computational Environment and Libraries

In the development of classification models using ML and deep learning techniques, a variety of open-source Python libraries were employed for data preprocessing, model training, evaluation, and visualization. All procedures were conducted within the Python 3.11 programming environment. The libraries utilized are categorized as follows:

Data Processing and Analysis

- Pandas (v1.5): For creating and manipulating data frames
- Numpy (v1.23): For numerical operations and vectorized calculations

ML Model Development

- Scikit-learn (v1.2): For implementing ML algorithms and performance evaluation
- H_2O (.frame, .model) (v3.46.0.6): H_2OAutoML

Model Evaluation and Visualization

- Matplotlib (v3.6): For data visualization and plotting
- Sklearn (.metrics, .ensemble) (v1.2): For performance metrics such as confusion matrix, ROC-AUC, and precision-recall (PR)-AUC
- Shap (v0.47): For SHapley Additive exPlanations (SHAP) analysis and feature importance visualization

Following model training, performance evaluation was conducted using the designated test dataset.

Performance Evaluation

Scikit-learn, Pandas, NumPy, SciPy, StatsModels, and Matplotlib/Seaborn-among Python's most robust libraries for ML and statistical analysis-were employed in this project. The modeling process underwent a comprehensive evaluation, including hyperparameter tuning and model selection through internal cross-validation. Model performance was assessed using multiple evaluation metrics. The following criteria were used for classification:

1. Classification Performance Metrics

- AUC-ROC
- AUC-PR
- Confusion matrix analysis
- Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR) F1 score, odds ratio

2. Model interpretability metrics

- Feature importance analysis
- SHAP graphs

3. Validation results of the predictive models were analyzed to ensure a comprehensive assessment. This structured and multifaceted evaluation approach provides a robust framework for predicting treatment modality outcomes based on laboratory-derived data.

RESULTS

Dataset Description and Data Pre-processing

The dataset used in this study included a total of 1,972 records, consisting of 1,494 entries in the training set, 373 in the test set, and an additional 105 records in the validation set. All datasets contained hemogram parameters alongside demographic data, allowing for comprehensive baseline characterization.

Baseline demographic characteristics of the study population are presented in Table 1. The mean age was 46.08±30.13 years in the training set, 44.47±30.51 years in the test set, and significantly higher in the validation set at 65.68±16 years (p<0.001). When stratified by sex, no significant differences were observed in mean age between male and female participants within each subset (all p>0.05). The reason the mean age was significantly higher in the AH compared to the main building is that the data here were obtained from patients mainly hospitalized in the palliative care ward.

Regarding sex distribution, males comprised 53.4% of the training set, 52.3% of the test set, and 61% of the validation set, while females made up 46.6%, 47.7%, and 39%, respectively. These differences were not statistically significant (p=0.277), suggesting a relatively balanced gender distribution across the subsets.

Descriptive statistics for hemogram and related variables are presented in Table 2. Among all measured biomarkers, basophil count (BASO) were the only variable showing a statistically significant difference between the datasets (p<0.001). Other parameters, including white blood cell count, neutrophils, lymphocytes, monocytes, eosinophils, hemoglobin, CRP, and PCT, did not show significant variation across the training, test, and validation cohorts (all p>0.05). This indicates general homogeneity in these biomarkers across subsets, enhancing the comparability of model training and validation.

"The performance of RF, H_2O AutoML, and ensemble models was comparatively evaluated based on their predictive capabilities, classification metrics, and interpretability. Classification metrics such as F1 score, sensitivity, specificity, and AUC-ROC were used to assess the models' ability to discriminate between classes."

