npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01426-9

Predicting appropriateness of antibiotic treatment among ICU patients with hospital-acquired infection

Ella Goldschmidt[1,2,7], Ella Rannon[3,7], Daniel Bernstein[4], Asaf Wasserman[4], Michael Roimi[5], Anat Shrot[6], Dan Coster[1,2,8] ✉ & Ron Shamir[1,8] ✉

Antimicrobial resistance is a rising global health threat, leading to ineffective treatments, increased mortality and rising healthcare costs. In ICUs, inappropriate empiric antibiotic therapy is often given due to treatment urgency, causing poor outcomes. This study developed a machine learning model to predict the appropriateness of empiric antibiotics for ICU-acquired bloodstream infections, using data from the MIMIC-III database. To address missing values and dataset imbalances, novel computational methods were introduced. The model achieved an AUROC of 77.3% and AUPRC of 40.4% on validation, with similar results on external datasets from MIMIC-IV and Rambam Hospital. The model also predicted mortality risk, identifying a 30% mortality rate in high-risk patients versus 16.8% in low-risk groups. External validation on the eICU database showed a comparable gap, with mortality rates at 24% for high-risk and 7.7% for low-risk groups. Our study demonstrates the potential of machine learning models to predict inappropriate empiric antibiotic treatment.

Infectious diseases are considered one of the major health risks worldwide. Although the development of antimicrobial drugs has transformed the treatment of bacterial infections, the massive increase in antibiotic consumption has led to the emergence of bacterial resistance, thus reducing antibiotic efficacy[1–3]. Consequently, both the Centers for Disease Control and Prevention and the World Health Organization declared antibiotic resistance as a threat to human health[4,5] and have created guidelines for appropriate antibiotic administration[2,6,7].

Nowadays, culture incubation is the gold standard for bacterial pathogen assessment. The process takes 48 to 72 h. Typically, a gram stain is completed after 24 h, organism identification is obtained after another 24 to 48 h, and antimicrobial susceptibility testing (AST) profile is received after 72 h[8,9]. However, since early antibiotic intervention is a critical determinant of patients' survival, patients are often treated with empiric antibiotic therapy, where antibiotics are administered prior to receiving blood culture and AST results. This treatment is based on the clinician's preliminary evaluation of the patient's health state, infection history, and local bacterial resistance patterns[10,11]. Nevertheless, such treatment might be inappropriate, as the antibiotic administered might not be suitable to the pathogen. In particular, ICU-acquired infections are more likely to be resistant to a broad spectrum of antibiotics[12].

Recently, it has been shown that inappropriate antibiotic therapy (IAT) is associated with a higher incidence of treatment failure, higher mortality rate, and a prolonged hospital stay, which can also result in higher healthcare cost[13,14]. Moreover, in severe cases of bloodstream infections and cases of septic shock, a life-threatening condition characterized by a significant drop in blood pressure following an infection, IAT was found to be the most important factor in ICU patients' outcome[15]. Therefore, it is essential to develop methods for rapid identification of treatment appropriateness in ICU patients.

In standard ICU practice for patients suspected of having infection, blood cultures are first taken, and a specific empiric antibiotic therapy is administered for 7 days. If subsequent medical results reveal that the antibiotic is not appropriate, treatment is adjusted. Results available at different times after culture collection (Fig. 1a) include: (a) Gram staining results, usually available after 24 h; (b) organism identification from the culture, typically returned after 48 h[16]; and (c) the full antibiogram results, which is a susceptibility test for various antimicrobials, available after approximately 72 h[16,17]. The antibiogram provides the definitive answer. Therefore, a predictive model that assesses the appropriateness of empirical antibiotic treatment before the 72-h mark could alert physicians to reassess the antibiotic choice, potentially improving patient outcomes. However, to the best

[1]Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv, Israel. [2]Faculty of Medicine, Tel-Aviv University, Tel Aviv, Israel. [3]The Shmunis School of Biomedicine and Cancer Research, Tel-Aviv University, Tel Aviv, Israel. [4]Department of Internal Medicine "E", Tel-Aviv Sourasky Medical Center, Tel Aviv, Israel. [5]Intensive Care Unit, Rambam Health Care Campus, Haifa, Israel. [6]Independent researcher, Haifa, Israel. [7]These authors contributed equally: Ella Goldschmidt, Ella Rannon. [8]These authors jointly supervised this work: Dan Coster, Ron Shamir. ✉e-mail: dancoster@gmail.com; rshamir@tau.ac.il

