

RESEARCH

Open Access



Deep learning analysis of 12-lead electrocardiograms for bloodstream infection prediction: a multi-center validation study

Po-Huang Chen^{1,2}, Shang-Yang Li^{3,4}, Dung-Jang Tsai^{4,5}, Yan-Ru Wang⁶, Shiue-Wei Lai^{1,2}, Ping-Ying Chang^{1,2}, Jia-Hong Chen^{1,2}, Ming-Shen Dai^{1,2}, Chin-Sheng Lin⁷, Chin Lin^{4,5*} and Cho-Hao Lee^{1,2*}

Abstract

Background Prompt diagnosis of bloodstream infections (BSIs) is critical for antimicrobial stewardship but hindered by blood culture delays of 48 h or more. We developed and validated a deep learning model to predict BSI risk from standard 12-lead electrocardiograms (ECGs).

Methods This study adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines. In this retrospective, multi-center study, we trained a deep learning model to predict clinically significant BSIs (excluding common contaminants) using ECGs from adult emergency department (ED) patients at an academic medical center in Taiwan (April 2012 to March 2022; total $N = 163,826$) and validated it in three independent cohorts: internal ($n = 22,205$), external ($n = 22,709$), and MIMIC-IV ($n = 66,937$). For the MIMIC-IV cohort, BSI was identified using the derived suspicion of infection framework with a positive blood culture flag.

Results The model achieved AUCs of 0.821 (95% CI 0.804–0.838), 0.800 (0.782–0.818), and 0.756 (0.743–0.768) across the three test cohorts, respectively. At a high-sensitivity rule-out threshold the model classified 19–25% of patients as low risk (negative predictive value [NPV] 99.8%); at the Youden point, the sensitivity and specificity were 73.2% and 76.8% in the internal test set, with an NPV of 99.2%. The model showed superior discrimination compared with XGBoost and logistic regression baselines. Decision-curve analysis showed net benefit over heart rate alone and the default treat-all and treat-none strategies across the clinically relevant low-threshold range. Exploratory analyses indicated that the ECG-derived risk signal was elevated across Gram-positive, Gram-negative, and fungal infections; pathogen-specific prediction will require a dedicated model (Supplementary Figure S3). The primary model stratified 90-day mortality risk among infected patients (adjusted HR 3.43 [1.93–6.09] for high- vs. low-risk).

Conclusions Deep learning analysis of standard 12-lead ECGs obtained at ED triage can predict clinically significant BSI with discrimination superior to conventional ECG parameters. The model discriminated best (highest AUC) in

*Correspondence:

Chin Lin
xup6fup0629@gmail.com
Cho-Hao Lee
drleechohao@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

patients with low clinical suspicion ($qSOFA=0$), although its incremental value in this low-prevalence subgroup is modest. Prospective validation is needed before clinical implementation.

Keywords Artificial intelligence, Electrocardiogram, Bloodstream infection, Deep learning, Antimicrobial stewardship, Bacteremia, Risk stratification, Clinical decision support

Background

Bloodstream infection (BSI), defined as the presence of a viable pathogen in blood culture (including both bacteremia and fungemia), is a medical emergency with mortality rates ranging from 13% to 21%, even with modern treatments [1, 2]. The standard diagnostic method, blood culture, takes at least 48 hours to confirm results [3]. Since each hour of delay in appropriate antibiotic treatment raises mortality by about 7.6% in patients with septic shock [4], clinicians must make empiric treatment decisions with incomplete data [5]. This challenge is especially serious in ‘occult’ BSI [6]. Elderly, immunocompromised, or multimorbid patients often present with atypical symptoms, evading tools like $qSOFA$ and common biomarkers, which lack sufficient accuracy for standalone detection of BSI [7, 8]. However, unnecessary antibiotic use drives resistance, side effects, and higher costs [9]. Current risk-stratification approaches for BSI rely on clinical scoring systems (e.g., Shapiro criteria), biomarkers (procalcitonin, C-reactive protein), or combinations thereof. These are used in practice but, individually, none is accurate enough to determine the need for blood cultures or empiric antibiotics without additional clinical assessment [7, 10].

Recent progress in artificial intelligence (AI) has pushed ECG analysis beyond traditional cardiac monitoring [11]. AI-ECG models can now detect non-cardiac conditions including electrolyte disorders, anemia, and reduced ejection fraction by identifying subtle signals that are not apparent to human readers [12–14]. Systemic infections induce complex cardiovascular adaptations that alter cardiac electrophysiology, often before overt clinical symptoms appear, offering a window for early detection [15, 16].

Building on this foundation, we hypothesized that deep learning could detect BSI by analyzing standard 12-lead ECGs. Previous studies have explored machine learning for sepsis prediction using clinical data [17, 18], and Kwon et al. showed the potential of ECG-based deep learning for sepsis screening [19]. However, the evolving definition of sepsis (from SIRS to SOFA criteria [20]) introduces heterogeneity in these models’ endpoints, whereas BSI provides an objective, microbiologically defined target.

Here, we present a deep learning model that analyzes 12-lead ECGs to predict BSI risk at the time of hospital presentation. We conducted a multi-center, retrospective study to train and validate this AI-ECG system across

distinct patient populations, including internal and external validation in two hospitals within the same healthcare system and a geographically separate critical care dataset from the United States. We also examined the model’s interpretability by identifying key ECG features associated with BSI risk. Our aim is to provide clinicians with a rapid, non-invasive screening aid for BSI that supports antibiotic stewardship by reducing unnecessary therapy in low-risk cases.

Methods

Study design and data sources

We conducted a retrospective, multi-center study approved by the Institutional Review Board of Tri-Service General Hospital in Taipei, Taiwan (IRB No. C202105049) with a waiver of informed consent due to the retrospective nature of the study. This study was conducted and reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines for prediction model development and validation [21]. A completed TRIPOD checklist is provided in Supplementary Table S4. Data were collected from two hospital settings: an academic medical center in Neihu District (Hospital A) and a community hospital in Zhongzheng District (Hospital B) spanning April 2012 to March 2022. We included adult emergency department patients (aged > 20 years, the age of legal majority in Taiwan during the study period) who underwent both an ECG and a blood culture during the ED encounter. The ECG used for prediction was the triage ECG; all ECGs were obtained within 1 h of triage (median time from triage to ECG 0.2–0.3 h across cohorts; Supplementary Table S7). The prediction endpoint was a clinically significant BSI confirmed by a blood culture obtained within 1 day (24 h) of that ECG. The index time was defined as the patient’s emergency department triage time.

To assess generalizability beyond Taiwan, we incorporated data from the MIMIC-IV database (2008–2019) [22], drawn from a single US academic centre (Beth Israel Deaconess Medical Center, Boston, Massachusetts). MIMIC-IV is geographically homogeneous and not nationally representative of the United States, but it provides a geographically and demographically distinct, cross-ethnic external test cohort. The MIMIC-IV database was accessed under PhysioNet credentialed data use agreement.

Patient cohorts and data processing

We used five cohorts across two hospitals and MIMIC-IV. From Hospital A, we established a development set (36,988 patients; 64,822 ECGs) and a tuning set (14,987 patients; 26,376 ECGs). For final evaluation, we used three test cohorts: an internal test set from Hospital A (22,205 patients), an external test set from Hospital B (22,709 patients), and the MIMIC-IV test set (66,937 patients). Each test cohort used a single ECG per patient. Patient populations were strictly separated between development and validation phases to prevent data leakage. The development and tuning cohorts incorporated multiple ECGs per patient to capture temporal variation in cardiac signals, while test cohorts used only the first ECG per patient to replicate real-world clinical scenarios.

Standardized 10-second, 12-lead ECGs were acquired using Philips instrumentation (500 Hz). In addition to raw waveforms, we extracted 31 diagnostic patterns (e.g., arrhythmias, conduction disturbances) and 8 quantitative measurements from the system outputs (complete list provided in Supplementary Table S5). Missing data (< 9%) underwent multiple imputation using multivariate imputation by chained equations (MICE) with predictive mean matching (5 imputations). Imputed features were used exclusively for the parallel XGBoost model, not for the deep learning model which processed raw waveforms [23–25].

