Journal of Microbiology, Immunology and Infection 56 (2023) 782—792

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jmii.com

Original Article

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

Yu-Hsin Chang [a,b], Chiung-Tzu Hsiao [c], Yu-Chang Chang [c], Hsin-Yu Lai [c], Hsiu-Hsien Lin [c], Chien-Chih Chen [d], Lin-Chen Hsu [e], Shih-Yun Wu [f], Hong-Mo Shih [a,b,g,**], Po-Ren Hsueh [b,c,h,*], Der-Yang Cho [i,***]

[a] *Department of Emergency Medicine, China Medical University Hospital, Taichung, Taiwan*
[b] *School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan*
[c] *Department of Laboratory Medicine, China Medical University Hospital, Taichung, Taiwan*
[d] *Department of Laboratory, Wei-Gong Memorial Hospital, Miaoli City, Taiwan*
[e] *Department of Laboratory, An-Nan Hospital, China Medical University, Tainan, Taiwan*
[f] *School of Medicine, Chang Gung University, Taoyuan, Taiwan*
[g] *Department of Public Health, China Medical University, Taichung, Taiwan*
[h] *Division of Infectious Diseases, Department of Internal Medicine, China Medical University Hospital, China Medical University, Taichung, Taiwan*
[i] *Department of Neurosurgery, China Medical University Hospital, Taichung, Taiwan*

Received 30 April 2023; received in revised form 6 May 2023; accepted 6 May 2023
Available online 23 May 2023

* Corresponding author. Departments of Laboratory Medicine and Internal Medicine, China Medical University Hospital, No. 2, Yude Road, North District, Taichung, 40447, Taiwan.

** Corresponding author. Department of Emergency Medicine, China Medical University Hospital, No. 2, Yude Road, North District, Taichung, 40447, Taiwan.

*** Corresponding author. Department of Neurosurgery, China Medical University Hospital, No. 2, Yude Rd., North Dist., Taichung, 404332, Taiwan.

E-mail addresses: homoe042002@hotmail.com (H.-M. Shih), hsporen@gmail.com (P.-R. Hsueh), d5057@mail.cmuh.org.tw (D.-Y. Cho).

https://doi.org/10.1016/j.jmii.2023.05.001

KEYWORDS
Bacteremia early prediction;
Machine learning;
Cell population data;
Complete blood count;
Cell differential count

Abstract *Background:* Bacteremia is a life-threatening complication of infectious diseases. Bacteremia can be predicted using machine learning (ML) models, but these models have not utilized cell population data (CPD).

Methods: The derivation cohort from emergency department (ED) of China Medical University Hospital (CMUH) was used to develop the model and was prospectively validated in the same hospital. External validation was performed using cohorts from ED of Wei-Gong Memorial Hospital (WMH) and Tainan Municipal An-Nan Hospital (ANH). Adult patients who underwent complete blood count (CBC), differential count (DC), and blood culture tests were enrolled in the present study. The ML model was developed using CBC, DC, and CPD to predict bacteremia from positive blood cultures obtained within 4 h before or after the acquisition of CBC/DC blood samples.

Results: This study included 20,636 patients from CMUH, 664 from WMH, and 1622 patients from ANH. Another 3143 patients were included in the prospective validation cohort of CMUH. The CatBoost model achieved an area under the receiver operating characteristic curve of 0.844 in the derivation cross-validation, 0.812 in the prospective validation, 0.844 in the WMH external validation, and 0.847 in the ANH external validation. The most valuable predictors of bacteremia in the CatBoost model were the mean conductivity of lymphocytes, nucleated red blood cell count, mean conductivity of monocytes, and neutrophil-to-lymphocyte ratio.

Conclusions: ML model that incorporated CBC, DC, and CPD showed excellent performance in predicting bacteremia among adult patients with suspected bacterial infections and blood culture sampling in emergency departments.

Introduction

Bacteremia is a life-threatening condition resulting from the presence of viable bacteria in the bloodstream.[1—3] A previous study has indicated that patients in the emergency department (ED) with bacteremia have a higher 30-day mortality rate than those with negative blood cultures.[1] The higher mortality rate associated with bacteremia is frequently linked to delayed or inappropriate use of anti-infective treatment.[2,3] Currently, blood culture remains the gold standard test for identifying the causative agent of bacteremia, and studies have reported that the time-to-positivity ranges from approximately 16 to 25 h on average.[4—7] In addition, one-third to over half of positive blood cultures may be the result of contamination, which occurs when bacteria that are not in the bloodstream are introduced into the culture bottle during blood sampling.[8—10] Due to the ambiguity regarding the clinical relevance of potential contaminants, patients may need to stay in the hospital for longer periods, receive unnecessary antibiotic treatment, and experience additional laboratory testing,[11,12] resulting in significantly higher costs in pharmacy charges, laboratory charges, and indirect costs.[13] Thus, the accurate and timely detection of bacteremia is critical in clinical practice.

Over the past few decades, new-generation haematology analysers have measured quantitative information on the morphological and functional characteristics of leukocytes to generate cell population data (CPD) for leukocyte differential count (DC).[14] Compared to conventional CBC/DC test which only provides the numbers of different blood cells, CPD can offer a more in-depth view of blood cells regarding the cellular volume, granularity, complexity, transparency, composition, and membrane surface of the cells.[15] CPD has been already applied in infectious disease, including early diagnosis of COVID-19, screening tools for viral infection in children and discriminating the etiologies of fever.[16—18] Several studies have indicated that there is a significant change in some parameters of CPD in response to bacterial infection, but they mainly utilized statistical analysis methods to assess the differences.[19—23] However, the CPD report contains many numbers that may require further interpretation before clinical utility. Therefore, machine learning (ML) models that can process large-scale information may play a critical role.

ML has been widely applied in medicine, and recent studies have shown that ML has the potential to detect bacteremia with greater efficacy than traditional scales, such as the quick Sequential Organ Failure Assessment (qSOFA) score and systemic inflammatory response syndrome (SIRS).[24,25] However, whether ML models trained with CPD can effectively predict bacteremia and whether these models can be applied to the population in the ED are still unclear. The objective of the present study was to establish ML models for the early prediction of bacteremia among adult patients in the ED using CPD, complete blood count (CBC), and DC. The prevalence of bacteremia and distribution of data may vary across different hospitals. Therefore, in order to evaluate the generalizability of the model, we included two different hospitals for external validation. In addition, the present study used explainable AI to depict how the ML model makes decisions.