Table 1. The baseline characteristics of the study population

Characteristics	Train set (n=1494) Value±SD	Test set (n=373) Value±SD	Validation set (n=105) Value±SD	p value
Age (years)	46.08±30.13	44.47±30.51	65.68±16	<0.001
Male	46.03±30.07	43.05±31.81	65.69±15.51	
Female	46.15±30.22	46.02±29.03	65.66±16.95	
Sex				0.277
Male	798 (53.4%)	195 (52.3)	64 (61%)	
Female	696 (46.6%)	178 (47.7%)	41 (39%)	
Blood culture positivity rate	363 (24.3 %)	91 (24.4 %)	36 (34.29 %)	

SD: Standart deviation

Table 2. Descriptive statistics of the hemogram and related variables

Variable	Unit	Train set Value±SD	Test set Value±SD	Validation set Value±SD	p value
White blood cell	(×10^9/L)	13.34±15.14	12.63±7.94	10.28±5.39	0.075
Neutrophil	(×10^9/L)	8.9±6.6	8.99±6.79	8.06±5.15	0.420
Lymphocyte	(×10^9/L)	3.35±12.18	2.74±3.69	1.30±0.76	0.154
Monocyte	(×10^9/L)	1.09±2.17	1.09±0.77	0.77±0.42	0.257
Eosinophil	(×10^9/L)	0.28±0.68	0.27±0.32	0.13±0.22	0.060
Basophil	(×10^9/L)	0.12±0.18	0.11±0.09	0.04±0.03	**<0.001**
Hemoglobin	(g/dL)	10.97±2.6	11.07±2.59	10.64±2.28	0.327
C-reactive protein	(mg/L)	78.23±91.44	81.97±95.94	94.03±78.09	0.217
Procalcitonin	(ng/mL)	3.48±14.05	3.36±10.68	3.76±9.14	0.794

SD: Standart deviation

Comparison of Classification Performance Metrics

The analysis began with the RF model prior to evaluating other models. To determine the optimal classification threshold, multiple cut-off values (0.3, 0.5, and 0.7) were evaluated based on their corresponding F1 scores. Among these, a threshold of 0.3 yielded the highest F1 score, indicating a better balance between precision and recall. Consequently, the analysis proceeded using this threshold for subsequent model evaluation.

Following the initial modeling phase, the H_2O AutoML framework was employed to systematically explore a wide range of algorithms and hyperparameter configurations. Among the candidate models generated, a gradient boosting machine (GBM) emerged as the most performing, striking an optimal trade-off between discrimination and calibration metrics-specifically AUC-ROC and log loss. The selected model (ID: GBM_grid_1_AutoML_9_20250326_201624_model_8) achieved an AUC-ROC of 0.942 and a log loss of 0.283, indicating both high classification accuracy and well-calibrated probabilistic outputs.

In the current dataset, the RF model demonstrated superior sensitivity, whereas the H_2O AutoML framework yielded higher specificity. Given the complementary strengths of these models, an ensemble approach combining both was hypothesized to offer enhanced overall performance. Accordingly, further performance analyses were conducted using the ensemble model.

Table 3 presents a comprehensive comparison of the three models-RF (threshold=0.30), H_2O AutoML, and the ensemble model-based on various diagnostic and predictive performance metrics. Among these, the Ensemble Model demonstrated the most balanced and robust performance across nearly all evaluated criteria. In terms of sensitivity, the RF model achieved the highest value (0.80), indicating its superior ability to correctly identify positive cases. However, the H_2O AutoML model excelled in specificity (0.98) and PPV (0.90), highlighting its strength in correctly identifying negative cases and reducing false positives. The ensemble model, which was developed by combining the strengths of the two approaches, achieved a high sensitivity (0.78) close to RF, and a specificity (0.97) comparable to AutoML, reflecting its effectiveness in maintaining a strong trade-off among both metrics.

Furthermore, the ensemble model yielded the highest odds ratio (121.59) and F1 score (0.84) among the three, indicating a superior overall discriminatory power and a well-balanced PR relationship. Its PLR of 27.50 and NLR of 0.23 also suggest a high diagnostic utility. These results support the rationale for using an ensemble strategy, as it effectively leverages the complementary advantages of the individual models.