The dataset was enriched with patient demographics (age, sex, body mass index) and comorbidity profiles from electronic medical records. Comorbidities included diabetes mellitus, hypertension, hyperlipidemia, chronic kidney disease, acute myocardial infarction, stroke, coronary artery disease, heart failure, atrial fibrillation, and chronic obstructive pulmonary disease, identified using ICD-9 and ICD-10 codes [26].

Blood cultures were obtained at the discretion of the treating emergency physician when systemic infection was clinically suspected; the study cohort therefore comprised patients with suspected infection who had both a blood culture and an ECG. To define clinically significant BSI, all positive blood cultures were reviewed based on established clinical guidelines. A clinically significant BSI was defined as the isolation of a recognized pathogen from at least one blood culture bottle. The prediction endpoint was a clinically significant BSI confirmed by a blood culture obtained within 1 day (24 h) of the index ECG; the per-patient prevalence of this endpoint was 2.1% (development), 2.2% (internal), 2.0% (external), and 1.8% (MIMIC-IV). For reference, the proportion of patients with any non-contaminant positive culture in the linked record, irrespective of timing, was approximately 7%. A pre-defined list of organisms commonly considered contaminants, including coagulase-negative staphylococci, *Staphylococcus epidermidis*, *Diphtheroids*,

Bacillus spp., *Lactobacillus* spp., *Cutibacterium acnes*, *Micrococcus* spp., *Corynebacterium* spp., *Propionibacterium acnes*, and *Paenibacillus* spp., were excluded from the positive cohort unless the same organism was isolated from two or more separate blood culture sets drawn within a 48-hour period. Cases of fungemia were included as part of the BSI endpoint. In an exploratory subgroup analysis we examined whether the primary BSI risk score discriminated each pathogen class (Gram-positive, Gram-negative, fungal) from other patients (Supplementary Figure S3).

For the MIMIC-IV cohort, BSI was identified using the derived ‘suspicion of infection’ framework, which identifies clinical episodes where antibiotic orders and blood culture specimens co-occurred within a predefined time window [27]. A positive BSI was defined as any blood culture yielding microbial growth (positive_culture = 1). Unlike the Taiwan cohorts, organism-level adjudication for contaminant exclusion was not performed in the MIMIC-IV dataset, as the derived table does not provide species-level identification. Consequently, pathogen subclassification (Gram-positive, Gram-negative, fungal) was not available for this cohort. This less stringent endpoint definition may include contaminant organisms, potentially reducing the signal-to-noise ratio.

Deep learning model architecture and training

The deep learning model architecture was adapted from our previously validated AI-ECG mortality prediction system, the technical details of which have been described elsewhere [28, 29]. Briefly, the model is a convolutional neural network with an attention mechanism, originally trained on more than 450,000 ECGs with all-cause mortality as the primary label using survival data with censored events. The model processes raw 12-lead ECG waveforms (5,000 data points per lead at 500 Hz, yielding a $12 \times 5,000$ input matrix). During training, random 4,096-sample segments were used for data augmentation; for prediction, overlapping segments with averaged outputs were used to maximize coverage of the 10-second recording. For the present study, the model was retrained with BSI as the prediction target, using a sigmoid activation function for binary classification (BSI probability), in contrast to the survival prediction output used in the original mortality model. Training was performed using the Adam optimizer (batch size = 32, initial learning rate = 0.001) with dynamic learning rate reduction, early stopping, and L2 regularization (10^{-4}). No specific class imbalance handling was applied; the model was trained on the natural BSI prevalence (~ 7%) in the development cohort; this training label was defined at the ECG level (a positive blood culture during the emergency department encounter) and is therefore higher than the approximately 2% per-patient, 24-hour evaluation

endpoint used in the Results. Training was performed on an NVIDIA DGX-1 platform. Training code for the base architecture is publicly available at <https://github.com/lmshepherd/ECGSurvNet>.

Model evaluation and statistical analysis

We evaluated model performance using multiple metrics. Discriminative performance was assessed using the area under the receiver operating characteristic curve (AUC) and precision-recall characteristics (PRAUC). 95% confidence intervals for AUC were calculated using the DeLong method [30]. Clinical applicability was evaluated through sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at the optimal threshold defined by Youden's index (sensitivity + specificity - 1).

To identify subgroups with the greatest clinical benefit, we conducted stratified analyses across patient subgroups defined by age, sex, Charlson Comorbidity Index (CCI; 0 = no comorbidity, 1–3 = mild, 4–6 = moderate, >=7 = severe), and qSOFA score. These analyses identified populations most likely to benefit from AI-ECG BSI screening.

We assessed model calibration by plotting predicted probabilities against observed BSI frequencies using the Hosmer-Lemeshow goodness-of-fit test with decile grouping. Decision-curve analysis estimated the net benefit of the model for the decision of whether to obtain a blood culture and initiate empiric antibiotics. Predicted probabilities for the deep-learning model and for heart rate alone were obtained by logistic recalibration on the development set, and net benefit was compared against heart rate alone and against the conventional treat-all and treat-none reference strategies across the clinically relevant range of threshold probabilities near the observed prevalence (approximately 1–10%).

We assessed prognostic value using Kaplan-Meier analysis, with patients stratified by AI-ECG risk categories. Risk categories were defined based on the model's prediction score using thresholds derived from the tuning set: low-risk (score < 0.827), medium-risk (0.827 to 1.625), and high-risk (score > 1.625). We investigated mortality outcomes both in the overall population and within the BSI subgroup. The independent association between model predictions and outcomes was quantified using multivariable Cox proportional hazard models, adjusting for age and sex, with results expressed as hazard ratios (HR) and 95% confidence intervals. Adjustment was limited to age and sex to assess the model's independent prognostic value with minimal covariate adjustment; a fully adjusted model was not pursued because many clinical variables (e.g., qSOFA, vital signs) are themselves inputs that the ECG model may capture.

To enhance interpretability, we used an XGBoost model to rank the importance of 8 quantitative and 4 diagnostic ECG features (selected from the full set of 39 features) based on information gain. Feature importance was assessed using two-fold cross-validation, with the XGBoost model trained to predict the deep learning model's BSI risk scores (regression objective). R-squared values represent the squared Pearson correlation between predicted and observed scores, averaged across folds. Of note, the XGBoost model serves as a parallel interpretable model rather than a direct explanation of the deep learning model's internal representations. To quantify the incremental value of deep learning over simpler approaches, we compared the deep learning model's discrimination with that of XGBoost (using all 39 ECG features or 8 quantitative features only), logistic regression (heart rate, corrected QT interval, QRS wave axis, and T wave axis), and heart rate alone (Supplementary Table S8). To evaluate the model's rule-out performance independent of disease prevalence, we report positive and negative likelihood ratios and project NPV across a range of assumed BSI prevalences (Supplementary Table S9).

All statistical analyses were performed using R version 4.5.2 with the following packages: pROC for AUC calculation and DeLong confidence intervals, survival for Kaplan-Meier and Cox regression analyses, xgboost for the interpretable model, mice for multiple imputation, ResourceSelection for calibration assessment, and ggplot2 for visualization. The deep learning model was developed using Python with TensorFlow/Keras on an NVIDIA DGX-1 platform.

Results

Patient characteristics

Our study included ECGs from a development set (64,822 ECGs from 36,988 patients) and three distinct test cohorts (Fig. 1). Using one ECG per patient, the BSI prevalence was 2.1% in the development set, 2.2% in the internal test set, 2.0% in the external test set, and 1.8% in the MIMIC-IV cohort (1,191 patients with positive blood cultures). As detailed in Table 1 and Supplementary Tables S1–S3, a clear gradient of clinical severity was observed across the model's risk strata. Patients classified as high-risk were consistently older, had higher qSOFA scores, and a greater burden of comorbidities compared to low-risk patients (all $p < 0.001$).