Materials and methods

Study design and participants

The present study collected data from three hospitals. For derivation and prospective validation, data were obtained from China Medical University Hospital (CMUH). CMUH is a 1700-bed, urban, academic, tertiary care hospital in Taichung city, which is in central Taiwan. There are approximately 150,000 to 160,000 ED visits annually at CMUH. For external validation, data were acquired from the ED in Wei-Gong Memorial Hospital (WMH) and Tainan Municipal An-Nan Hospital (ANH). With a total of 872 beds, WMH is located in Miaoli, Taiwan, and it serves as an academic, regional hospital with approximately 55,000 ED visits annually. ANH, an academic regional hospital in Tainan, Southern Taiwan, has a capacity of 925 beds and serves an annual volume of 50,000 in the ED. Approval for conducting the present study was granted by the Institutional Ethics Committee of China Medical University Hospital (Reference No. CMUH112-REC3-043). As the present study involved minimal risk to the subjects, informed consent was waived.

For patients with suspected infectious disease, emergency physicians generally perform CBC, DC and blood cultures. The number of blood culture sets is mainly based on the clinical judgment of physicians. However, most physicians usually follow the routine practices of their respective hospitals, which leads to variations in the distribution of the number of blood culture sets among different hospitals. The CBC was analysed on a Beckman Coulter DxH 900 (Beckman Coulter, Miami, Florida, USA), and the collection of blood culture followed the Clinical and Laboratory Standards Institute (CLSI) guidelines.[26] The detection and identification of blood samples were performed using the BACTEC™ FX system (Becton Dickinson Microbiology Systems, Sparks, MD, USA) with a positive recovery rate between 13.31% and 17.68%.[7]

For the derivation cohort, we retrospectively enrolled adult patients (age 20 years or older) who had a CBC test in the ED at CMUH during the following periods: May 1, 2021 to July 31, 2022 and March 1, 2022 to December 31, 2022. The exclusion criteria were as follows: 1) patients without blood culture test; 2) patients without information about WBC or any DC; 3) patients with a WBC count of zero, and (4) if sampling time of all blood culture was more than 4 h from sampling times of CBC/DC. External validation was performed with the same inclusion and exclusion criteria during the following periods: December 1, 2022, to January 31, 2023, in WMH and October 1, 2022 to January 31, 2023, in ANH. Prospective validation was conducted at CMUH from February 15, 2023 to April 15, 2023.

Data source

The clinical information was obtained from the electronic medical record of CMUH, including demographic information, such as age, gender, and laboratory tests involving blood culture, CPD, and CBC/DC. The CBC test included the following measurements: white blood cell (WBC) count, haemoglobin, haematocrit, red blood cell (RBC) count, platelet count, platelet distribution width (PDW), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC), and monocyte distribution width (MDW). DC included the percentage of lymphocytes (LY), monocytes (MO), segmented neutrophils (NE), eosinophils (EO), and basophils (BA). In addition, sampling time was also recorded in each examination.

When performing DC analysis, the CPD was obtained routinely from a Beckman Coulter DxH900 analyser which is a quantitative, multiparameter, automated haematology analyser.[27] Volume, conductivity, and scattergram (VCS) technology is employed to evaluate the WBC differential, nucleated RBC count, and immature granulocytes. VCS technology involves measuring impedance to determine cell volume, analysing the internal composition of the cell and nucleus-to-cytoplasm ratio using radiofrequency current, and assessing cellular granularity through five light-scatter measurements, which together constitute the CPD data. The detailed items of the CPD are listed in Supplementary Table S1.

Pre-processing

Because some values derived from the blood analysis were not presented in the laboratory report in some hospitals, we calculated these values directly from the raw data, including absolute neutrophil count, PLR, and NLR. The missing values of the DC was replaced with zeros, including the percentage of band cells, lymphocytes, monocytes, segmented neutrophils, eosinophils, and basophils. For other blood and CPD parameters, the median value of the training set was used to impute the missing values. The imputation method and missing values of parameters in each cohort are displayed in Supplementary Table S2.

Training pipeline

The training process is shown in Fig. 1. For internal validation, we divided the CMUH cohort from 2021 to 2022 into 80% for training and 20% for testing. Within the training set, we performed 5-fold cross-validation to assess performance. To prevent the algorithms from being biased towards higher values, all features, which were continuous, were scaled before training the models. A standard scaler was utilized in the present study, which made the mean of the data zero and the standard deviation one. Subsequently, because there was an imbalance in the distribution of positive and negative labels, the resampling technique of synthetic minority oversampling technique (SMOTE)-edited nearest neighbour (ENN) was applied. SMOTE generates synthetic samples of the minority class by interpolating between the feature vectors of minority class instances, which increases the representation of the minority class.[28] In contrast, ENN removes noisy samples from the majority class by examining the neighbours of each sample and removing those that are misclassified.[29] These adjustments were made to ensure that the number of patients between the two groups was more balanced to prevent AI algorithms from being biased in favour of the majority class in the imbalanced dataset. For external and prospective