Figure 2a illustrates the ROC and PR curves of the three ML models evaluated in this study: RF, H_2O AutoML, and the ensemble model. All models demonstrated excellent discriminative performance, with identical AUC-ROC and AUC-PR values of 0.95/0.89, indicating a strong ability to

Table 3. Diagnostic and predictive performance metrics for machine learning models

	Random forest (Th=0.30)	H_2O automated ML	Ensemble model	Validation set
Sensitivity	0.80 (0.71-0.87)	0.63 (0.52-0.72)	0.78 (0.68-0.85)	0.78 (0.68-0.85)
Specificity	0.94 (0.91-0.96)	0.98 (0.95-0.99)	0.97 (0.95-0.99)	0.93 (0.90-0.96)
Positive predictive value	0.81 (0.73-0.89)	0.90 (0.83-0.97)	0.90 (0.83-0.97)	0.85 (0.78-0.92)
Negative predictive value	0.94 (0.90-0.96)	0.89 (0.86-0.93)	0.93 (0.90-0.969	0.89 (0.86-0.92)
Positive likelihood ratio	13.31 (8.30-21.33)	29.44 (13.13-66.00)	27.50 (13.77-54.92)	10.73 (5.64-18.95)
Negative likelihood ratio	0.21 (0.14-0.32)	0.38 (0.29-0.50)	0.23 (0.15-0.33)	0.24 (0.15-0.33)
Odds ratio	63.22 (31.02-128.80)	77.12 (30.93-192.26)	121.59 (51.42-287.46)	44.8 (21.97-93.01)
F1 score	0.81 (0.73-0.87)	0.74 (0.66-0.81)	0.84 (0.77-0.89)	0.81 (0.74-0.86)
Matthews correlation coefficient	0.745 (0.660-0.822)	0.694 (0.603-0.775)	0.790 (0.714-0.861)	0.721 (0.669-0.787)

Th: Threshold, ML: Machine learning

distinguish between positive and negative cases. Despite the equal AUC values, the slight variations in curve shapes across models reflect differences in threshold behavior and confidence calibration.

Figure 2b provides the confusion matrices for the respective models, further highlighting their classification behaviors.

The RF model (threshold=0.30) achieved a relatively higher sensitivity, correctly identifying 73 out of 91 actual positive cases, albeit at the expense of slightly more false positives (17 cases). In contrast, the H_2O AutoML model prioritized specificity, yielding only 6 false positives while missing true positives (34 false negatives). The ensemble model balanced these trade-offs effectively, reducing false negatives compared to AutoML (20 cases) while maintaining a high specificity (274 true negatives).

These findings confirm that although overall discriminative capacity was similar across models (as reflected in the AUC metrics), the ensemble model provided the most favorable balance between sensitivity and specificity-an important consideration in scenarios requiring both reliable detection and minimization of false alarms.