AI-ECG model performance

The AI-ECG model showed strong discriminative ability across all test sets (Fig. 2; Supplementary Table S6); because discrimination summarised by AUROC is influenced by the high proportion of true negatives at low prevalence, we interpret it alongside the precision-recall,

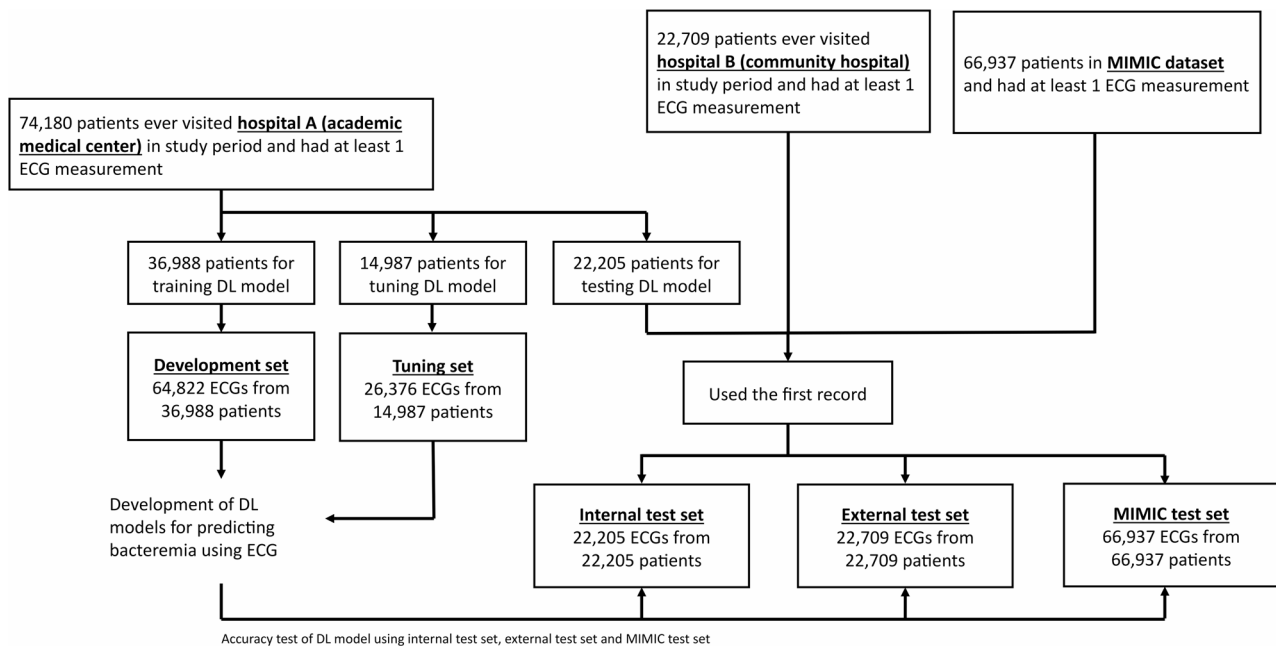


Fig. 1 Study design and dataset partitioning. Schematic illustration of dataset construction and analysis workflow. The study cohort was divided into five non-overlapping patient populations: a development cohort, a tuning cohort, an internal test cohort from an academic medical center, an external test cohort from a community hospital, and the MIMIC-IV test cohort. Patient assignment to mutually exclusive datasets prevented data leakage between model development and validation phases

predictive-value, likelihood-ratio, and decision-curve analyses reported below. The model achieved AUC values of 0.821 (95% CI 0.804–0.838) in the internal test set, 0.800 (0.782–0.818) in the external test set, and 0.756 (0.743–0.768) in the MIMIC-IV test set. Because the intended use is to rule out infection, we evaluated a range of operating points rather than a single Youden-optimal threshold (Supplementary Table S6). At the Youden point (sensitivity 61.0–73.2%), 27–39% of true BSI would be missed, which is inappropriate for a rule-out test. We therefore additionally report a high-sensitivity rule-out threshold: at approximately 98% sensitivity the model classified 19–25% of patients as low risk while missing only 0.3–0.4 true BSI per 1000 ED visits (negative likelihood ratio 0.07–0.10), and at approximately 92% sensitivity it classified 43–49% as low risk while missing 1.6–1.9 per 1000. The corresponding negative predictive values were 99.8% and 99.6–99.7%, respectively. The choice of threshold reflects the acceptable miss rate for a given clinical setting. Compared with baseline models built from the same ECGs, the deep learning model showed superior discrimination on both test sets: an XGBoost model using 39 ECG features achieved an internal AUC of 0.790 and an external AUC of 0.774, whereas the deep learning model achieved 0.821 internally and 0.800 externally. Heart rate alone achieved AUCs of 0.786 and 0.760, and logistic regression using four ECG parameters achieved 0.792 and 0.765, all below the deep learning model and confirming that it extracts predictive

information from the raw waveform beyond conventional ECG parameters (Supplementary Table S8).

Model performance across patient subgroups

Stratified analysis revealed distinct patterns of model performance across patient subgroups (Fig. 3). The model performed best in patients under 60 years of age (AUC > 0.80 across all test sets) and maintained clinically meaningful discrimination in elderly patients. Comorbidity burden analysis showed stronger performance in patients with no or mild comorbidities compared to those with severe comorbidity. Most notably, the model achieved the highest discrimination in patients with qSOFA = 0 (AUC 0.819 [95% CI 0.795–0.843], 0.790 [0.764–0.815], and 0.754 [0.739–0.769] across the internal, external, and MIMIC-IV test sets, respectively), with performance decreasing as qSOFA scores increased. Sex did not significantly affect model performance. In the MIMIC-IV dataset, performance was consistent across racial and ethnic groups: White (AUC 0.732, $n = 42,223$), Black (0.800, $n = 11,890$), Hispanic (0.782, $n = 3,830$), Asian (0.817, $n = 2,805$), and Other (0.772, $n = 6,189$). All ECGs were obtained within 1 h of triage (median 0.2–0.3 h; Supplementary Table S7), so the reported discrimination reflects true point-of-care prediction at presentation.

Table 1 Baseline characteristics across the development, tuning, and test cohorts (internal, external, and MIMIC-IV)