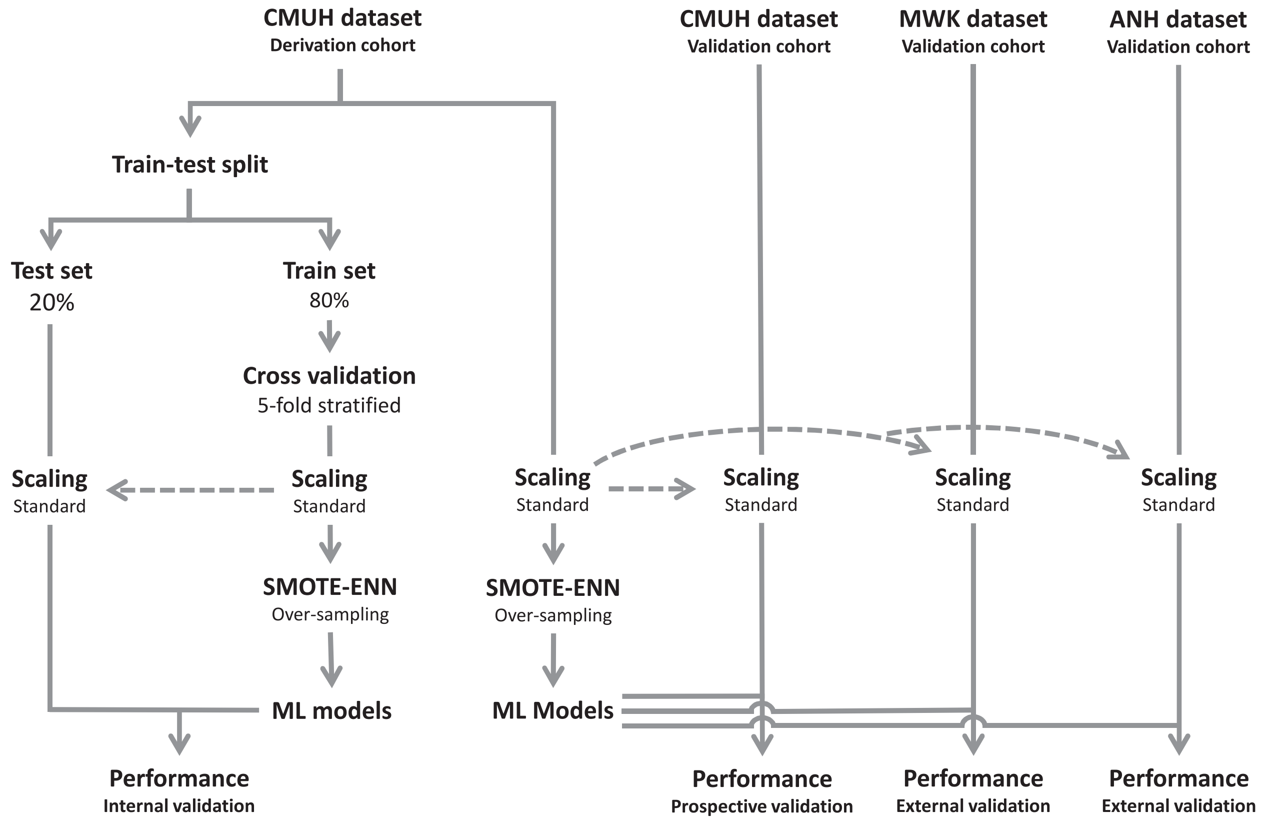

Figure 1. Training pipeline among different cohorts. CMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital; SMOTE, synthetic minority oversampling technique; ENN, edited nearest neighbour; ML, machine learning.

validation, we used the entire CMUH dataset from 2021 to 2022 for model training and recorded the performance on each cohort separately.

In the present study, we developed binary classifiers using five most common used machine learning models as follows: extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), categorical boosting (CatBoost), random forest (RF), and logistic regression (LR).[30—34] These models are not only extensively integrated into machine learning models for predicting medical issues but are also open-source resources that are convenient to implement and perform various tasks.

Feature selection and hyperparameter tuning were not executed in the present study.

Performance evaluation

The main objective of the present study was to establish ML models through CBC/DC and CPD to provide early prediction of bacteremia from positive blood cultures obtained within 4 h before or after the acquisition of CBC/DC blood samples. We labelled the sample as positive if there were microorganisms in the blood culture, and we labelled the sample as negative if there was contamination or no growth of microorganisms. The definition of contamination of blood cultures was according to the CLSI guidelines.[26]

To evaluate the performance of the present models, we utilized several metrics, including the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), precision (positive predictive value), recall (sensitivity), F1 score, negative predictive value, and specificity. Additionally, we trained the model using different combinations of data types to assess the impact on performance. DeLong's test was applied to compare the AUROC between individual models.[35,36] To interpret the output of the models, we applied the SHapley Additive exPlanations (SHAP) method using the SHAP python package (version 0.41.0). The results are presented as bee swarm plots, in which each dot corresponds to an individual data point in the model.[37]

Experimental environment

The pre-processing of data was conducted using MATLAB (version R2021a). To construct and train the ML models, Python (version 3.8.10) on the Google Colab platform was applied.[38]

Results

Participant characteristics

The present study initially identified 76,426 CBC/DC samples in the derivation cohort of CMUH, but after exclusion, the final dataset used for developing the ML model

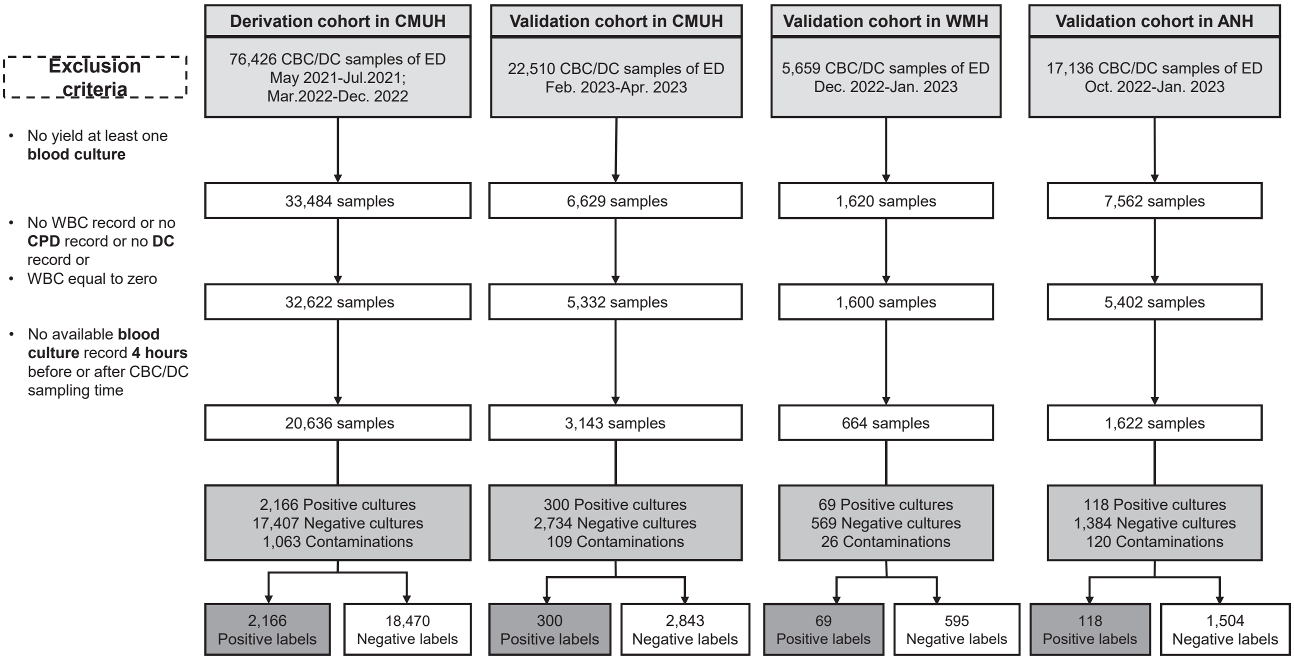

Figure 2. Flow chart of patient recruitment in different datasets. CMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital; CBC, complete blood count; CPD, cell population data; DC, differential count.

consisted of 2166 cases with positive labels and 18,470 cases with negative labels. The prospective validation cohort of CMUH included 3143 cases with both CBC/DC and blood culture samples after exclusion, comprising 300 positive labels and 2843 negative labels. Additionally, external validation was conducted with cases of 69 positive and 595 negative labels from WMH as well as cases of 118 positive and 1504 negative labels from ANH. The inclusion and exclusion process for each cohort is illustrated in Fig. 2.

The mean age of patients in the derivation and internal validation cohort was 64 ± 29 years, while it was slightly younger in the prospective validation cohort (62 ± 34 years) and older in the WMH cohort (69 ± 37 years) and ANH cohort (66 ± 31 years). Each cohort had nearly an equal proportion of females, making up almost half of the included subjects. The average WBC count was 9.3 × 10^9/L, and the average proportion of neutrophils was 79.5% in the derivation cohort, with similar CBC/DC distribution results observed in each cohort. Table 1 presents the patient demographics and CBC/DC examination results for the derivation cohort, prospective validation cohort, and external validation cohorts from WMH and ANH.