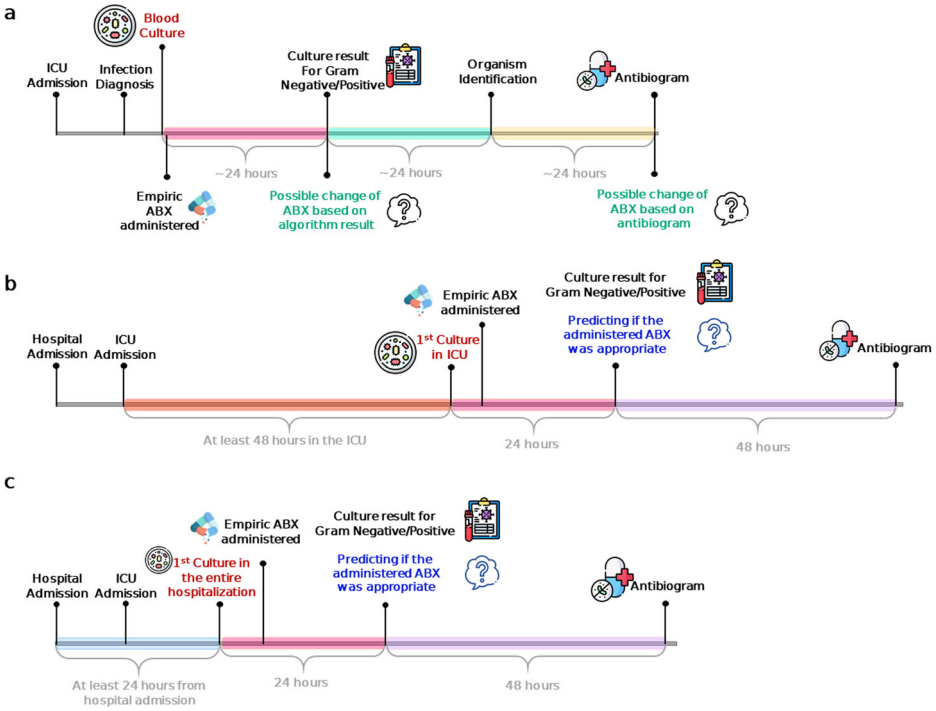

Fig. 1 | Prediction timeline and the model pipeline. a General timeline of culture results and empirical antibiotic treatment in ICU. After an infection diagnosis is available, a blood culture is taken (BCT), and then the empirical antibiotic (ABX) treatment starts. Gram staining results are available 24 h after the blood culture is taken (pink interval). Organism identification requires an extra 24 h (green interval), and the antibiogram culture results take an additional 24 h to be processed and returned from the laboratory (yellow interval). **b**, **c** Patients meeting either one of the following criteria were included in the cohort: **b** At least 48 h passed from ICU admission until the first BCT (orange interval). **c** At least 24 h passed from hospital admission to BCT (light blue interval). Our model uses the data collected until 24 h after BCT and then returns the prediction whether the ABX administered was appropriate or not. At that point (marked in blue), the prediction gives the physician the opportunity to reassess the ABX therapy and modify it if needed, 48 h earlier than the antibiogram results.

of our knowledge, as of now, no machine learning model has been developed to predict antibiotic appropriateness in ICU patients with hospital-acquired infections.

In this study, we developed a machine learning model that predicts the appropriateness of antibiotic empirical treatments based on electronic medical records (EMRs) of ICU patients with hospital-acquired infection. Our prediction is made 24 h after the blood culture is taken (Fig. 1b, c), approximately 24 h after the empirical antibiotic has been administered[18], and some 48 h before the antibiogram is available, allowing the physician an early opportunity to reassess the antibiotic choice. Unlike previous models that tried to make the prediction at the time of culture collection[19,20] or at first antibiotic administration[21,22], we use also clinical data obtained during the consecutive 24 h (not including additional data from the microbial lab such as gram staining) and make the prediction at this point. We hypothesized that by that time, the patient's lab measurements and vital signs are already affected by the antibiotic intervention and therefore give better indication of whether the antibiotic treatment is appropriate.

In the process of method development, we also devised novel computational methods and a flexible pipeline to deal with challenges that often arise when dealing with EMR data, such as missing values and imbalanced data. The methods are described in detail and can be adopted for other models that use EMRs.

Results

Cohort description

We used MIMIC-III, an open-access, anonymized database of EMRs of ICU patients, to develop, validate, and test our model. Data from 53,423 distinct ICU stays of adult patients admitted to Beth Israel Deaconess Medical Center (Boston, MA, USA) between 2001 and 2012 are included in the database[23,24]. The dataset contains for each patient stay time-independent (static) features, such as age, gender, ethnicity, weight, height, and a large variety of time-dependent (dynamic) features that are measured during hospitalization, including vital signs, lab measurements, and drug administrations. We used 55 continuous features (Table 1), 7 drug features (Supplementary Table 1) that were created by aggregating 242 drugs into 11 drug categories (Supplementary Table 2), and 39 categorical features (Supplementary Table 3). For all features, only values recorded before the *prediction time* (henceforth abbreviated as PT), set to 24 h after the time the blood culture was taken (abbreviated as BCT), were considered. We considered for our cohort only patients with suspected hospital-acquired infection (Fig. 1b, c). Overall, the training set consisted of 105 patients divided into two classes. The *inappropriate treatment group*, defined as those who received antibiotic treatment to which the pathogen was resistant, consisted of 22 patients. The remaining 83 patients received antibiotic treatment to which the pathogen was sensitive and therefore were included in the *appropriate treatment group*. The validation set comprised 30 patients, including 8 who were administered inappropriate treatment and 22 who received appropriate treatment.

Administered antibiotics analysis

Analysis of blood culture results and the drugs administered to the patients in our cohort revealed that Coagulase-positive *Staphylococcus aureus* was the most common pathogen, detected in 51% of the patients (69/135). The most common antibiotics administered to patients with that organism were *vancomycin* (50.7%, 35/69) and *levofloxacin* (23%, 16/69). Overall, the most common antibiotic administered to patients was *vancomycin* (57%, 77/135, Supplementary Fig. 1).

Furthermore, the AST results of those blood cultures revealed the most common pairing of an organism and the antibiotic tested on it. Of the antibiotics tested on Coagulase-positive *Staphylococcus aureus, Gentamicin* had 43 cultures (1/43 had a resistant outcome), *Oxacillin* had 43 (23/43 resistant), and *Levofloxacin* had 42 (25/42 resistant). The most resistant bacteria was *Enterococcus faecium*, which was resistant to at least one type of antibiotic 73% of the times it was observed (38/52), and the antibiotic that had the highest incidence of resistance was *erythromycin*, which experienced resistance 70.8% of the times (34/48) (Supplementary Fig. 2).