Variable	Development (n = 36,988)	Tuning (n = 14,987)	Internal (n = 22,205)	External (n = 22,709)	MIMIC-IV (n = 66,937)	p
Demographics						
Age, years	58.7 ± 18.9	58.5 ± 18.9	58.8 ± 18.9	60.9 ± 20.3	58.9 ± 19.9	< 0.001
Sex, male	18,888 (51.1)	7,582 (50.6)	11,171 (50.3)	11,432 (50.3)	31,649 (47.3)	< 0.001
qSOFA score	0.2 ± 0.5	0.2 ± 0.5	0.2 ± 0.5	0.2 ± 0.5	0.1 ± 0.4	< 0.001
Emergency Triage Level						
Triage level 1	3,422 (9.3)	1,423 (9.5)	1,957 (8.8)	1,067 (4.7)	9,122 (13.6)	< 0.001
Triage level 2	11,178 (30.2)	4,456 (29.7)	6,683 (30.1)	6,295 (27.7)	34,204 (51.1)	
Triage level 3	21,713 (58.7)	8,798 (58.7)	13,144 (59.2)	14,919 (65.7)	23,453 (35.0)	
Triage level 4–5	675 (1.8)	310 (2.1)	421 (1.9)	428 (1.9)	158 (0.2)	
Vital Signs						
Heart rate, bpm	88.2 ± 22.0	87.9 ± 21.8	88.2 ± 21.7	87.7 ± 21.5	85.7 ± 21.1	< 0.001
Respiratory rate, /min	19.2 ± 5.1	19.2 ± 4.4	19.2 ± 4.7	18.8 ± 4.3	18.0 ± 3.1	< 0.001
SBP, mmHg	137.9 ± 27.5	138.0 ± 27.4	138.0 ± 27.4	139.7 ± 27.4	136.1 ± 25.1	< 0.001
DBP, mmHg	80.1 ± 17.3	80.3 ± 17.2	80.1 ± 17.3	76.0 ± 17.2	77.0 ± 16.0	< 0.001
Temperature, °C	36.7 ± 1.0	36.7 ± 1.0	36.7 ± 1.0	36.6 ± 1.0	36.7 ± 0.6	< 0.001
Blood Culture Results						
Positive	2,535 (6.9)	1,018 (6.8)	1,537 (6.9)	1,654 (7.3)	-	< 0.001
Gram-positive	698 (1.9)	290 (1.9)	444 (2.0)	474 (2.1)	-	
Gram-negative	1,549 (4.2)	609 (4.1)	922 (4.2)	1,016 (4.5)	-	
Polymicrobial	141 (0.4)	57 (0.4)	92 (0.4)	88 (0.4)	-	
Fungal	147 (0.4)	62 (0.4)	79 (0.4)	76 (0.3)	-	
Negative	34,453 (93.1)	13,969 (93.2)	20,668 (93.1)	21,055 (92.7)	-	
Comorbidities (Charlson Comorbidity Index)						
Myocardial infarction	1,403 (3.8)	601 (4.0)	873 (3.9)	925 (4.1)	7,295 (10.9)	< 0.001
Heart failure	3,280 (8.9)	1,361 (9.1)	1,972 (8.9)	2,543 (11.2)	10,189 (15.2)	< 0.001
Peripheral vascular disease	2,058 (5.6)	814 (5.4)	1,201 (5.4)	1,743 (7.7)	5,033 (7.5)	< 0.001
Cerebrovascular disease	6,786 (18.3)	2,791 (18.6)	4,085 (18.4)	5,022 (22.1)	6,514 (9.7)	< 0.001
Dementia	2,691 (7.3)	1,087 (7.3)	1,615 (7.3)	2,579 (11.4)	2,035 (3.0)	< 0.001
Chronic pulmonary disease	5,452 (14.7)	2,198 (14.7)	3,342 (15.1)	5,687 (25.0)	11,945 (17.8)	< 0.001
Rheumatic disease	997 (2.7)	385 (2.6)	586 (2.6)	1,048 (4.6)	1,790 (2.7)	< 0.001
Peptic ulcer disease	6,948 (18.8)	2,788 (18.6)	4,249 (19.1)	5,891 (25.9)	1,614 (2.4)	< 0.001
Mild liver disease	5,261 (14.2)	2,127 (14.2)	3,216 (14.5)	4,773 (21.0)	4,727 (7.1)	< 0.001
DM without complications	7,192 (19.4)	2,899 (19.3)	4,370 (19.7)	5,758 (25.4)	13,716 (20.5)	< 0.001
DM with complications	3,408 (9.2)	1,403 (9.4)	2,036 (9.2)	3,067 (13.5)	4,732 (7.1)	< 0.001
Hemiplegia/paraplegia	1,008 (2.7)	403 (2.7)	635 (2.9)	764 (3.4)	1,629 (2.4)	< 0.001
Renal disease	3,957 (10.7)	1,585 (10.6)	2,400 (10.8)	3,101 (13.7)	9,031 (13.5)	< 0.001
Malignancy	4,431 (12.0)	1,762 (11.8)	2,693 (12.1)	3,223 (14.2)	6,823 (10.2)	< 0.001
Moderate/severe liver disease	507 (1.4)	205 (1.4)	327 (1.5)	300 (1.3)	1,518 (2.3)	< 0.001
Metastatic tumor	1,592 (4.3)	603 (4.0)	946 (4.3)	971 (4.3)	2,899 (4.3)	0.574
AIDS	18 (0.0)	7 (0.0)	18 (0.1)	35 (0.2)	540 (0.8)	< 0.001

Data are presented as mean ± SD or n (%). SBP = systolic blood pressure; DBP = diastolic blood pressure; DM = diabetes mellitus; AIDS = acquired immunodeficiency syndrome; BSI = bloodstream infection. Emergency triage level follows Taiwan ED 5-level system (1 = most urgent). Blood culture organism-level data were not available for the MIMIC-IV cohort (indicated by '-'); BSI in MIMIC-IV was defined using the derived 'suspicion of infection' framework with positive blood culture flag (see Methods). MIMIC-IV BSI prevalence: 1,191 (1.8%). Per-patient BSI prevalence after deduplication: internal 2.2%, external 2.0%, MIMIC-IV 1.8%. p values from chi-square test for categorical and ANOVA for continuous variables

Prognostic value in mortality prediction

Kaplan-Meier survival analysis showed that the AI-ECG model effectively stratified mortality risk (Fig. 4; Supplementary Figure S2). In the all-cause mortality analysis, high-risk patients had significantly higher 90-day mortality. In the internal test set, the high-risk group's mortality was 13.0% compared to 0.9% for the low-risk group

(adjusted HR 10.41, 95% CI 8.13–13.33). This risk stratification was consistent across all external test cohorts. The model maintained significant prognostic value among patients with confirmed BSI: high-risk patients had higher 90-day mortality than low-risk patients (internal set: 10.5% vs. 3.6%; adjusted HR 3.43, 95% CI 1.93–6.09). The medium-risk tier followed the expected

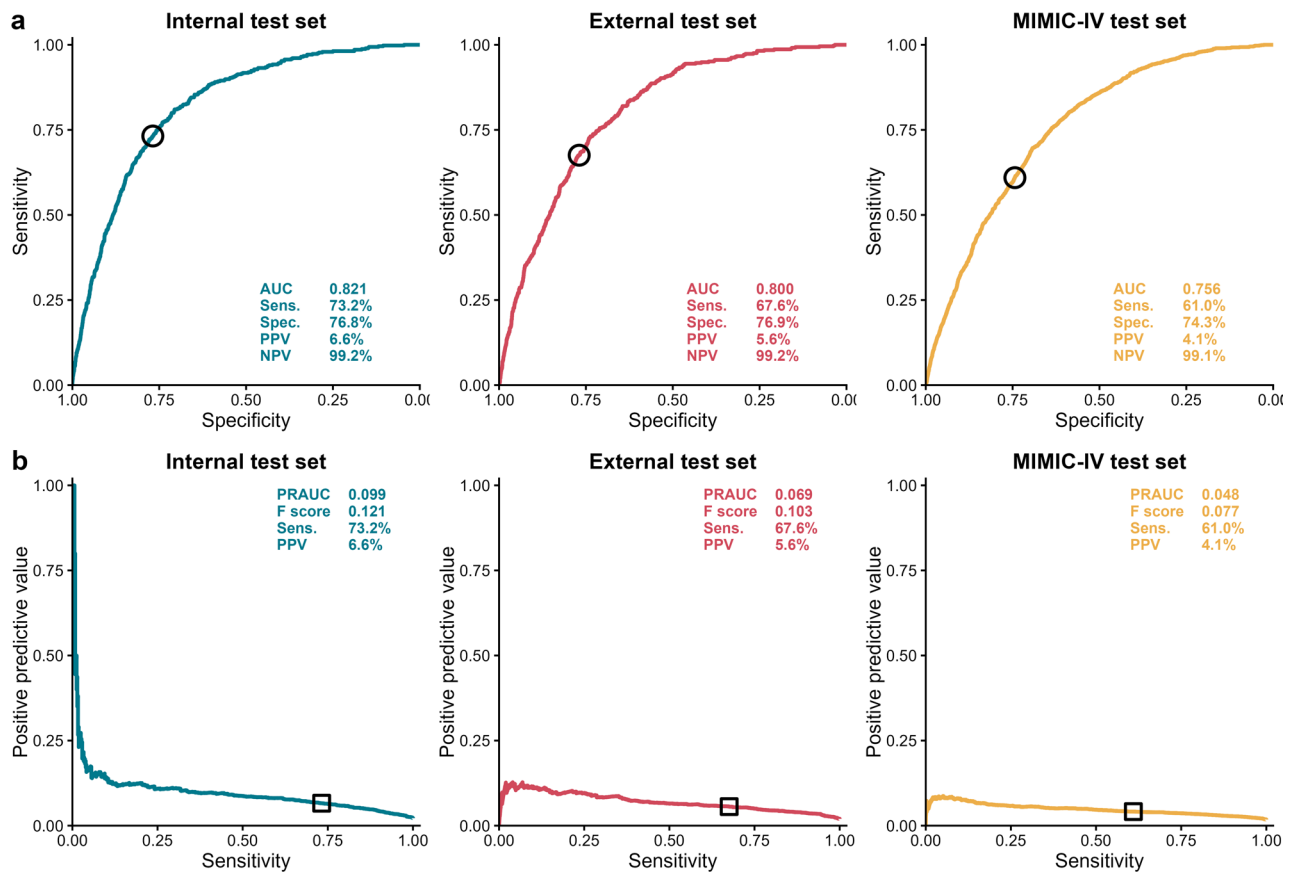


Fig. 2 Discriminative performance across test cohorts. Receiver operating characteristic (ROC) curves (**a**) and precision-recall curves (**b**) for the deep learning model across the internal test set (AUC 0.821, 95% CI 0.804–0.838), external test set (AUC 0.800, 0.782–0.818), and MIMIC-IV test set (AUC 0.756, 0.743–0.768). $N=22,205$ (internal), 22,709 (external), and 66,937 (MIMIC-IV). Circle markers denote the Youden-optimal operating point; the recommended high-sensitivity rule-out operating points and the full threshold spectrum are provided in Supplementary Table S6

gradient in the full cohort (adjusted HR 3.1–4.9 versus low); however, within the smaller BSI-positive subgroup the medium and high tiers contained few patients with overlapping confidence intervals, and their survival curves cross (for example, external test set: medium HR 3.37 [1.67–6.81] vs. high HR 3.04 [1.39–6.64]). The three-tier gradient should therefore be interpreted from the full cohort, where it is robust, rather than from the under-powered BSI-positive subgroup.