In the derivation cohort of 20,636 cases, 60.8%, 39.0%, and 0.2% had one set, two sets, and three sets of blood cultures, respectively, within 4 h before and after the CBC/DC exam. The prospective validation cohort from CMUH had a similar distribution of blood culture sets. However, a higher proportion of one set of blood cultures (87.5%) was observed in the WMH cohort, while the ANH cohort had a higher proportion of two sets of blood culture samples (71.9%). In the derivation cohort, 10.5% had a positive blood culture result, 84.3% had a negative result, and 5.2% were suspected to have a contaminated blood culture result. The proportion of positive blood culture results was slightly lower in the CMUH prospective cohort and the WMH cohort (9.5% and 10.3%, respectively), while the lowest proportion of positive blood culture results (7.3%) was noted in the ANH cohort. *Escherichia coli* was the most commonly identified pathogen in all four cohorts. In the derivation cohort from the CMUH, and validation cohort from WMH and ANH, *Klebsiella pneumoniae* accounted for the second-highest proportion of positive blood culture results, followed by *Staphylococcus aureus*. In the prospective validation cohorts from CMUH, *S. aureus* accounted for 16.0% of all positive blood culture results, followed by *K. pneumoniae* and *Proteus mirabilis* (9.7% and 3.0%, respectively). Table 2 provides more details on the distribution of blood culture samples and identified pathogens.

Supplementary Table S1 shows the missing values of all parameters among each cohort. There were some missing values among the DC of the cohort in CMUH. Among all hospitals, band cells showed large missing values because they were reported as blank if not detected during blood analysis, and after pre-processing, they presented as missing values. Except for band cells, MDW accounted for the highest number of missing values.

Performance

Table 3 outlines the performance of the ML models for predicting bacteremia using CBC/DC and CPD data through cross-validation. The CatBoost model achieved a slightly better AUROC (0.844 ± 0.002) than the other models, with the LGB model coming in second (0.842 ± 0.001). The CatBoost model also had a higher AUPRC (0.447 ± 0.003) than the LGB model (0.435 ± 0.008). The F1-score for the CatBoost model was 0.445, with a specificity of 0.826 and

Table 1 Basic characteristics of patients in different cohorts.

Datase	CMUH		WMH	ANH
	Derivation cohort	Validation cohort	Validation cohort	Validation cohort
Case number	20,636	3143	664	1622
Age, median (IQR)	64.0 (29.0)	62.0 (34.0)	69.0 (37.0)	66.0 (31.0)
Female	10,305 (50.0)	1528 (48.6)	332 (48.5)	805 (49.6)
Complete Blood Count and DC, Median (IQR)				
WBC ($10^3/\mu$L)	9.3 (6.2)	9.5 (6.2)	9.4 (7.0)	9.4 (6.1)
NE (%)	79.5 (17.2)	80.2 (16.7)	79.6 (13.7)	76.9 (17.9)
MO (%)	7.0 (4.6)	7.0 (4.6)	7.0 (5.9)	7.0 (5.8)
LY (%)	10.5 (12.1)	10.1 (11.6)	10.0 (10.0)	12.5 (13.4)
BA (%)	0.4 (0.4)	0.4 (0.4)	0.3 (0.5)	0.4 (0.3)
EO (%)	0.5 (1.3)	0.5 (1.3)	0.3 (1.0)	0.7 (1.4)
Band cells (%)	5.8 (9.4)	5.1 (13.4)	6.0 (14.0)	NA
Hgb (g/dL)	12.3 (3.4)	12.5 (3.4)	12.4 (3.7)	12.5 (3.1)
RBC ($10^6/\mu$L)	4.15 (1.15)	4.2 (1.2)	4.2 (1.2)	4.3 (1.1)
NRBC (%)	0 (0.1)	1.6 (1.5)	0.0 (0.1)	0.1 (0.1)
Plt ($10^3/\mu$L)	222.0 (124.0)	223.0 (119.0)	213.0 (121.0)	222.0 (115.0)
NLR	7.5 (10.6)	7.9 (10.8)	7.9 (9.0)	6.2 (8.5)
PLR	20.5 (27.3)	21.0 (27.9)	20.8 (23.0)	17.7 (21.7)
MDW	21.2 (5.7)	21.4 (5.7)	NA	20.5 (5.0)
PDW (fL)	16.8 (0.8)	16.8 (0.8)	16.9 (0.9)	16.7 (0.9)
MCV (fL)	88.6 (7.7)	88.6 (7.3)	87.8 (9.7)	88.2 (7.3)
MCH (pg)	30.1 (3)	30.2 (2.8)	29.9 (4.1)	30.1 (2.9)
MCHC (g/dL)	33.8 (1.2)	34.0 (1.2)	33.9 (1.5)	34.0 (1.4)
Hct (%)	36.4 (9.7)	36.8 (9.8)	36.7 (9.7)	36.9 (8.5)

CMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital; DC, differential count; IQR, interquartile range; WBC, white blood cell; NE, neutrophil; MO, monocyte; LY, lymphocyte; BA, basophil; EO, eosinophil; Hgb, haemoglobin; RBC, red blood cell; NRBC, nucleated red blood cell; Plt, platelet; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; MDW, monocyte distribution width; PDW, platelet distribution width; MCV, mean corpuscular volume; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular haemoglobin concentration; Hct, haematocrit; IQR, interquartile range.

Table 2 Blood culture numbers, results, and bacteria identified in each cohort.

Dataset	CMUH		WMH	ANH
	Derivation cohort	Validation cohort	Validation cohort	Validation cohort
Case number	20,636	3143	664	1622
Blood culture number within 4 h to CBC/DC sampling time, No (%)				
One	12,543 (60.8)	2000 (63.6)	581 (87.5)	454 (28.0)
Two	8049 (39.0)	1133 (36.0)	78 (11.7)	1166 (71.9)
Three	44 (0.2)	10 (0.3)	5 (0.8)	2 (0.1)
bacteremia, No. (%)				
Yes[a]	2166 (10.5)	300 (9.5)	69 (10.3)	118 (7.3)
No[a]	17,407 (84.3)	2734 (87.0)	569 (85.7)	1384 (85.3)
Contamination	1063 (5.2)	109 (3.5)	26 (3.9)	120 (7.4)
Bacteria identified, No. (%)				
E. coli	901 (41.6)	120 (40.0)	32 (46.4)	56 (47.5)
K. pneumoniae	328 (15.1)	29 (9.7)	14 (20.3)	21 (17.8)
S. aureus	244 (11.3)	48 (16.0)	7 (10.1)	8 (6.8)
P. aeruginosa	75 (3.5)	9 (3.0)	1 (1.4)	1 (0.8)
P. mirabilis	59 (2.7)	10 (3.3)	2 (2.9)	4 (3.4)
S. enteritidis	56 (2.6)	5 (1.7)	2 (2.9)	1 (0.8)
E. faecium	39 (1.8)	8 (2.7)	0 (0.0)	1 (0.8)
A. baumanni	20 (0.9)	3 (1.0)	0 (0.0)	1 (0.8)

[a] Cases with contamination were not includedCMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital; No., number; CBC, complete blood count; DC, differential count.