Appropriate antibiotic treatment model

Our model aimed to predict the risk of administering an inappropriate antibiotic treatment to an ICU inpatient. In order to select the optimal model, we used five iterations of stratified 5-fold cross-validation over the training set. In each iteration, we evaluated the model using the mean area under the receiver-operator characteristics curve (AUROC) and the area under the precision-recall curve (AUPRC) over all five folds. We then

Table 1 | Statistics of the continuous features used in our pipeline

Feature (unit)	Inappropriate			Appropriate			*p*-value
	N	Mean ± SD	Time from BCT	*N*	Mean ± SD	Time from BCT (mean ± SD)	
Age (years)	22	69.45 ± 15.19		83	65.52 ± 16.76		0.67
Admission to prediction (hours)	22	186.88 ± 110.86		83	134.39 ± 109.6		0.37
ICU admission to prediction (hours)	22	122.38 ± 88.24		83	106.79 ± 97.27		0.81
Alanine aminotransferase (IU/L)	16	171.06 ± 283.72	20.11 ± 52.06	61	112.69 ± 404.41	31.49 ± 51.16	0.83
Alkaline phosphatase (IU/L)	16	109.94 ± 40.41	27.66 ± 56.25	59	83.41 ± 34.32	42.4 ± 81.71	0.29
Anion gap (mEq/L)	22	14.09 ± 3.96	−14.65 ± 8.28	83	13.25 ± 3.57	−11.97 ± 9.2	0.73
Arterial pH (pH)	22	7.38 ± 0.1	−9.79 ± 27.4	76	7.42 ± 0.06	−3.98 ± 36.56	0.47
Aspartate aminotransferase (IU/L)	16	201.31 ± 421.97	20.11 ± 52.06	61	68.59 ± 100.96	31.49 ± 51.16	0.6
BUN (mg/dL)	21	35.52 ± 25.81	−12.63 ± 16.87	82	33.06 ± 20.12	−11.83 ± 9.13	0.91
Base excess (mEq/L)	20	−1.75 ± 5.28	−13.75 ± 16.19	75	1.4 ± 4.36	−1.68 ± 40.75	0.27
Basophils (%)	19	0.07 ± 0.15	86.26 ± 91.59	56	0.15 ± 0.18	51.71 ± 84.07	0.41
Bicarbonate (mEq/L)	22	22.23 ± 5.09	−12.02 ± 20.4	83	25.41 ± 4.47	−9.44 ± 17.72	0.24
tCO_2 (mEq/L)	21	22.52 ± 4.99	−14.98 ± 12.54	81	26.38 ± 4.78	−8.74 ± 32.1	0.24
Calcium (mg/dL)	22	7.9 ± 0.53	−12.43 ± 16.41	82	8.24 ± 0.68	−8.6 ± 19.39	0.27
Chloride (mEq/L)	22	105.68 ± 6.18	−17.28 ± 6.79	83	103.63 ± 5.38	−13.03 ± 7.59	0.51
Creatine kinase (CK) (IU/L)	15	305.27 ± 639.25	80.76 ± 74.79	59	405.9 ± 580.84	61.07 ± 103.45	0.86
Creatinine (mg/dL)	22	1.47 ± 1.15	−16.45 ± 6.65	81	1.47 ± 1.26	−10.65 ± 10.02	1
Eosinophils (%)	19	0.49 ± 0.62	86.23 ± 91.6	59	0.7 ± 0.87	50.46 ± 83.27	0.63
Glucose (mg/dL)	22	135.23 ± 43.56	−19.58 ± 5.63	83	143.88 ± 40.3	−18.94 ± 5.74	0.77
Heart rate (BPM)	22	92.36 ± 17.91	−20.55 ± 7.17	83	91.01 ± 17.14	−22.39 ± 3.83	0.93
Hematocrit (%)	22	29.96 ± 4.53	−16.55 ± 6.3	83	29.83 ± 4.25	−12.7 ± 8.43	0.97
Hemoglobin (g/dL)	22	10.1 ± 1.57	−15.47 ± 6.22	83	10.11 ± 1.43	−10.95 ± 8.47	0.99
INR	22	1.6 ± 0.61	−8.0 ± 25.15	82	1.37 ± 0.44	7.28 ± 34.87	0.45
Ionized calcium (mmol/L)	17	1.12 ± 0.07	−5.38 ± 29.32	63	1.14 ± 0.08	0.46 ± 30.24	0.81
Lactate (mmol/L)	19	2.33 ± 1.77	−1.77 ± 28.68	71	1.85 ± 1.19	20.16 ± 49.91	0.65
Lymphocytes (B) (%)	19	7.85 ± 6.25	86.23 ± 91.6	58	8.45 ± 6.11	54.39 ± 84.95	0.91
MCH (pg)	22	30.03 ± 1.78	−11.68 ± 13.07	81	30.88 ± 1.9	−10.55 ± 9.22	0.37
MCHC (%)	22	33.61 ± 1.74	−15.21 ± 6.1	83	33.82 ± 1.43	−10.65 ± 8.29	0.87
MCV (fL)	22	88.98 ± 5.07	−15.21 ± 6.1	80	91.17 ± 4.94	−11.08 ± 8.3	0.41
Magnesium (mg/dL)	22	2.07 ± 0.37	−13.79 ± 16.02	83	2.04 ± 0.27	−10.88 ± 13.66	0.91
Monocytes (B) (%)	19	3.89 ± 1.92	87.2 ± 90.53	56	3.86 ± 2.59	53.43 ± 86.25	0.98
NBP diastolic (mmHg)	22	56.68 ± 16.74	−20.56 ± 7.18	83	58.54 ± 13.51	−22.46 ± 3.88	0.88
NBP mean (mmHg)	22	78.38 ± 20.88	−20.55 ± 7.19	83	77.64 ± 15.12	−22.45 ± 3.84	0.97
NBP systolic (mmHg)	22	124.55 ± 30.92	−8.24 ± 39.91	83	121.47 ± 22.44	−17.8 ± 18.22	0.9
Neutrophils (%)	19	80.64 ± 12.2	94.64 ± 88.86	59	83.02 ± 7.87	53.58 ± 84.73	0.79
Oxygen saturation (%)	22	97.05 ± 3.42	−21.21 ± 7.02	83	97.54 ± 2.88	−22.36 ± 3.99	0.84
PEEP set (cmH_2O)	20	6.95 ± 3.07	−4.75 ± 24.73	59	6.47 ± 2.93	−12.17 ± 35.1	0.84
PT (s)	22	16.04 ± 3.39	−7.09 ± 25.77	82	14.67 ± 2.89	7.62 ± 35.02	0.44
PTT (s)	22	39.6 ± 20.72	−8.59 ± 25.11	82	34.16 ± 11.39	6.6 ± 35.26	0.63
Phosphorous (mEq/L)	21	3.42 ± 1.13	−14.47 ± 9.38	81	3.37 ± 1.09	−6.77 ± 21.84	0.96
Platelets (K/μL)	22	171.27 ± 109.02	−15.57 ± 6.31	83	181.0 ± 88.73	−5.46 ± 36.5	0.91
Potassium (mEq/L)	22	4.0 ± 0.58	−16.44 ± 6.62	83	4.02 ± 0.5	−14.1 ± 7.8	0.98
RDW (%)	22	16.43 ± 1.86	−15.21 ± 6.1	81	15.23 ± 1.84	−10.95 ± 8.34	0.24
Red blood cells (m/μL)	22	3.42 ± 0.58	−15.21 ± 6.1	82	3.27 ± 0.49	−10.81 ± 8.37	0.65
Respiratory rate (BPM)	22	22.41 ± 5.28	−19.85 ± 7.61	83	20.7 ± 5.63	−22.4 ± 3.8	0.54
Sodium (mEq/L)	22	138.0 ± 4.84	−17.28 ± 6.79	83	138.57 ± 4.51	−13.67 ± 7.85	0.87
Temperature C (°C)	22	37.09 ± 0.73	−19.72 ± 7.01	83	37.35 ± 0.84	−20.91 ± 4.76	0.51
Total bilirubin (mg/dL)	15	3.87 ± 4.85	29.37 ± 58.6	61	1.23 ± 1.61	30.73 ± 54.06	0.37
No. of previous cultures (*N*)	22	1.05 ± 1.56		83	0.35 ± 1.19		0.39