Critical ECG features for bloodstream infection prediction

An XGBoost model was used to identify key predictive ECG features (Fig. 5). Heart rate emerged as the dominant feature, followed by corrected QT interval, QRS wave axis, and T wave axis. This feature set was highly consistent across test cohorts, with the selected features explaining approximately three-quarters of the variance in the deep learning model's continuous BSI risk output ($R^2=73.42\%$ and 74.46% for the internal and external test sets, respectively, based on two-fold cross-validated XGBoost regression). Patients with BSI had significantly different ECG parameters compared to

those without, including higher heart rates and longer QRS durations (all $p<0.001$). The features identified as important by XGBoost likely overlap with, but may not fully represent, the features learned by the deep learning model from raw waveform data.

Exploratory pathogen-class analysis

In exploratory subgroup analyses, the primary BSI risk score showed similar one-vs-rest discrimination for Gram-positive, Gram-negative, and fungal infections (Supplementary Figure S3), consistent with a pan-pathogen signal rather than discrimination between pathogen classes.

Model calibration and clinical utility

Calibration curves showed good agreement between predicted BSI probabilities and observed infection frequencies, assessed using the Hosmer-Lemeshow test with decile grouping, particularly in the clinically relevant 0–15% probability range (Supplementary Figure S4). Decision-curve analysis evaluated the decision of whether to obtain a blood culture and initiate empiric

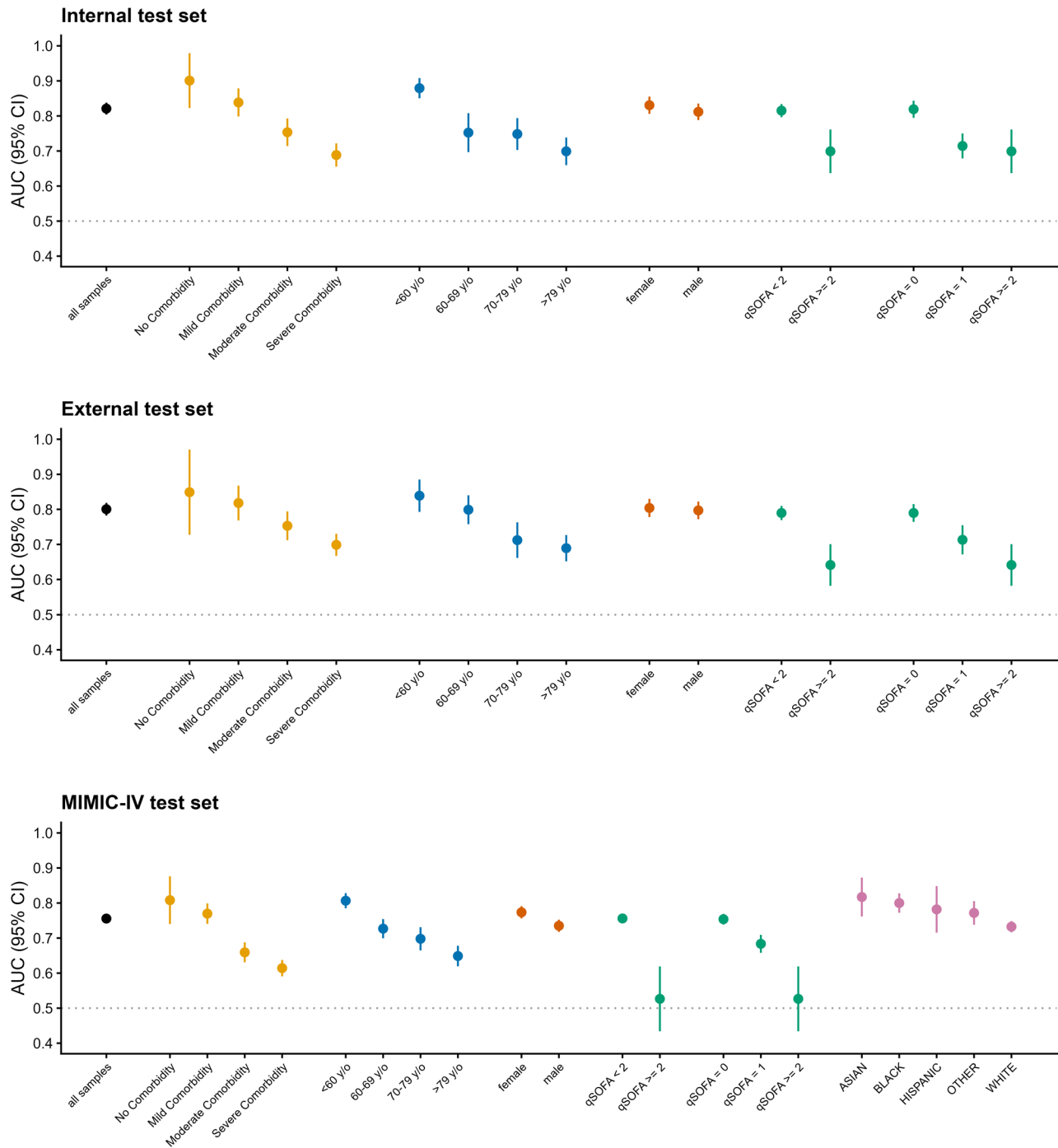
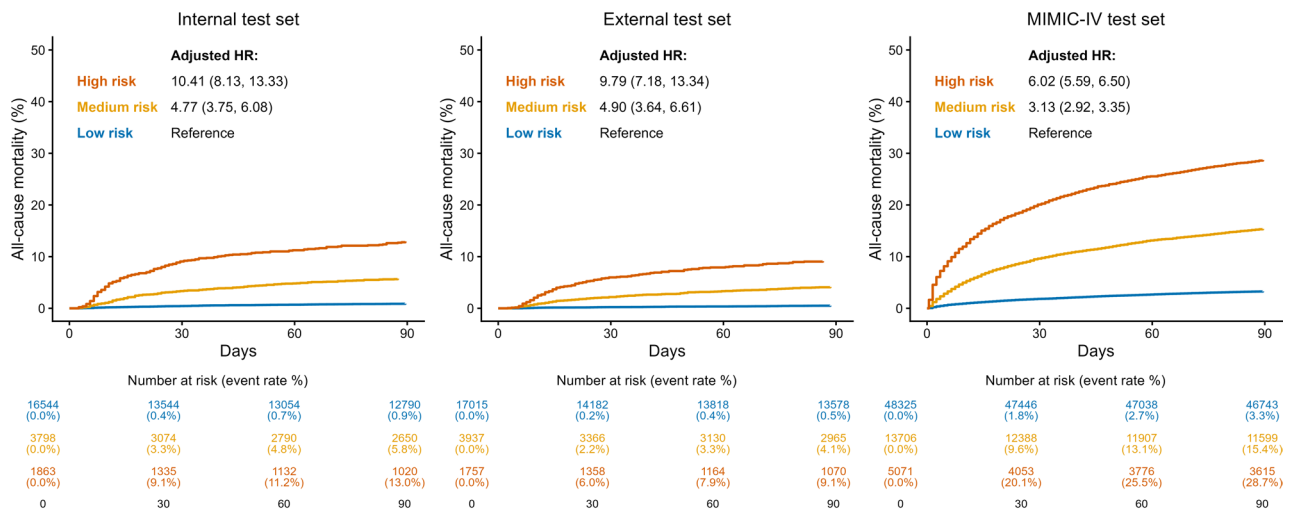


Fig. 3 Subgroup analysis across patient characteristics. Point estimates (with 95% confidence intervals) of model performance (AUC), shown as points with error bars, stratified by demographic and clinical characteristics: age categories, gender, Charlson comorbidity index, qSOFA score, and, in the MIMIC-IV cohort, race and ethnicity. Results are presented separately for the internal test, external test, and MIMIC-IV test cohorts

antibiotics. Across the clinically relevant range of threshold probabilities near the observed prevalence (approximately 1–10%), the AI-ECG model provided the highest net benefit of the strategies examined, exceeding treat-all, treat-none, and heart rate alone (Supplementary Figure S5); treat-all conferred negative net benefit beyond approximately 2%. The incremental net benefit of the

deep-learning model over heart rate alone was small but consistent across this range, indicating that the model adds information beyond a single vital sign for this decision.