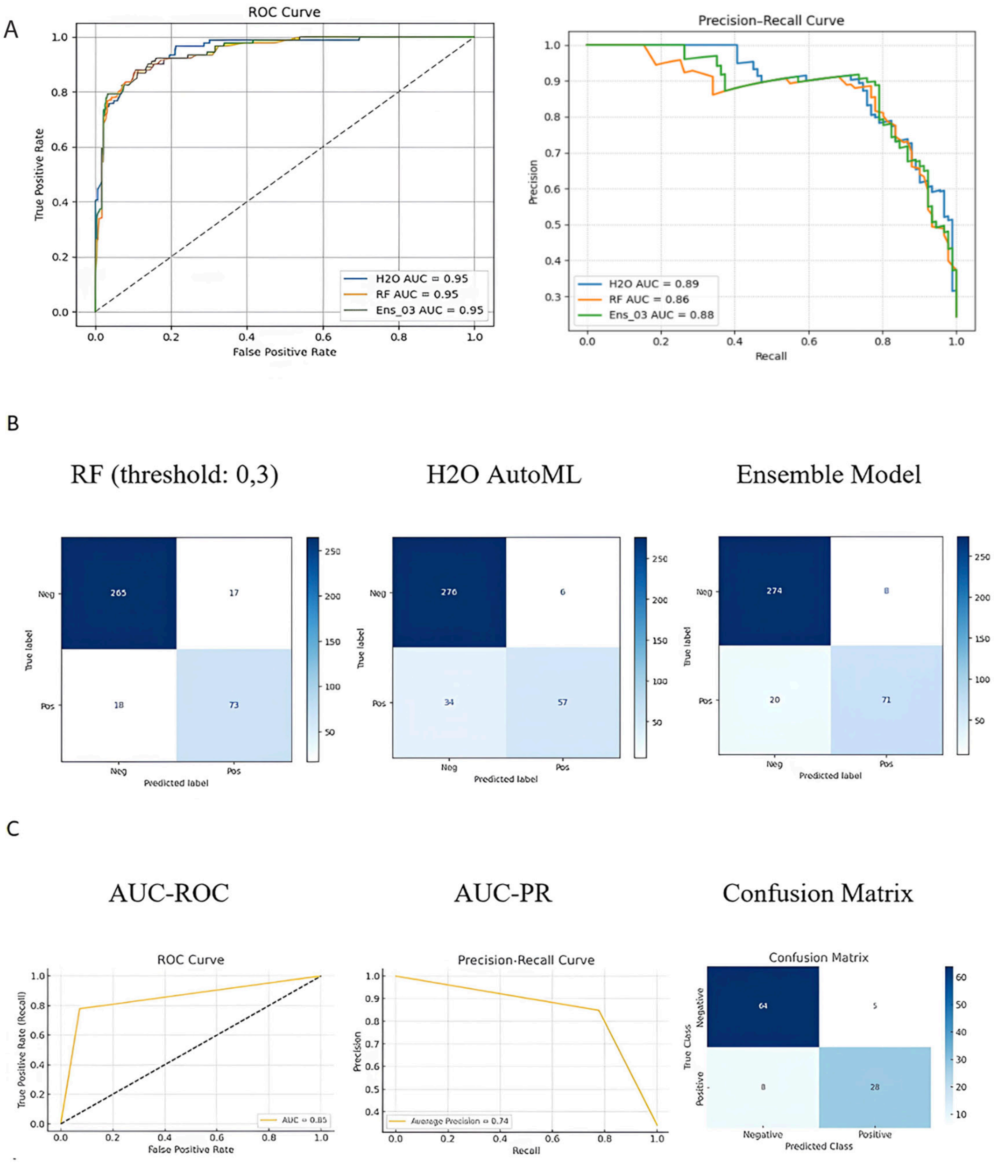

Figure 2. Model performance outputs. (A) AUC-ROC and AUC-PR plots for machine learning models. (B) Confusion matrixes for machine learning models. (C) Performance metrics of validation set

ROC: Receiver operating characteristic, AUC: Area under the curve, RF: Random forest, AutoML: Automated machine learning, PR: Precision-recall

Validation Results of the Models

In accordance with the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) recommendations, external validation was conducted to ensure the generalizability and robustness of our models. The performance outcomes of the validation set are summarized in Table 3 and Figure 2c.

The ensemble model developed in this study was further evaluated on an independent validation set to assess its generalizability beyond the original test data, in accordance with IFCC guidelines that recommend external validation for diagnostic algorithms. The model demonstrated strong classification performance, achieving a sensitivity of 0.78 and a specificity of 0.93, indicating a balanced ability to detect both positive and negative cases. The area under the ROC curve (AUC-ROC) was 0.85, while the area under the PR curve (AUC-PR) reached 0.74, reflecting solid discriminative power even in the presence of class imbalance. Additional metrics such as a PPV of 0.85, a NPV of 0.89, a PLR of 10.73, and an odds ratio of 44.8, further emphasize the clinical relevance of the model's predictions. The F1 score of 0.81 and Matthews correlation coefficient (MCC) of 0.721 confirm the model's robustness and diagnostic accuracy. These results are consistent with the findings obtained from the test set, further supporting the model's reliability across different data sources.