Table 3 Model performance of cross-validation in internal validation.

Models	AUROC	AUPRC	F1-score	Sensitivity	Specificity	PPV	NPV
CatBoost	0.844	0.447	0.445	0.715	0.826	0.323	0.962
LGBM	0.842	0.435	0.435	0.710	0.820	0.313	0.961
XGB	0.839	0.437	0.439	0.696	0.829	0.321	0.959
LR	0.838	0.391	0.323	0.882	0.586	0.198	0.977
RF	0.834	0.391	0.391	0.776	0.746	0.262	0.966

CatBoost, categorical boosting; LGBM, light gradient boosting machine; XGBoost, extreme gradient boosting; RF random forest classifier; LR logistic regression; PPV positive predictive value; NPV negative predictive value; AUROC area under receiver—operator curve; AUPRC area under precision-recall curve; CMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital.

Table 4 Model performance of Catboost and LGBM in the derivation, prospective internal validation, and external validation cohorts.

Cohort	CatBoost		LGBM	
	AUROC	AUPRC	AUROC	AUPRC
CMUH-derivation	0.844	0.447	0.842	0.435
CMUH-validation	0.812	0.419	0.820	0.409
WMH	0.844	0.363	0.837	0.367
ANH	0.847	0.426	0.857	0.437

CatBoost, categorical boosting; LGBM, light gradient boosting machine; AUROC area under receiver-operator curve; AUPRC area under precision-recall curve; CMUH, China Medical University Hospital; WMH, Wei-Gong Memorial Hospital; ANH, Tainan Municipal An-Nan Hospital.

an NPV of 0.962. Fig. 3 shows the ROC curves and precision-recall curves for all developed ML models, including the CatBoost, LGB, XGB, LR, and RF models.

To validate the ML model performance, the CatBoost model and LGB model, which outperformed the other models, were selected. In the CMUH prospective validation cohort, the CatBoost model showed an AUROC of 0.812 and an AUPRC of 0.419, while the LGB model had an AUROC of 0.82 and an AUPRC of 0.409. In the WMH external validation cohort, the CatBoost model had an AUROC of 0.844 with an AUPRC of 0.363, while the LGB model showed slightly lower AUROC and AUPRC. The ANH external validation cohort yielded an AUROC of 0.847 with an AUPRC of 0.426 for the CatBoost model, and the LGB model demonstrated slightly higher AUROC (0.857) and AUPRC (0.437) values (see Table 4).

Feature importance

The analysis of feature importance through the SHAP value in CatBoost and LGBM is shown in Fig. 4, and only the top 15 features are displayed. A positive SHAP value indicates a higher likelihood of bacteremia, while a negative value suggests a lower likelihood. In the CatBoost classifier (Fig. 4A), the five most important features were mean conductivity of lymphocyte (MN_C_LY), NRBC, mean conductivity of monocyte (MN_C_MO), NLR, and standard

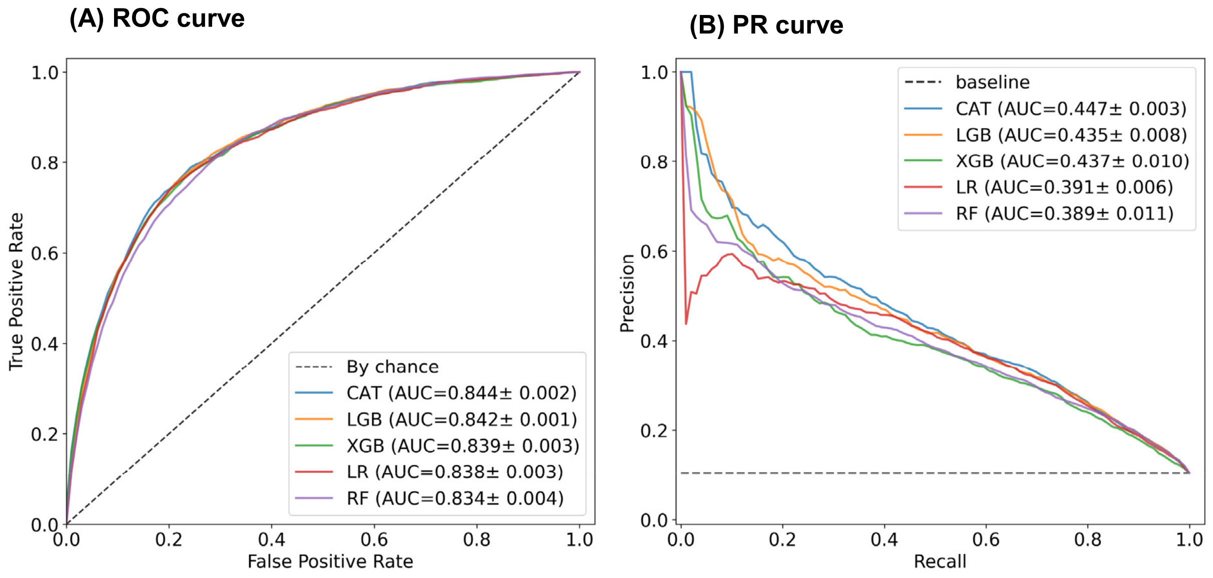

Figure 3. Receiver operating characteristic (ROC) curves and precision-recall (PR) curves of different models. CAT, categorical boosting; LGB, light gradient boosting; XGB, extreme gradient boosting; RF random forest classifier; LR logistic regression.