All-cause mortality



All-cause mortality among patients with bacteremia

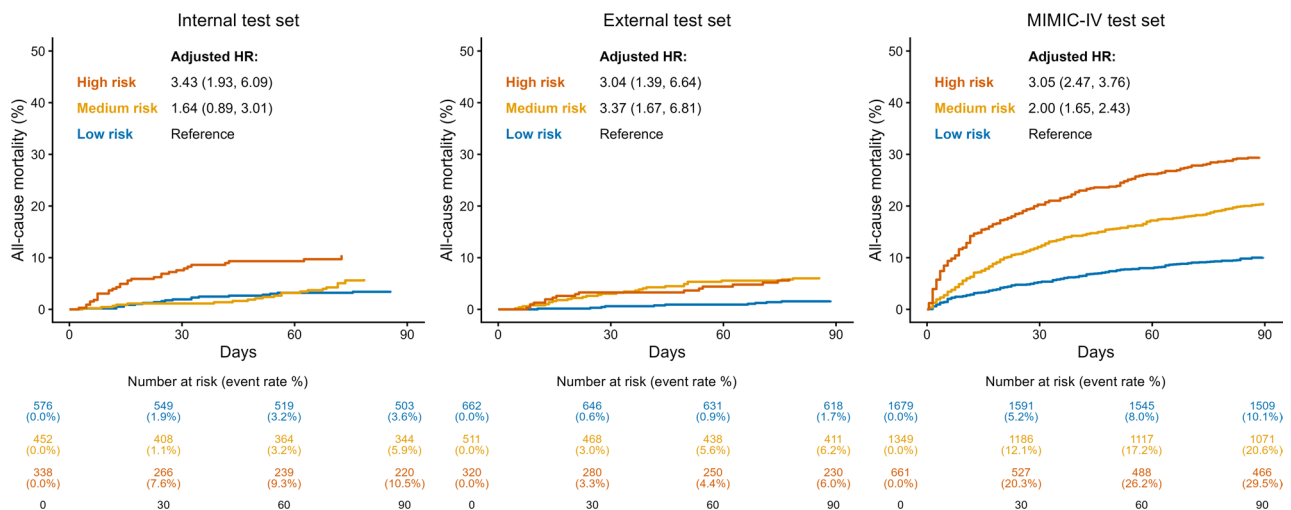


Fig. 4 Survival analysis by AI-ECG risk classification. Kaplan-Meier survival curves for 90-day all-cause mortality, stratified by AI-ECG risk categories: high-risk (prediction score > 1.625), medium-risk (0.827 to 1.625), and low-risk (< 0.827). Upper panels present mortality in the overall population; lower panels display mortality within BSI-positive patients. Adjusted hazard ratios (adjusted for age and sex) with 95% confidence intervals are shown for each risk category relative to the low-risk reference group

Discussion

Our study shows that deep learning analysis of standard 12-lead ECGs can predict clinically significant BSI with meaningful discrimination in emergency department settings. After excluding common contaminants, the model achieved robust performance in internal and external Taiwanese test cohorts (AUC 0.821 and 0.800) and maintained good performance in a large U.S. dataset (AUC 0.756). The model's consistently high negative predictive value (>99%) suggests its primary utility as a rapid rule-out tool, potentially aiding antimicrobial stewardship. Our analyses also reveal that the model's predictions are well-calibrated and offer net clinical benefit per decision-curve analysis, and that the ECG-derived risk signal is elevated across Gram-positive, Gram-negative,

and fungal infections (an exploratory, pan-pathogen observation rather than discrimination between pathogen classes; Supplementary Figure S3).

ECG signatures of bloodstream infection

The model's ability to detect BSI from the ECG is rooted in the complex cardiovascular changes induced by systemic infection, such as altered heart rate and repolarization, which were key predictive features in our model [16, 20, 31, 32]. The ECG-derived signal was present across pathogen classes, suggesting a shared cardiovascular response to bloodstream infection; whether ECGs can discriminate between pathogen classes is unresolved and would require a dedicated multi-class model, which we identify as a priority for future work.

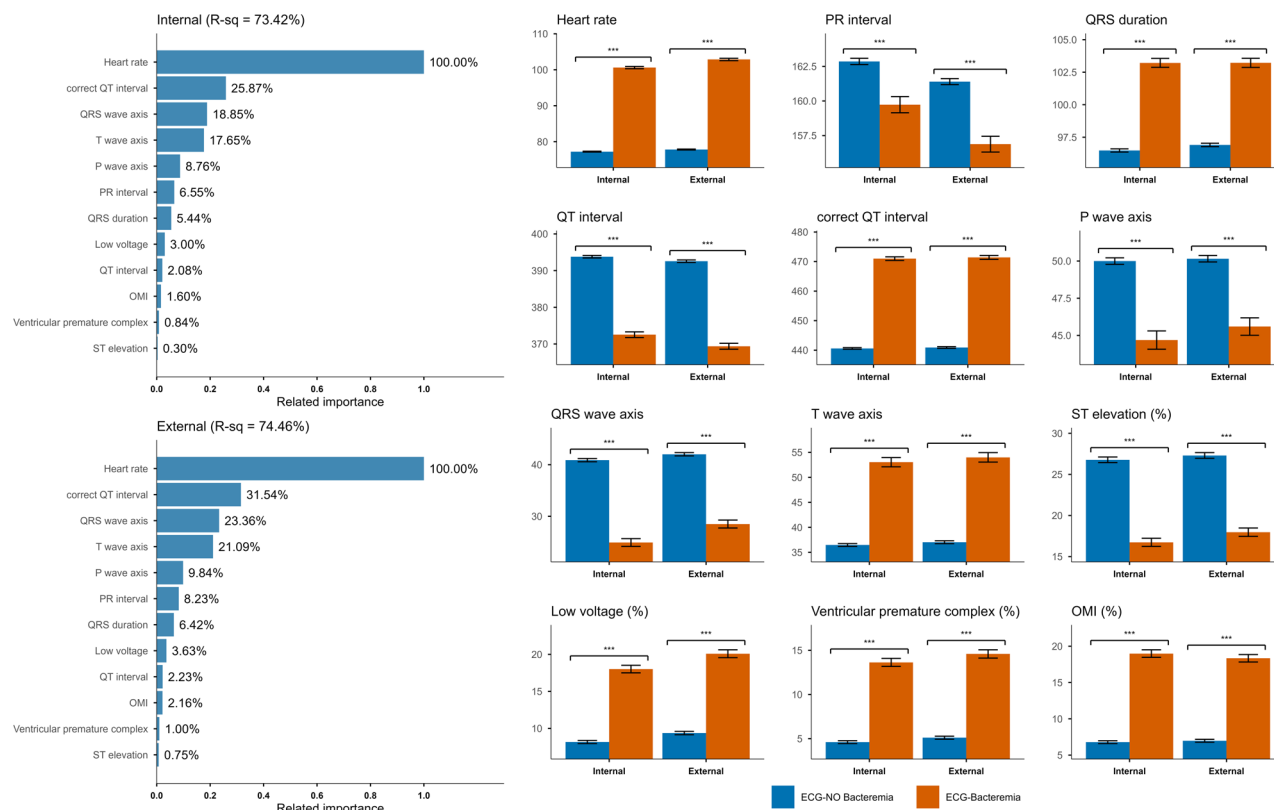


Fig. 5 Key ECG features for bloodstream infection prediction. Left panels show relative feature importance based on information gain derived from XG-Boost models. R-squared values indicate the proportion of variance explained. Right panels provide comparative analysis of key ECG parameters between patients with and without bloodstream infection, demonstrating statistically significant differences

Model performance and generalizability

The model discriminated best in patients with qSOFA = 0, the subgroup in which clinical uncertainty is often highest [33]; however, this subgroup also has the lowest BSI prevalence (1.4%), so its high NPV (99.7%) is only modestly above the null value of 98.6% (incremental gain approximately + 1.1% point; negative likelihood ratio approximately 0.2). The model's value in this subgroup therefore lies in a moderate reduction of an already low post-test probability rather than in the NPV alone, and it should support, not replace, clinical judgment. Subgroup prevalences and predictive values are reported in Supplementary Table S6. The model's reduced performance in patients with higher qSOFA scores or advanced age aligns with previous reports of reduced prognostic utility of clinical scores in such populations [34].