Table 1 (continued) | Statistics of the continuous features used in our pipeline

Feature (unit)	Inappropriate			Appropriate			*p*-value
	N	Mean ± SD	Time from BCT	*N*	Mean ± SD	Time from BCT (mean ± SD)	
No. of resistant cultures BCT (*N*)	22	1.36 ± 2.65		83	0.55 ± 3.31		0.61
White blood cells (K/μL)	22	13.46 ± 8.0	−13.21 ± 10.1	83	12.83 ± 5.93	−6.73 ± 21.11	0.92
Fraction of fever measurements out of all temperature measurements (%)	22	24 ± 21		83	34 ± 28		0.39
pCO_2 (mmHg)	20	37.15 ± 6.13	−14.73 ± 16.24	75	40.83 ± 8.16	−1.93 ± 40.67	0.29
pH (U) (pH)	20	5.68 ± 0.78	26.79 ± 69.36	72	5.68 ± 0.75	18.52 ± 40.37	0.99
pO_2 (mmHg)	21	108.86 ± 43.99	−14.69 ± 15.81	75	109.6 ± 35.38	−0.64 ± 41.49	0.98

For each feature, the table shows, in each class, the number of patients with the feature, the mean and standard deviation of the feature's value, the mean and standard deviation of the duration (in hours) between measurement time and blood culture time (BCT), and the *p*-value of two-sided *t*-test results between feature values of the two classes, after false discovery rate (FDR) correction[72]. For each feature, only the last values before prediction time were taken into account for this table. All lab measurements are blood-based except the vital signs and measurements marked with U, which are urine-based.