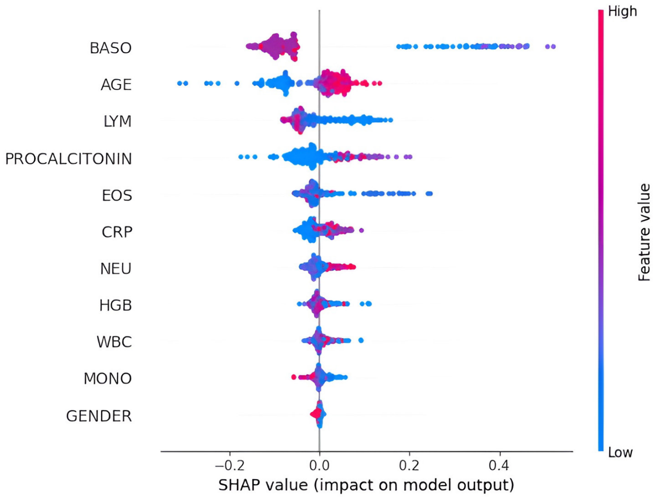

Figure 3. SHAP Summery plot of feature contributions

BASO: Basophil count, LYM: Lymphocyte, EOS: Eosinophil, CRP: C-reactive protein, NEU: Neutrophil, HGB: Hemoglobin, WBC: White blood cell, MONO: Monocyte, SHAP: SHapley Additive exPlanations

Interpretability and Threshold-Based Diagnostic Performance of Key Variables

To improve the interpretability of the model, SHAP analysis was applied to evaluate the contribution of individual features to model predictions. As illustrated in Figure 3, the most impactful variable was "BASO", followed by "AGE", "LYM", and "PROCALCITONIN". Although "BASO" ranked highest, its SHAP value distribution was narrow and centered near zero, indicating frequent but limited directional impact. In contrast, AGE and "PROCALCITONIN" displayed broader SHAP distributions, suggesting stronger influence on model output when elevated. Features such as GENDER, MONO, and WBC ranked lower in importance, showing minimal effect.

Complementing the SHAP results, ROC-based threshold analysis (Table 4) revealed that age (threshold=46.0) provided the highest sensitivity (0.92), with moderate specificity (0.55). PCT (threshold=0.26) showed a sensitivity of 0.78 and specificity of 0.68, indicating balanced diagnostic value. In contrast, basophil demonstrated high

Table 4. Receiver operating characteristic-derived diagnostic thresholds of selected variables

Feature	Threshold	Sensitivity	Specificity
Basophil	0.4	0.03	0.98
Gender	2.0	0.0	1.0
C-reactive protein	99.6	0.69	0.75
Eosinophil	4.5	0.0	1.0
Hemoglobin	6.7	1.0	0.03
Lymphocyte	60.0	0.0	1.0
Monocyte	5.0	0.01	1.0
Neutrophil	6.9	0.65	0.5
Procalcitonin	0.26	0.78	0.68
White blood cell	22.6	0.09	0.93
Age	46.0	0.92	0.55

specificity (0.98) but low sensitivity (0.03), while gender, eosinophil, lymphocyte , and monocyte achieved perfect or near-perfect specificity but negligible sensitivity.