Although the cohort definition permitted an ECG and a blood culture during the ED encounter, analysis of ECG timestamps showed that all ECGs (100% in every cohort) were obtained within 1 h of triage (median 0.2–0.3 h; Supplementary Table S7), confirming that the reported discrimination reflects prediction at the point of care.

The lower AUC observed in the MIMIC-IV cohort (0.756) likely reflects several factors: differences in ECG hardware (Philips in Taiwan vs. potentially multiple

systems in MIMIC-IV), different patient demographics (U.S. vs. Taiwan), different healthcare contexts (critical care database vs. all-comer ED), and the inclusion of ICU patients who may already be receiving treatment. The less precise BSI endpoint in MIMIC-IV also likely contributed. While our Taiwan cohorts employed adjudicated blood cultures with meticulous contaminant exclusion, the MIMIC-IV endpoint relied on a binary positive culture flag without organism-level review [27], potentially introducing label noise that would be expected to attenuate measured discrimination. The lower BSI prevalence in the MIMIC-IV cohort (1.8% vs. 2.0–2.2% in Taiwan) further reflects the different case ascertainment strategy: the suspicion of infection framework requires concurrent antibiotic and culture orders within a predefined time window, which may undercount BSI events compared to direct blood culture review. Additionally, the MIMIC-IV cohort includes a broader ED population, many of whom may not have had blood cultures ordered, diluting the BSI prevalence.

Clinical utility and implications for antimicrobial stewardship

Although the model's AUROC values were favourable, AUROC at the low target prevalence of this setting (~

2%) is dominated by the large number of true negatives and can overstate clinical usefulness; we therefore base the assessment of clinical utility below on prevalence-sensitive measures, namely predictive values, likelihood ratios, and decision-curve net benefit, rather than on AUROC alone. The model's clinical value as a rule-out tool is supported by its high NPV (99.2% in the internal test set). However, this metric must be interpreted in context of disease prevalence [35]. At the observed test set BSI prevalence (~ 2%), the null NPV (predicting all negative) is already 97.8%, and the model's incremental gain is + 1.4%. The negative likelihood ratio of 0.35 indicates moderate rule-out value. Under the assumption that sensitivity and specificity remain constant, the model's incremental NPV would increase with prevalence (Supplementary Table S9): for example, an illustrative + 4.4% points at 7% prevalence and + 9.2 at 15%. This projection is illustrative only; because a different case-mix can change the spectrum of disease and therefore the likelihood ratios themselves, the operating characteristics in a higher-prevalence population must be established prospectively rather than assumed. This suggests the model's rule-out utility is greatest in clinical settings with higher pretest probability.

In our internal test cohort, applying the model at a high-sensitivity rule-out threshold (approximately 98% sensitivity) classified 24% of patients as low-risk while missing only 0.4 true BSI per 1,000 ED visits, whereas the Youden-optimal threshold classified 76% as low-risk but missed 6 per 1,000 and is therefore not recommended for a rule-out application. Similar safety-efficiency trade-offs were observed in the external and MIMIC-IV test sets (Supplementary Table S6). Beyond rule-out, the deep learning model provides value that simpler approaches cannot. Compared with heart rate alone (AUC 0.786/0.760), logistic regression using four ECG parameters (AUC 0.792/0.765), and an XGBoost model using 39 ECG features (AUC 0.790/0.774), the deep learning model achieved AUCs of 0.821/0.800, a consistent improvement across both the internal and external test sets (Supplementary Table S8). That the deep learning model outperforms feature-based models built from the same waveforms indicates it captures predictive information in the raw signal not represented by conventional ECG parameters. The deep learning model's ability to stratify mortality risk among infected patients (Fig. 4) further supports its role as an integrated risk stratification tool rather than a standalone rule-out test.

Although we did not directly compare the AI-ECG model with conventional biomarkers, indirect comparison with published data provides context. Procalcitonin (PCT) for predicting bacteremia has reported AUCs ranging from 0.71 to 0.79 in meta-analyses [7], while C-reactive protein alone achieves AUCs of approximately

0.65–0.75 [36]. The Shapiro clinical prediction rule for bacteremia in ED patients achieved a sensitivity of 97% and specificity of 29% [10]. Clinical prediction rules for bacteremia have also shown moderate accuracy [31]. Our AI-ECG model achieves comparable or superior discrimination (AUC 0.756–0.821) while requiring no laboratory results and providing instantaneous output. However, a direct head-to-head comparison in the same patient population is needed, and future studies should evaluate whether combining AI-ECG predictions with conventional biomarkers improves diagnostic accuracy.

Study limitations and future directions

Our study has several limitations. First, the retrospective design means that only patients with both an ECG and a blood culture were included. Although ECGs are obtained routinely at triage and were available for most patients with a blood culture, requiring a blood culture restricts the cohort to patients with clinically suspected infection; patients with an ECG but no blood culture (the majority of ED patients) were excluded, and the model's performance in an unselected ED population is unknown. Second, because all ECGs were obtained within 1 h of triage (Supplementary Table S7), the model's rule-out value as measured by NPV is partially driven by the low BSI prevalence (~ 2%) in these test cohorts after deduplication to one ECG per patient. At this prevalence, the incremental NPV over a null strategy is modest (+ 1.4% in internal test set; Supplementary Table S9). The model's rule-out utility is expected to be greater in populations with higher BSI prevalence, such as febrile ED patients or immunocompromised populations. Third, the model was developed entirely on a Taiwanese population. While MIMIC-IV provides cross-ethnic validation, the performance gap suggests population-specific calibration may be needed. Fourth, the XGBoost feature importance analysis serves as a parallel interpretable model and does not directly explain the deep learning model's internal representations. Fifth, we did not directly compare our model with established biomarkers such as procalcitonin or C-reactive protein, as these were not consistently available across all cohorts (see indirect comparison above). Sixth, all data are from overlapping time periods; temporal validation (training on earlier data, validating on later data) was not performed. Technical variations between sites, such as different ECG systems, could also affect generalizability. Finally, the model's performance declined in patients with complex comorbidities and higher qSOFA scores. Future research should prioritize prospective validation studies, including silent prospective validation (model runs in background without clinical action), integration studies with clinical workflows, and cost-effectiveness analysis.

Conclusions

This multi-center study shows that deep learning analysis of standard 12-lead ECGs, obtained at the time of emergency department triage, can predict clinically significant BSI with discrimination superior to conventional ECG parameters and logistic regression baselines. The model discriminated best (highest AUC) in patients with low clinical suspicion (qSOFA = 0), although its incremental value in this low-prevalence subgroup is modest; its rule-out value would be expected to increase with prevalence, although this requires prospective confirmation. Because all ECGs were acquired at ED triage, model performance reflects true point-of-care prediction, and baseline-model comparisons confirmed that deep learning outperforms simpler models. Prospective validation in real-world clinical workflows is needed before implementation.

Abbreviations

AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
BSI	Bloodstream infection
CCI	Charlson Comorbidity Index
CI	Confidence interval
CNN	Convolutional neural network
CRP	C-reactive protein
DCA	Decision curve analysis
DL	Deep learning
ECG	Electrocardiogram
ED	Emergency department
HR	Hazard ratio
ICD	International Classification of Diseases
LR	Likelihood ratio
MICE	Multivariate imputation by chained equations
NPV	Negative predictive value
PCT	Procalcitonin
PMM	Predictive mean matching
PPV	Positive predictive value
PRAUC	Area under the precision-recall curve
qSOFA	Quick Sequential Organ Failure Assessment
SIRS	Systemic inflammatory response syndrome
SOFA	Sequential Organ Failure Assessment
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-026-03657-0>.