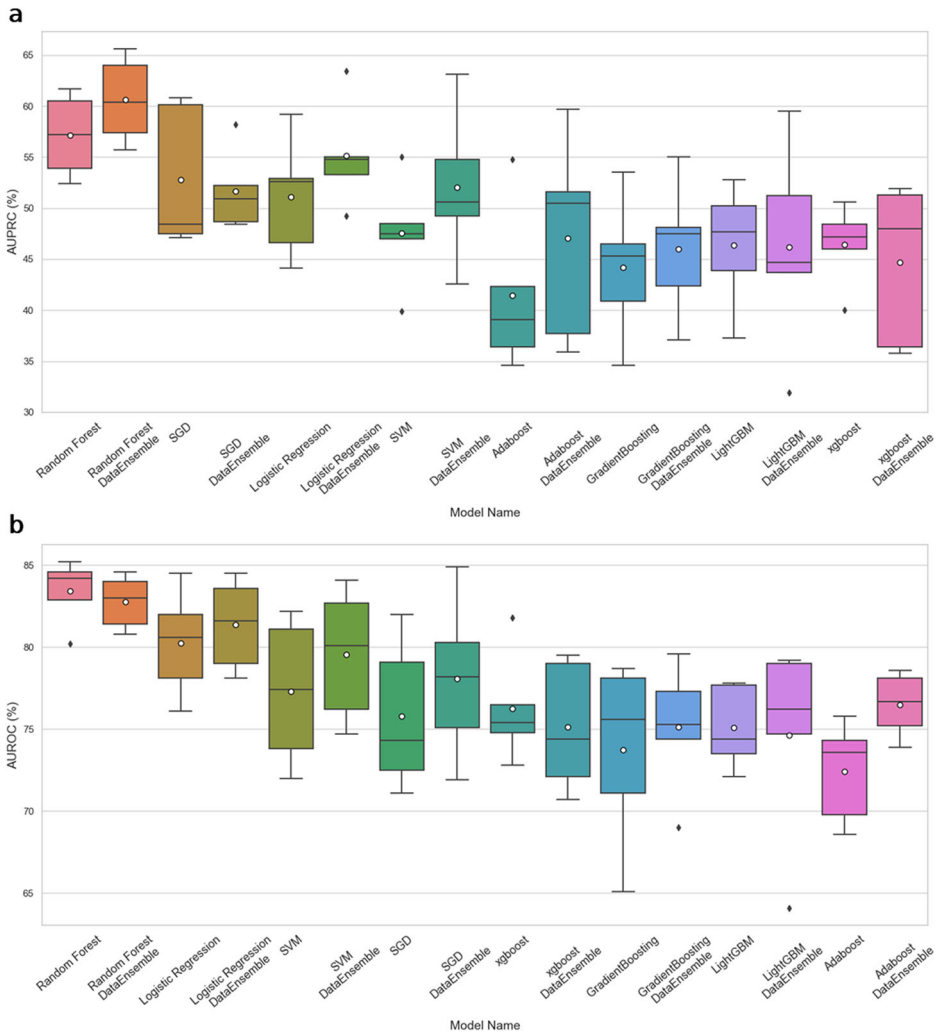

Fig. 2 | Performance of eight prediction models on the training set. Performance of eight machine learning models with and without the 'DataEnsemble' balancing approach for predicting antibiotic appropriateness. Model performance was evaluated using five iterations of 5-fold cross-validation over the training set. The horizontal line indicates the median, the white circle indicates the mean, the box indicates the IQR, the boundaries of the whiskers are the minimum and maximum values, and the black points indicate outliers. **a** AUPRC. **b** AUROC. The models are sorted by the mean AUPRC and AUROC.

averaged these metrics over the five iterations. Each model was evaluated with and without a novel training approach using balanced cohort ('DataEnsemble', see "Methods"). The Random Forest DataEnsemble model had the best performance (Fig. 2) with an AUROC of 82.76 ± 1.46% and an AUPRC of 60.61 ± 3.76% on the training set (Fig. 3). Notably, for seven out of the eight models the DataEnsemble received better median AUPRC scores compared to the original model. Thus, Random Forest DataEnsemble was chosen as the final prediction model.

MIMIC-III validation

We retrained the Random-Forest model with the selected parameters on the entire training set and applied it on the validation set (Table 2, Fig. 4a, b). After a consultation with clinicians, a classification threshold of 0.45 (i.e., classifying all samples with risk score ≥ 0.45 as positive) was chosen. For that threshold the model achieved a positive predictive value (PPV) of 50%, negative predictive value (NPV) of 86%, sensitivity of 62% and specificity of 82%. In addition, the model achieved an AUROC score of 77.3% and an

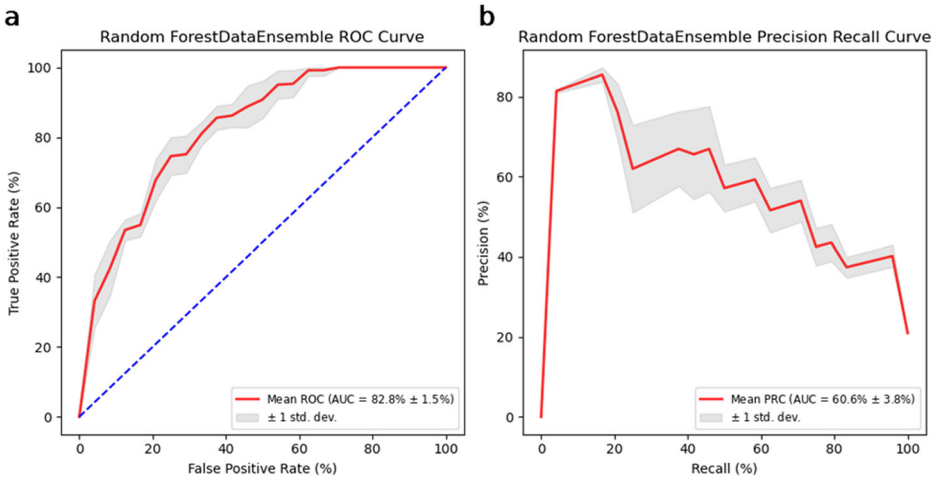

Fig. 3 | Mean performance of the Random Forest DataEnsemble model on five iterations of 5-fold cross-validation. a AUROC. **b** AUPRC. The red line is the mean, the gray area is ± one standard deviation from the mean.