DISCUSSION

The demographic characteristics and baseline laboratory findings of our cohort provide critical context for interpreting model behavior and clinical performance. Our median age of 46 years, male predominance, and blood culture positivity rate of 24.3% are generally consistent with prior studies on similar hospital populations.[1-3] This rate is slightly higher than in some multicenter analyses that report rates ranging from 6.6% to 12%, likely due to different inclusion criteria or local epidemiology.[14-16]

Although the mean age in the validation set was significantly higher than in the training and test sets ($p<0.001$), the model demonstrated robust performance across multiple metrics. The validation results, including high sensitivity (0.78), specificity (0.93), PPV (0.85), and an F1 score of 0.81, indicate effective generalizability without signs of overfitting. Moreover, the positive and NLRs (10.73 and 0.24, respectively) and a strong odds ratio (44.8) support the model's diagnostic strength. Nonetheless, the notable age discrepancy suggests a potential distributional shift, which could impact external validity. Therefore, monitoring model performance across different age groups in future applications is recommended to ensure consistent generalizability.

Notably, serum PCT and CRP levels were significantly elevated among culture-positive patients, which aligns with previous findings. Jeong et al.[17] reported median PCT and CRP levels of 3.2 ng/mL and 132 mg/dL, respectively, in patients with bacteremia, significantly higher than in non-bacteremia groups (0.4 ng/mL and 82.2 mg/dL). Nasimfar et al.[18] similarly observed that septic children had markedly elevated PCT (3.42 ng/mL) and CRP (55.18 mg/L) levels compared to controls. These results support the early diagnostic potential of these biomarkers. Liaudat et al[19] further demonstrated that PCT outperformed CRP and white blood cell count in predicting culture positivity using principal component analysis.

The ensemble ML model developed in our study achieved excellent predictive metrics (AUC-ROC: 0.95, F1 score: 0.84, MCC: 0.79), outperforming most previously reported models. In contrast, other ML approaches using CatBoost or RF have typically yielded AUCs between 0.75 and 0.85.[6,7,20,21] Crucially, our model maintained high performance in external validation (AUC: 0.85; sensitivity: 0.78; specificity: 0.93), which strengthens its generalizability.

Interpretability remains a cornerstone of clinical ML implementation. Our approach integrated SHAP analysis for global feature importance with ROC-derived thresholds for clinical usability. Age and PCT stood out as dual anchors of model strength-ranking high in SHAP impact and exhibiting clear diagnostic cut-offs. This is consistent with findings by Galli et al. [22], who confirmed that PCT provides greater specificity and sensitivity than CRP in critically ill patients, and by Morgan et al.[23], who demonstrated that PCT-guided strategies improve antibiotic stewardship in febrile neutropenia.

BASO was the top-ranked SHAP feature, yet exhibited poor clinical sensitivity. This highlights the distinction between algorithmic influence and practical diagnostic utility, an important concept in applied ML.[5,13] Features such as eosinophils and sex had high specificity but low sensitivity, indicating value in reducing false positives.

The combined use of SHAP analysis and ROC-derived thresholds offers a powerful dual approach for interpreting model behavior in both algorithmic and clinical domains. While BASO emerged as the top-ranking feature in SHAP importance, its limited diagnostic utility, particularly due to low sensitivity, suggests it plays a secondary role in actual decision-making. Conversely, age and PCT were notable not only for their strong SHAP contributions but also for yielding clinically meaningful threshold values, reinforcing their role as primary diagnostic indicators within the model.

Interestingly, some variables, such as gender and eosinophils, demonstrated high specificity yet poor sensitivity. This suggests they are better suited for ruling out false positives rather than identifying true disease states, underscoring that statistical importance does not always equate to clinical utility. These findings illustrate the nuanced roles that features may play: a variable may be statistically dominant in shaping predictions (via SHAP) but lack practical impact at the bedside (via threshold behavior), or vice versa.[4,24]

Clinically, the implementation of our model could have a significant impact. ML tools have been shown to reduce unnecessary blood cultures and antibiotic use while maintaining diagnostic accuracy. For instance, Boerman et al.[25] and Martin et al.[20] report, up to 60% reduction in unnecessary cultures using ML models in ED and PICU settings. Similar benefits were observed in studies using real-time prediction tools based on vital signs or lab panels.[26]