Supplementary Material 1

Acknowledgements

None.

Author contributions

PHC: Conceptualization, Data curation, Formal analysis, Writing - original draft. SYL: Data curation, Software. DJT: Software, Methodology. YRW: Data curation. SWL, PYC, JHC, MSD: Data curation, Writing - review & editing. CSL: Methodology, Supervision. CL: Conceptualization, Methodology, Supervision, Writing - review & editing. CHL: Conceptualization, Supervision, Writing - review & editing.

Funding

This study was supported by the Tri-Service General Hospital (TSGH-D-114168, TSGH-D-115176 to P.-H.C.) and National Defense Medical University (MND-MAB-C02-115011 to C. L.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Data availability

The datasets from Tri-Service General Hospital used in this study are available from the corresponding author upon reasonable request. The MIMIC-IV dataset is publicly available through the PhysioNet repository (<https://physionet.org/content/mimiciv/1.0/>) following completion of required training and data use agreement.

Declarations

Ethics approval and consent to participate

This retrospective, multi-center study was approved by the Institutional Review Board of Tri-Service General Hospital (TSGH IRB No. C202105049), Taipei, Taiwan. The requirement for informed consent was waived by the IRB due to the retrospective nature of the study. The MIMIC-IV database was accessed under PhysioNet credentialed data use agreement. All methods were performed in accordance with the Declaration of Helsinki and relevant guidelines and regulations.

Consent for publication

Not applicable (no individual patient data are presented).

Competing interests

The authors declare no competing interests.

Author details

¹Division of Hematology and Oncology, Department of Internal Medicine, Tri-Service General Hospital, College of Medicine, National Defense Medical University, Taipei, Taiwan

²Department of Oncology, Tri-Service General Hospital, College of Medicine, National Defense Medical University, Taipei, Taiwan

³School of Public Health, College of Public Health, National Defense Medical University, Taipei, Taiwan

⁴Digital Medical Center, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan

⁵School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan

⁶Division of Infectious Diseases and Tropical Medicine, Department of Internal Medicine, Tri-Service General Hospital, College of Medicine, National Defense Medical University, Taipei, Taiwan

⁷Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, College of Medicine, National Defense Medical University, Taipei, Taiwan

Received: 28 March 2026 / Accepted: 22 June 2026

Published online: 29 June 2026

References

1. Kumar A, Ellis P, Arabi Y, Roberts D, Light B, Parrillo JE, et al. Initiation of inappropriate antimicrobial therapy results in a fivefold reduction of survival in human septic shock. *Chest*. 2009;136:1237–48.
2. Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med*. 2017;43:304–77.
3. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit Care Med*. 2014;42:1749–55.
4. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med*. 2006;34:1589–96.

5. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *N Engl J Med*. 2017;376:2235–44.
6. van Duin D. Diagnostic challenges and opportunities in older adults with infectious diseases. *Clin Infect Dis*. 2012;54:973–8.
7. Hoeboer SH, van der Geest PJ, Nieboer D, Groeneveld AB. The diagnostic accuracy of procalcitonin for bacteraemia: a systematic review and meta-analysis. *Clin Microbiol Infect*. 2015;21:474–81.
8. Serafim R, Gomes JA, Salluh J, Póvoa P. A Comparison of the Quick-SOFA and Systemic Inflammatory Response Syndrome Criteria for the Diagnosis of Sepsis and Prediction of Mortality: A Systematic Review and Meta-Analysis. *Chest*. 2018;153:646–55.
9. Kuti EL, Patel AA, Coleman CI. Impact of inappropriate antibiotic therapy on mortality in patients with ventilator-associated pneumonia and blood stream infection: a meta-analysis. *J Crit Care*. 2008;23:91–100.
10. 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.
11. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18:465–78.
12. Lou YS, Lin CS, Fang WH, Lee CC, Wang CH, Lin C. Development and validation of a dynamic deep learning algorithm using electrocardiogram to predict dyskalemas in patients with multiple visits. *Eur Heart J Digit Health*. 2023;4:22–32.
13. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25:70–4.
14. Lee CH, Liu WT, Lou YS, Lin CS, Fang WH, Lee CC, et al. Artificial intelligence-enabled electrocardiogram screens low left ventricular ejection fraction with a degree of confidence. *Digit Health*. 2022;8:20552076221143249.
15. Deutschman CS, Tracey KJ. Sepsis: current dogma and new perspectives. *Immunity*. 2014;40:463–75.
16. Liu W, Shao R, Zhang S, Jin L, Liu R, Chen P, et al. Characteristics, predictors and outcomes of new-onset QT prolongation in sepsis: a multicenter retrospective study. *Crit Care*. 2024;28:115.
17. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice. *Crit Care Med*. 2019;47:1485–92.
18. Goh KH, Wang L, Yeow AYY, Poh H, Li K, Yeow JLL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun*. 2021;12:711.
19. Kwon JM, Lee YR, Jung MS, Lee YJ, Jo YY, Kang DY, et al. Deep-learning model for screening sepsis using electrocardiography. *Scand J Trauma Resusc Emerg Med*. 2021;29:145.
20. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315:801–10.
21. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55–63.
22. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammber A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1.
23. Lin CS, Lee YT, Fang WH, Lou YS, Kuo FC, Lee CC, et al. Deep Learning Algorithm for Management of Diabetes Mellitus via Electrocardiogram-Based Glycated Hemoglobin (ECG-HbA1c): A Retrospective Cohort Study. *J Pers Med*. 2021;11:725.
24. Hung Y, Lin C, Lin CS, Lee CC, Fang WH, Lee CC, et al. Artificial Intelligence-Enabled Electrocardiography Predicts Future Pacemaker Implantation and Adverse Cardiovascular Events. *J Med Syst*. 2024;48:67.
25. Lin C, Chau T, Lin CS, Shang HS, Fang WH, Lee DJ, et al. Point-of-care artificial intelligence-enabled ECG for dyskalemia: a retrospective cohort analysis for accuracy and outcome prediction. *NPJ Digit Med*. 2022;5:8.
26. Chang CH, Lin CS, Luo YS, Lee YT, Lin C. Electrocardiogram-Based Heart Age Estimation by a Deep Learning Model Provides More Information on the Incidence of Cardiovascular Disorders. *Front Cardiovasc Med*. 2022;9:754909.
27. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009–2014. *JAMA*. 2017;317:1471–80.
28. Lin CS, Liu WT, Tsai DJ, Lou YS, Chang CH, Lee CC, et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nat Med*. 2024;30:1461–70.
29. Tsai DJ, Tsai WT, Lin CS, Lou YS, Fang WH, Lee CC, et al. Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. *Digit Health*. 2023;9:20552076231187247.
30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–45.
31. Su CP, Chen TH, Chen SY, Ghiang WC, Wu GH, Sun HY, et al. Predictive model for bacteremia in adult patients with blood cultures performed at the emergency department: a preliminary report. *J Microbiol Immunol Infect*. 2011;44:449–55.
32. Punja M, Mark DG, McCoy JV, Javan R, Pines JM, Brady W. Electrocardiographic manifestations of cardiac infectious-inflammatory disorders. *Am J Emerg Med*. 2010;28:364–77.
33. Duncan CF, Youngstein T, Kirrane MD, Lonsdale DO. Diagnostic Challenges in Sepsis. *Curr Infect Dis Rep*. 2021;23:22.
34. Ishikawa Y, Fukushima H, Tanaka H, Yoshida S, Yokoyama M, Matsuoka Y, et al. Prognostic Accuracy of the qSOFA Score for In-Hospital Mortality in Elderly Patients with Obstructive Acute Pyelonephritis: A Multi-Institutional Study. *Diagnostics (Basel)*. 2021;11:2277.
35. Denny KJ, Gartside JG, Alcorn K, Cross JW, Maloney S, Keijzers G. Appropriateness of antibiotic prescribing in the Emergency Department. *J Antimicrob Chemother*. 2019;74:515–20.
36. Simon L, Gauvin F, Amre DK, Saint-Louis P, Lacroix J. Serum procalcitonin and C-reactive protein levels as markers of bacterial infection: a systematic review and meta-analysis. *Clin Infect Dis*. 2004;39:206–17.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.