Table 2 | Performance values of the Random Forest DataEnsemble model on the validation set for different risk score thresholds

Threshold	PPV	NPV	Sensitivity	Specificity
0.1	27%	100%	100%	0%
0.15	28%	100%	100%	23%
0.2	31%	100%	100%	23%
0.25	31%	100%	100%	23%
0.3	32%	100%	100%	23%
0.35	47%	94%	88%	68%
0.4	50%	84%	75%	73%
0.45	**50%**	**86%**	**62%**	**82%**
0.5	56%	82%	62%	82%
0.55	33%	76%	25%	82%
0.6	25%	70%	12%	86%

Given in bold are the values for the selected threshold.
PPV positive predictive value, *NPV* negative predictive value.

AUPRC score of 40.4%. Those values were lower than those obtained on the training set, however, it is to be expected that a predictor's performance will be reduced when validated against new samples.

The prediction time was chosen as 24 h after BCT since empiric antibiotic is administered to patients only after the BC is taken[25], and it is usually administered at least until the antibiogram is received[17]. The data from the additional 24 h between BCT and PT may already contain clues about the patient's response to the administered antibiotics. Therefore, we assessed the contribution of utilizing these data to the prediction. To accomplish this, the same Random Forest model DataEnsemble was applied using only data obtained prior to BCT. Accordingly, this model was only trained on procedures, drugs administered, vital signs, and lab measurements collected prior to BCT, and cultures taken at least 3 days prior to BCT. The demographic data was used as well since it is measured at hospital admission and is not updated during the stay. The resulting model exhibited poor performance, achieving AUROC 58% and AUPRC 29.8%, which mirrors the proportion of positive samples in the validation set at 26.67%. Subsequently, the pipeline parameters were optimized for the best mean AUPRC on the training set and the model was evaluated on the validation set. The results were similar, yielding AUROC 55.1% and AUPRC 27.7%. These findings show the importance of using data from the period following the drug administration to the patient. It is evident that training the model solely on pre-culture data without also using the data after the drug intervention results in near-random predictions, as the model lacks sufficient informative values.

We also wished to assess the contribution of data obtained before antibiotic administration to the prediction. For this goal, we applied the same Random Forest DataEnsemble model using only data obtained after the blood culture was taken (i.e., from BCT to PT). Here, the model was trained both on demographic data as well as on procedures, drug, vital signs and lab measurements occurring in the 24 h after BCT. Since culture results are returned 3 days after they are taken, taking culture results reported in this 24-h window would cause leakage of prior information. Therefore, culture features were excluded from this particular model. Again, the resulting model exhibited poor performance, achieving AUROC 61.4% and AUPRC 31%. Optimizing the pipeline parameters for the best mean AUPRC on the training set and evaluating the model on the validation set resulted in similar performance, yielding AUROC 60.2% and AUPRC 30.1%. Hence, relying solely on post-culture or pre-culture data for training the model leads to poor predictions, and incorporating information both from before and after drug administration greatly improves prediction quality.

Temporal validation

To further evaluate the performance of our model, we utilized MIMIC-IV[26] to form a temporal validation set. MIMIC-IV is an expanded and updated version of the MIMIC-III dataset, which was originally used for training and internal validation of our model. While MIMIC-III contains medical records spanning 2001 to 2012, MIMIC-IV expands this to include additional records up to 2019[26], accounting to total of 257,000 patients and 524,000 admission records[27]. To avoid data leakage and facilitate robust temporal validation, we extracted only patients who were hospitalized after 2012 and met our inclusion criteria, identifying a total of 65 patients. Among these, 55 received appropriate treatment and 10 were treated inappropriately.

We then applied the model that was trained on MIMIC-III data on the cohort from MIMIC-IV. It demonstrated an AUROC score of 73.2% and an AUPRC score of 42.3% (Fig. 5), which is consistent with the results obtained during our validation on MIMIC-III. Based on the classification threshold selected by clinicians according to the model's performance on the MIMIC-III validation set, the model achieved a PPV of 35%, an NPV of 93%, a sensitivity of 70% and a specificity of 80% on the MIMIC-IV dataset (Table 3).

External validation

To further assess the robustness of our approach, we evaluated it on a new dataset collected from the 18-bed surgical-medical ICU of Rambam Healthcare Campus (RHCC) in Haifa—a 960-bed tertiary care, academic, level 1 trauma center in Israel. The RHCC dataset, containing information