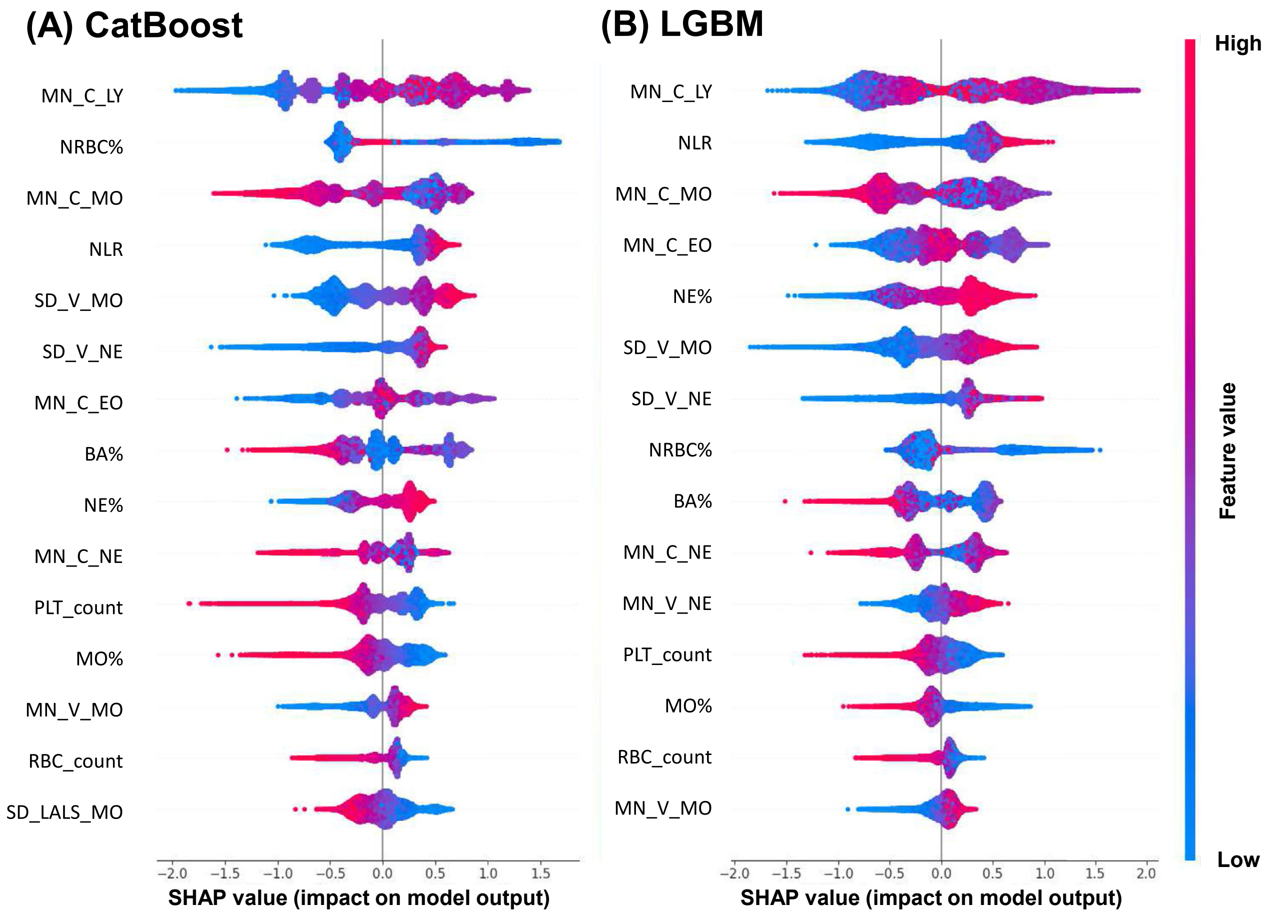

Figure 4. Feature importance derived from the SHAP values in the internal validation. There were 15 features arranged based on the SHAP values across all samples between the (A) CatBoost classifier and (B) LGBM classifier. Each feature is represented by a row, in which the SHAP value is displayed on the horizontal axis and a single point represents each data sample. Features with higher values are visualized in red, while those with lower values are shown in blue. The length of the feature bar indicates the magnitude and direction of the feature's impact on the model's output, in which a long bar signifies a large impact and a short bar signifies a small impact. SHAP, Shapley additive explanations; CatBoost, categorical boosting; LGBM, light gradient boosting machine; MN, mean; %, percentage; SD, standard deviation; C, conductivity; V, volume; LALS, low-angle light scatter; LY, lymphocyte; MO, monocyte; NE, neutrophil; EO, eosinophil; BA, basophil; NRBC, nucleated red blood cell; PLT, platelet; RBC, red blood cell.

Table 5 DeLong test to compare AUROC values of different inputs in the CatBoost classifier for internal validation.

Input type	AUROC (SD)	p value[a]	
CPD and CBC/DC	0.844 (0.002)	reference	
CPD	0.836 (0.002)	0.15	reference
CBC/DC	0.810 (0.003)	<0.001	<0.001

[a] p value was calculated to compare the AUROC value of the reference with other models from different input.
SD, standard deviation; CatBoost, categorical boosting; AUROC, area under receiver-operator curve; Std, standard deviation; CPD, cell population data; CBC, complete blood count; DC, differential count.

deviation of monocyte volume (SD_V_MO). In the LGBM classifier (Fig. 4B), the five most important features were MN_C_LY, NLR, MN_C_MO, mean conductivity of eosinophil (MN_C_EO), and percentage of NE (NE%). Of note, the CPD parameters accounted for more than half of the top ten important features in both models. By analysing the top five SHAP values of the CatBoost classifier, it was found that MN_C_LY, NLR, and SD_V_MO were positively correlated with bacteremia. Conversely, lower values of NRBC and MN_C_MO were associated with a higher risk of bacteremia.

Performance of different inputs for the Catboost classifier

In the present study, there were two types of features, namely, CPD and CBC/DC. We trained CatBoost using the following three input sources: 1) all features (CPD and CBC/DC), 2) CPD alone, or 3) CBC/DC alone. The results are presented in Table 5. Compared to the model trained with CBC/DC alone, the model trained with either all features or CPD alone showed a significantly higher AUROC ($p < 0.001$) through the DeLong test. The model trained with all features had the tendency to achieve a higher AUROC of 0.844

than models trained with CPD alone, but the difference was not statistically significant ($p = 0.15$).

Discussion

The present study indicated that by using the data derived from blood analysis (CPD and CBD/DC), ML models are capable of early prediction of bacteremia in blood culture obtained within 4 h before or after the acquisition of CBC/DC blood samples. Notably, this is the first study to establish ML models that combine CPD and CBC/DC. In internal validation, the CatBoost and LGBM classifiers obtained higher AUROC and AUPRC values, and they also performed well in external and prospective validation. Moreover, we discovered that incorporating CPD with CBC/DC significantly enhanced the predictive performance of the model.

Prior research has utilized multiple sources of data input for model training or establishment of predictive rules, showing good performance in the identification of bacteremia. These prior studies utilized models trained on medical notes, vital signs, and laboratory data in the ED.[24,25,39] However, medical notes rely heavily on subjective judgements and the quality of recording by the assessors. The laboratory data include various blood and biochemical parameters, which are dependent on the order of the physicians and consequently result in a considerable number of missing values. In addition, vital signs can be influenced by current treatments. For instance, blood pressure and heart rate can be affected by vasopressors or antihypertensive medications, and oxygen administration can affect blood oxygen saturation and respiratory rate. To address these issues and accurately reflect a patient's current condition, it is necessary to increase the scope of recorded data to include all relevant external factors; however, this can result in more missing values and lead to increased difficulty in implementation of the prediction model.