Nonetheless, our study has limitations. The retrospective single-center design introduces potential bias, though external validation provides partial mitigation. The exclusion of clinical signs, comorbidities, and vital data-known to enhance predictive performance in other models-limits our model's clinical depth.[9,27] In addition, the model does not distinguish true bacteremia from contamination, an issue commonly encountered in blood culture interpretation.[17,28]

Despite these limitations, our model offers a promising path forward. It uses only routine laboratory values and demographic features to deliver high performance with interpretable logic. Prospective multicenter studies and real-world deployment in clinical decision support systems (CDSS) are warranted to evaluate the true impact on diagnostic stewardship.

CONCLUSION

Our findings demonstrate that an interpretable ensemble ML model, leveraging routine hematologic and inflammatory parameters, can accurately predict blood culture positivity. By integrating SHAP-based model interpretability with clinically meaningful thresholds, the model provides both algorithmic transparency and bedside usability. Key predictors such as age and PCT consistently contributed to diagnostic performance across statistical and clinical domains. Given its external validity and reliance on readily available data, this model has the potential to be deployed in CDSS to improve diagnostic stewardship and reduce unnecessary blood cultures. Future prospective, multicenter validation studies are warranted to confirm these benefits and facilitate clinical implementation.

Ethics

Ethics Committee Approval: Ethical approval was obtained from the University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital Ethics Committee before initiating the study (decision number: 2025/03-27, date 10.04.2025).

Informed Consent: Retrospective study.

Acknowledgments

The authors would like to express their sincere gratitude to Prof. Dr. Ayfer Çolak for her invaluable academic guidance throughout the study. We also thank Prof. Dr. Savaş Yakan, Chief Physician of the hospital, for his administrative support and facilitation of the research process.

Footnotes

Authorship Contributions

Surgical and Medical Practices: F.D., Concept: F.D., Design: F.D., A.D., Data Collection or Processing: F.D., M.A., Analysis or Interpretation: F.D., M.A., Literature Search: F.D., M.A., A.D., Writing: F.D., A.D.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Mahmoud E, Al Dhoayan M, Bosaeed M, Al Johani S, Arabi YM. Developing machine-learning prediction algorithm for bacteremia in admitted patients. Infect Drug Resist. 2021;14:757-65.
2. Shapiro NI, Wolfe RE, Wright SB, Moore R, Bates DW. Who needs a blood culture? A prospectively derived and validated prediction rule. J Emerg Med. 2008;35:255-64.
3. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23:64.
4. Vijayakumar S, Nair SN, S AC, et al. AI Enhanced explainable early prediction of blood culture positivity in neutropenic patients using clinical and hematologic parameters. Comput Biol Med. 2025;189:109979.
5. Clark M. Prediction of clinical risks by analysis of preclinical and clinical adverse events. J Biomed Inform. 2015;54:167-73.
6. Campagner A, Agnello L, Carobene A, et al. Complete blood count and monocyte distribution width-based machine learning algorithms for sepsis detection: multicentric development and external validation study. J Med Internet Res. 2025;27:e55492.
7. Cheng M, Zhao X, Ding X, Gao J, Xiong S, Ren Y. Prediction of blood culture outcome using hybrid neural network model based on electronic health records. BMC Med Inform Decis Mak. 2020;20:121.
8. Zhang J, Liu W, Xiao W, Liu Y, Hua T, Yang M. Machine learning-derived blood culture classification with both predictive and prognostic values in the intensive care unit: a retrospective cohort study. Intensive Crit Care Nurs. 2024;80:103549.
9. Buchan K, Filannino M, Uzuner Ö. Automatic prediction of coronary artery disease from clinical narratives. J Biomed Inform. 2017;72:23-32.
10. van der Geest PJ, Mohseni M, Linssen J, Duran S, de Jonge R, Groeneveld ABJ. The intensive care infection score - a novel marker for the prediction of infection and its severity. Crit Care. 2016;20:180.
11. JetBrains Team. Learn IDE features. February 23, 2025. Available from: https://www.jetbrains.com/help/pycharm/feature-trainer.html
12. Fryda T, LeDell E, Gill N, Aiello S. H_2O: R Interface for the "H_2O" Scalable Machine Learning Platform. 2024. Available from: https://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/index.html
13. Topcu Dİ, Bayraktar N. Searching for the urine osmolality surrogate: an automated machine learning approach. Clin Chem Lab Med. 2022;60:1911-20.
14. Zhang XJ, Zhao H, Zhang D, et al. Blood cell count-derived inflammation indices as predictors of the osteoporotic risk of postmenopausal women. Eur Rev Med Pharmacol Sci. 2024;28:2207-16.
15. Chang YH, Hsiao CT, Chang YC, et al. Machine learning of cell population data, complete blood count, and differential count parameters for early prediction of bacteremia among adult patients with suspected bacterial infections and blood culture sampling in emergency departments. J Microbiol Immunol Infect. 2023;56:782-92.