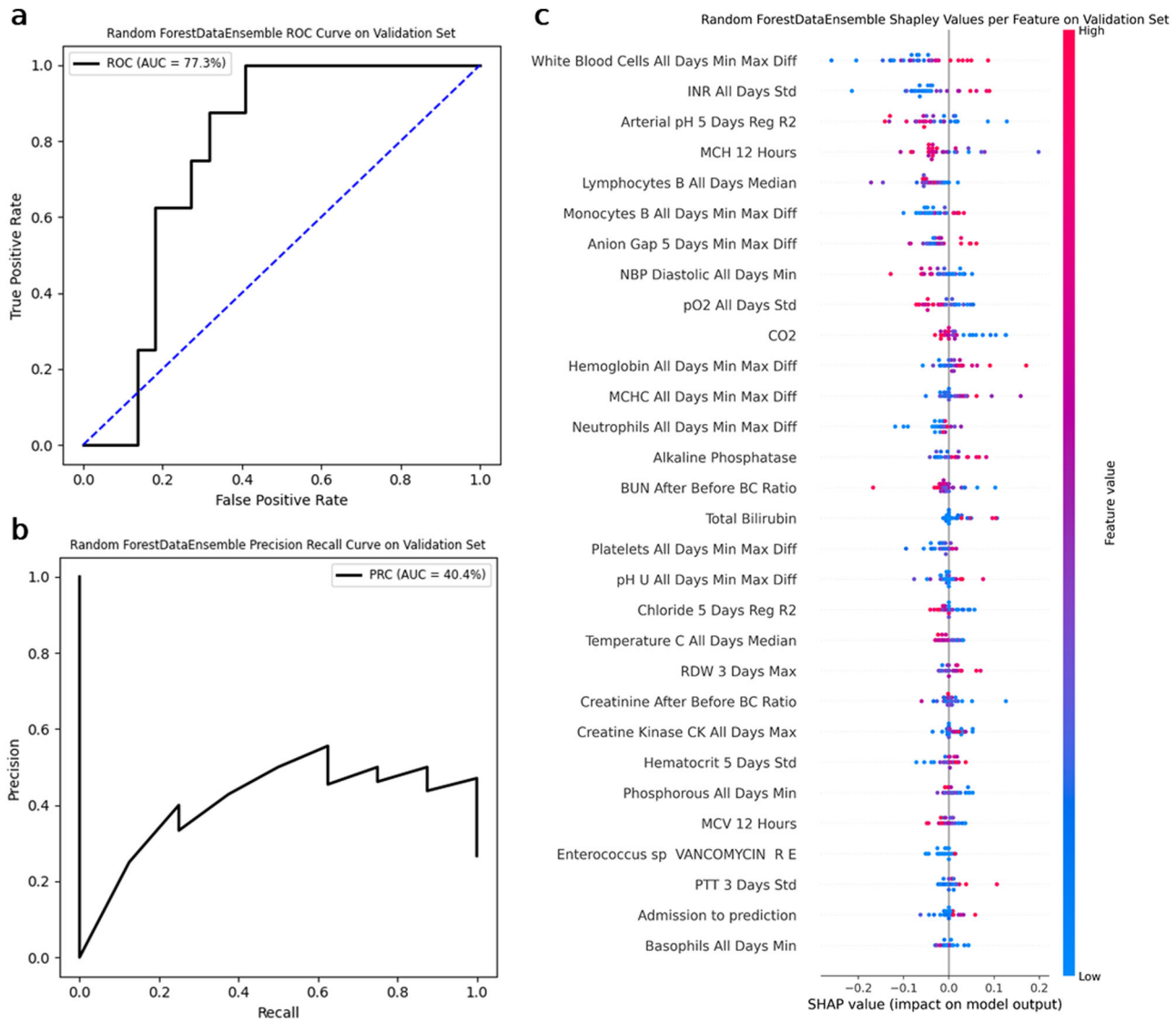

Fig. 4 | Performance of the Random Forest DataEnsemble model on the validation set. a AUROC, **b** AUPRC. **c** The thirty features with the highest absolute SHAP values. For each feature, the *X*-axis is the SHAP value, representing the contribution of that value to the model's decision. The features are ordered in descending mean absolute SHAP values. Each point corresponds to an observation where the color represents the feature value from blue (low value) to red (high value). The sign of the SHAP value indicates whether the feature observation contributes to positive or negative classification. All Days—timeframe of the entire hospitalization up to the prediction time (PT), 3/5 Days—timeframe of 3 or 5 days before PT, 12 h—measurement recorded approximately 12 h prior to PT, Min—minimal value, Max—maximal value, Min Max Diff—difference between the maximal and minimal values measured, Std—standard deviation, Reg R2—R^2 coefficient of a linear regression model fitted on values in the timeframe, After Before BC Ratio—ratio between the first value recorded after the blood culture was taken and the last value recorded before it, R—resistant culture, E—existence.

from January 2013 to December 2017, includes demographic information (age and gender), medical conditions, and chronic illnesses. It also covers laboratory measurements and vital signs (Supplementary Table 4), drug administrations (Supplementary Table 5), microbiology data, and clinical procedures (Supplementary Table 6). We extracted 161 patients who had acquired infections during their ICU stay and met our inclusion criteria. Among these patients, 106 received appropriate antibiotic treatment, while 55 were treated inappropriately. This dataset was split into training and test sets using a stratified approach, ensuring equal proportions of positive samples in both sets. This resulted in training and test sets consisting of 128 and 33 patients, respectively.

Since the new dataset differed substantially from the MIMIC-III cohorts in many ways, we adapted our model to the data while keeping most components of the methodology. This is a common practice in such situations[28,29]. Specifically, we trained a Random Forest DataEnsemble model on this dataset using the exact same data processing parameters chosen for the MIMIC-III model (see "Model's pipeline") and the same Random Forest hyperparameters employed on the final MIMIC-III model (see "Hyperparameter optimization"). Using the classification threshold determined by clinicians based on the model's performance with the MIMIC-III validation set, the model achieved a PPV of 42%, an NPV of 90%, a sensitivity of 91%, and a specificity of 36% on the RHCC dataset (Table 4). Furthermore, the model attained an AUROC score of 71.1% and an AUPRC score of 54.9% (Fig. 6), which aligns with the results obtained on the MIMIC-III validation set. Not surprisingly, applying the MIMIC-III model as is on the RHCC dataset resulted in poorer performance, likely due to differences in clinical practice and notable variations in patient characteristics, as well as the local microbial ecology[30]. See Supplementary Note 3 for more details.