Lien et al. discovered that CBC/DC can be used to train ML models to identify bacteria present in the blood culture ordered on the same day, and the addition of C-reactive protein (CRP) enhances the effectiveness of the model; they obtained an AUROC of 0.806 when the ML model was trained with CBC/DC count alone, which was similar to the present results (Table 5).[40] In the present study, we further applied the DeLong test to discriminate the significance between models with different inputs and revealed that CPD strengthened the predictive ability.

Previously, there were several studies statistically analysing the difference of CPD among normal control and patients with bacteremia, and indicated several CPD parameters were significantly increased in the bacteremia.[15,19,22] They also determined the cut-off values of each parameter, with some of them showed excellent performance in detecting bacteremia. However, these studies consisted of smaller dataset and did not apply their optimum cut-off value to prospective or external cohorts to validate its efficacy. On the contrary, by combining ML with CPD and CBC/DC, we recruited participants from three hospitals and applied our ML models among both external and prospective cohort to validate the performance.

In the CatBoost and LGBM classifiers, CPD composed over half of the top ten important features, including the mean conductivity and standard deviation of volume among different cells. Mean conductivity is measured through radiofrequency method, and it stands for the nucleus-to-cytoplasm ratio of the cells. Two of the top three most important features, the mean conductivity of lymphocyte and monocyte (MN_C_LY and MN_C_MO), have been reported to be related to bacterial infections and have also been found to be correlated with sepsis.[15,22,41,42] SD_V_MO, which is equivalent to monocyte distribution width (MDW), has been extensively investigated in recent years in the context of sepsis.[43,44] Cells of the innate immune system, including monocytes and polymorphonuclear leukocytes, serve as the first line of defence against infections.[45] Monocytes play a critical role in the immune response from the earliest stages and act as the primary defence against infections.[46] Moreover, monocytes are responsible for various immune functions, including antigen presentation, cytokine production, and phagocytosis. Monocytes are a heterogeneous group of cells with distinct phenotypes, nuclear morphologies, sizes, and functions. In infectious situations, this heterogeneity becomes more pronounced, resulting in variations in monocyte morphology due to functional changes in certain monocyte subsets.[47]

Because CPD are parameters routinely generated during analysis of CBC/DC without the requirement of additional blood samples, the application of CPD in the ML model offers some advantages. Obtaining a CPD report takes the same amount of time as obtaining a CBC/DC report, allowing the ML model to be executed earlier in the ED and enabling clinicians to initiate treatment or further testing in a timely manner. Second, having a single data source simplifies the integration of the predictive model into clinical practice because fewer data sources need to be connected. Additionally, since the data is device-generated rather than manually input, missing values are rare, which could have positive effect on performance. In addition, clinicians typically request a CBC test for most patients with suspected bacterial infection as part of the management in the ED. Therefore, using CPD as an additional marker for bacteremia is a cost-effective method that would not increase the cost or burden on the health care system.

However, the present study also had several limitations. First, only patients suspected of having an infectious disease and who had both CBC and blood culture exams were included, which prevented the evaluation of the utility of the ML model on other populations. Although a more precise definition of patients with infectious diseases can provide a more specific population for model development and may increase the model's predictive performance. However, limiting the definition of the patients enrolled in the present study may result in a decrease in the generalizability of our model. The purpose of this study to build a ML model that can offer predictive information on bacteremia for patients with varying degrees of suspicion of infectious disease. This can allow patients to benefit from the predictive model even without a clear infection source or specific symptoms (such as fever). In addition, limiting the inclusion criteria will lead to a decrease in the size of dataset, and increase the risk of overfitting among AI models. Second, the selection of study cases was based on the decisions of clinical physicians to perform blood culture exams, which may introduce some selection bias. However, physicians were unaware of the patients' inclusion in the study cohort, reducing possible selection bias. Third, host factors, such as diabetes and immunocompromised status, were not considered in the development of the model, but the aim of the model was to predict bacteremia without clinical information. Despite this, the ML model performed reliably. Finally, the study population was limited to a single country, and further validation through global cooperation is necessary to establish the model's generalizability.

In conclusion, the ML model that incorporated CBC, DC, and CPD showed excellent performance in predicting bacteremia among adult patients with suspected bacterial infections and blood culture sampling in emergency departments. This ML model will be further developed using CPD for all hospitalized patients to distinguish between Gram-positive and Gram-negative bacteria causing bloodstream infection, facilitating early detection of sepsis and predicting clinical prognosis.

Ethics statement

The present study was approved by the Institutional Review Board of China Medical University Hospital (referencing number, CMUH112-REC3-043).

Funding

None.

Declaration of competing interest

All authors have no conflicts of interest to declare.