16. Parente DM, Cunha CB, Mylonakis E, Timbrook TT. The clinical utility of methicillin-resistant staphylococcus aureus (MRSA) nasal screening to rule out MRSA pneumonia: a diagnostic meta-analysis with antimicrobial stewardship implications. Clin Infect Dis. 2018;67:1-7.

17. Jeong S, Park Y, Cho Y, Kim HS. Diagnostic utilities of procalcitonin and C-reactive protein for the prediction of bacteremia determined by blood culture. Clinica Chim Acta. 2012;413:1731-6.

18. Nasimfar A, Sadeghi E, Karamyyar M, Manesh L. Comparison of serum procalcitonin level with erythrocytes sedimentation rate, C-reactive protein, white blood cell count, and blood culture in the diagnosis of bacterial infections in patients hospitalized in Motahhari hospital of Urmia (2016). J Adv Pharm Technol Res. 2018;9:147-52.

19. Liaudat S, Dayer E, Praz G, Bille J, Troillet N. Usefulness of procalcitonin serum level for the diagnosis of bacteremia. Eur J Clin Microbiol Infect Dis. 2001;20:524-7.

20. Martin B, Payan M, Greer C, et al. 411: Machine learning to improve blood culture stewardship in the picu. Crit Care Med. 2025;53(1).

21. Mooney C, Eogan M, Ní Áinle F, et al. Predicting bacteraemia in maternity patients using full blood count parameters: a supervised machine learning algorithm approach. Int J Lab Hematol. 2021;43:609-15.

22. Galli F, Bindo F, Motos A, et al. Procalcitonin and C-reactive protein to rule out early bacterial coinfection in COVID-19 critically ill patients. Intensive Care Med. 2023;49:934-45.

23. Morgan JE, Phillips B. PAnTher Cub: procalcitonin-guided antibiotic therapy for febrile neutropenia in children and young people with cancer - a single-arm pilot study. BMJ Paediatr Open. 2022;6:001339.

24. Versbraegen N, Fouché A, Nachtegael C, et al. Using game theory and decision decomposition to effectively discern and characterise bi-locus diseases. Artif Intell Med. 2019;99:101690.

25. Boerman AW, Schinkel M, Meijerink L, et al. Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study. BMJ Open. 2022;12:053332.

26. Zoabi Y, Kehat O, Lahav D, Weiss-Meilik A, Adler A, Shomron N. Predicting bloodstream infection outcome using machine learning. Sci Rep. 2021;11:20101.

27. Baghdadi JD, Brook RH, Uslan DZ, et al. Association of a care bundle for early sepsis management with mortality among patients with hospital-onset or community-onset sepsis. JAMA Intern Med. 2020;180:707-16.

28. Kristóf K, Pongrácz J. Interpretation of blood microbiology results - function of the clinical microbiologist. EJIFCC. 2016;27:147-55.