References

1. Lindvig KP, Nielsen SL, Henriksen DP, Jensen TG, Kolmos HJ, Pedersen C, et al. Mortality and prognostic factors of patients who have blood cultures performed in the emergency department: a cohort study. *Eur J Emerg Med* 2016;**23**:166—72.
2. Opota O, Croxatto A, Prod'hom G, Greub G. Blood culture-based diagnosis of bacteraemia: state of the art. *Clin Microbiol Infect* 2015;**21**:313—22.
3. Bearman GM, Wenzel RP. Bacteremias: a leading cause of death. *Arch Med Res* 2005;**36**:646—59.
4. Lambregts MMC, Bernards AT, van der Beek MT, Visser LG, de Boer MG. Time to positivity of blood cultures supports early re-evaluation of empiric broad-spectrum antimicrobial therapy. *PLoS One* 2019;**14**:e0208819.
5. Ning Y, Hu R, Yao G, Bo S. Time to positivity of blood culture and its prognostic value in bloodstream infection. *Eur J Clin Microbiol Infect Dis* 2016;**35**:619—24.
6. Jacobs MR, Mazzulli T, Hazen KC, Good CE, Abdelhamed AM, Lo P, et al. Multicenter clinical evaluation of BacT/Alert Virtuo blood culture system. *J Clin Microbiol* 2017;**55**:2413—21.
7. Lin HH, Liu YF, Tien N, Ho CM, Hsu LN, Lu JJ. Evaluation of the blood volume effect on the diagnosis of bacteremia in automated blood culture systems. *J Microbiol Immunol Infect* 2013; **46**:48—52.
8. Gonsalves WI, Cornish N, Moore M, Chen A, Varman M. Effects of volume and site of blood draw on blood culture results. *J Clin Microbiol* 2009;**47**:3482—5.
9. Banerjee R, Teng CB, Cunningham SA, Ihde SM, Steckelberg JM, Moriarty JP, et al. Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing. *Clin Infect Dis* 2015; **61**:1071—80.
10. Salluzzo R, Reilly K. The rational ordering of blood cultures in the emergency department. *Qual Assur Util Rev* 1991;**6**:28—31.
11. Bates DW, Goldman L, Lee TH. Contaminant blood cultures and resource utilization. The true consequences of false-positive results. *JAMA* 1991;**265**:365—9.
12. van der Heijden YF, Miller G, Wright PW, Shepherd BE, Daniels TL, Talbot TR. Clinical impact of blood cultures contaminated with coagulase-negative staphylococci at an academic medical center. *Infect Control Hosp Epidemiol* 2011;**32**: 623—5.
13. Dempsey C, Skoglund E, Muldrew KL, Garey KW. Economic health care costs of blood culture contamination: a systematic review. *Am J Infect Control* 2019;**47**:963—7.
14. Urrechaga E. Reviewing the value of leukocytes cell population data (CPD) in the management of sepsis. *Ann Transl Med* 2020; **8**:953.
15. Park DH, Park K, Park J, Park HH, Chae H, Lim J, et al. Screening of sepsis using leukocyte cell population data from the Coulter automatic blood cell analyzer DxH800. *Int J Lab Hematol* 2011;**33**:391—9.
16. Vasse M, Ballester MC, Ayaka D, Sukhachev D, Delcominette F, Habarou F, et al. Interest of the cellular population data analysis as an aid in the early diagnosis of SARS-CoV-2 infection. *Int J Lab Hematol* 2021;**43**:116—22.
17. Urrechaga E, Bóveda O, Aguirre U, García S, Pulido E. Neutrophil cell population data biomarkers for acute bacterial infection. *J Pathol Infect Dis* 2018;**1**:1—7.
18. Jung YJ, Kim JH, Park YJ, Kahng J, Lee H, Lee KY, et al. Evaluation of cell population data on the UniCel DxH 800 Coulter Cellular Analysis system as a screening for viral infection in children. *Int J Lab Hematol* 2012;**34**:283—9.
19. Suresh PK, Minal J, Rao PS, Ballal K, Sridevi HB, Padyana M. Volume conductivity and scatter parameters as an indicator of acute bacterial infections by the automated haematology analyser. *J Clin Diagn Res* 2016;**10**:EC01—3.
20. Chaves F, Tierno B, Xu D. Quantitative determination of neutrophil vcs parameters by the coulter automated hematology analyzer. *Am J Clin Pathol* 2005;**124**:440—4.
21. DoĞAn Ö, ÇaliŞKan E, AltinÖZ Aytar A. Investigation of neutrophil volume, conductivity, and light-scattering parameters for early diagnosis of bacterial infections. *Val Health Sci* 2022;**12**(3):468—73.
22. Shekhar R, Pai S, Srinivasan VK, Srinivas V, Adhikary R, Bhavana MV. Alterations in leucocyte cell population data in bacteraemia: a study from a tertiary care hospital in India. *Int J Lab Hematol* 2021;**43**:e1—4.
23. Park SH, Park CJ, Lee BR, Nam KS, Kim MJ, Han MY, et al. Sepsis affects most routine and cell population data (CPD) obtained using the Sysmex XN-2000 blood cell analyzer: neutrophil-related CPD NE-SFL and NE-WY provide useful information for detecting sepsis. *Int J Lab Hematol* 2015;**37**:190—8.
24. Schinkel M, Boerman AW, Bennis FC, Minderhoud TC, Lie M, Peters-Sengers H, et al. Diagnostic stewardship for blood cultures in the emergency department: a multicenter validation and prospective evaluation of a machine learning prediction tool. *EBioMedicine* 2022;**82**:104176.
25. Choi DH, Hong KJ, Park JH, Shin SD, Ro YS, Song KJ, et al. Prediction of bacteremia at the emergency department during triage and disposition stages using machine learning models. *Am J Emerg Med* 2022;**53**:86—93.
26. Clinical and Laboratory Standards Institute. *Principles and procedures for blood cultures*. Wayne, PA: Approved Guideline. CLSI document M47-A, CLSI; 2007.

27. Serrando Querol M, Nieto-Moragas J, Marull Arnall A, Figueras MD, Jimenez-Romero O. Evaluation of the new beckmann coulter analyzer dxh 900 compared to sysmex xn20: analytical performance and flagging efficiency. *Diagnostics* 2021;**11**.
28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321—57.
29. Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans Syst Man Cybern* 1972;**2**:408—21. SMC-.
30. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*; 2016. p. 785—94.
31. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst* 2018;**31**.
32. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**.
33. Breiman L. Random forests. *Mach Learn* 2001;**45**:5—32.
34. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2013.
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988: 837—45.
36. Sun X, Xu W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;**21**: 1389—93.
37. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;**30**.
38. Bisong E, Google Colaboratory. In: Bisong E, editor. *Building machine learning and deep learning models on Google cloud platform: a comprehensive guide for beginners*. Berkeley, CA: Apress; 2019. p. 59—64.
39. Takeshima T, Yamamoto Y, Noguchi Y, Maki N, Gibo K, Tsugihashi Y, et al. Identifying patients with bacteremia in community-hospital emergency rooms: a retrospective cohort study. *PLoS One* 2016;**11**:e0148078.
40. Lien F, Lin HS, Wu YT, Chiueh TS. Bacteremia detection from complete blood count and differential leukocyte count with machine learning: complementary and competitive with C-reactive protein and procalcitonin tests. *BMC Infect Dis* 2022; **22**:287.
41. Urrechaga E, Boveda O, Aguirre U. Role of leucocytes cell population data in the early detection of sepsis. *J Clin Pathol* 2018;**71**:259—66.
42. Zhang W, Zhang Z, Pan S, Li J, Yang Y, Qi H, et al. The clinical value of hematological neutrophil and monocyte parameters in the diagnosis and identification of sepsis. *Ann Transl Med* 2021; **9**:1680.
43. Polilli E, Sozio F, Frattari A, Persichitti L, Sensi M, Posata R, et al. Comparison of monocyte distribution width (MDW) and procalcitonin for early recognition of sepsis. *PLoS One* 2020;**15**: e0227300.
44. Crouser ED, Parrillo JE, Seymour C, Angus DC, Bicking K, Tejidor L, et al. Improved early detection of sepsis in the ed with a novel monocyte distribution width biomarker. *Chest* 2017;**152**:518—26.
45. Peterson PK, Verhoef J, Schmeling D, Quie PG. Kinetics of phagocytosis and bacterial killing by human polymorphonuclear leukocytes and monocytes. *J Infect Dis* 1977;**136**:502—9.
46. Yona S, Jung S. Monocytes: subsets, origins, fates and functions. *Curr Opin Hematol* 2010;**17**:53—9.
47. Tak T, van Groenendael R, Pickkers P, Koenderman L. Monocyte subsets are differentially lost from the circulation during acute inflammation induced by human experimental endotoxemia. *J Innate Immun* 2017;**9**:464—74.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmii.2023.05.001.