Feature importance

Analysis of the features created for our model showed that none of the raw lab measurements and vital signs measurements were significant

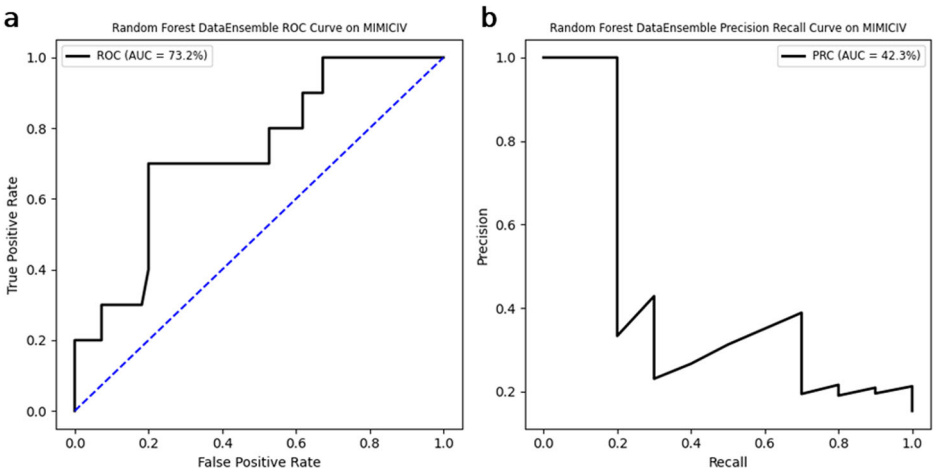

Fig. 5 | Performance of the Random Forest DataEnsemble model on the temporal validation set from MIMIC-IV. a AUROC, **b** AUPRC.

Table 3 | Performance values of the Random Forest Data Ensemble model that was trained on MIMIC-III on temporal validation set from MIMIC-IV for different risk score thresholds

Threshold	PPV	NPV	Sensitivity	Specificity
0.1	16%	100%	100%	13%
0.15	20%	100%	100%	25%
0.2	21%	91%	90%	38%
0.25	21%	93%	80%	44%
0.3	23%	92%	70%	80%
0.35	28%	93%	70%	80%
0.4	29%	93%	70%	80%
0.45	**35%**	**93%**	**70%**	**80%**
0.5	23%	87%	30%	82%
0.55	38%	88%	30%	93%
0.6	50%	87%	20%	100%

Given in bold are the values according to the threshold chosen for MIMIC-III.
PPV positive predictive value, *NPV* negative predictive value.

Table 4 | Performance values of the Random Forest DataEnsemble model trained and evaluated on the RHCC cohort for different risk score thresholds

Threshold	PPV	NPV	Sensitivity	Specificity
0.1	44%	100%	100%	0%
0.15	44%	100%	100%	0%
0.2	44%	100%	100%	0%
0.25	44%	100%	100%	0%
0.3	44%	100%	100%	36%
0.35	44%	100%	100%	36%
0.4	44%	89%	100%	36%
0.45	**42%**	**90%**	**91%**	**36%**
0.5	47%	80%	82%	55%
0.55	42%	73%	45%	68%
0.6	67%	71%	36%	91%

Given in bold are the values according to the threshold chosen for MIMIC-III.
PPV positive predictive value, *NPV* negative predictive value.

discriminators. However, previous cultures and especially resistant cultures were significantly associated with the inappropriate class (Supplementary Table 3). Moreover, the existence of any Ascites lab test is also associated with inappropriate classification.

The contribution of each feature to the model's risk score is estimated using SHAP values[31] (Fig. 4c). The most important features for the model were the difference between maximum and minimum white blood cell count (WBC) measured during the hospitalization, standard deviation of INR values measured during the entire hospitalization, R^2 of a regression model of arterial pH values in the 5-day timeframe before PT, and the mean corpuscular hemoglobin (MCH) measured 12 h before PT. Although total WBC count is a common laboratory marker for identifying patients with high risk for bacterial infection, studies have shown that WBC count had only minor discriminatory power in identifying patients with BI[32–34]. Most of the features that had a substantial impact on the model (23 of the top 30) were time-series features of vital signs and lab measurements which capture the trends in measurements over time. Moreover, the majority (28/30) used information from the 24 h after BC. Interestingly, very few (3/30) were raw measurements, and only one of the thirty most important features was related to prior cultures, and it ranked at a relatively low position 27.

We also analyzed the feature importance of the model trained on the RHCC dataset using the same approach as for the MIMIC-III model (Supplementary Fig. 3). Similar to the MIMIC-III model, most of the features that significantly influenced the model's predictions (21/30) were time-series features related to vital signs and lab measurements, while only a few (4/30) were raw measurement values. Additionally, three of these important features were related to drugs administered to patients.

Mortality prediction

The cohort used to train and test our model included only ICU inpatients who were diagnosed with bacterial infection, treated with an antibiotic, and had subsequent comprehensive antibiogram results. However, a study reported that only 19.5% of inpatients with bacteremia yielded positive blood cultures[35]. As a result, it is possible that a substantial number of patients had bacteremia but negative blood cultures, and thus they were excluded from our cohort. Such patients could still potentially benefit from a model that would assess the appropriateness of the antibiotic they received. Technically, however, our model cannot be evaluated for patients who had no antibiogram results.

As an indirect solution, we utilized our model to predict mortality, a metric that is available for all patients, regardless of the culture outcome. We measured the statistical relationship between the risk scores generated by our model, designed to predict antibiotic appropriateness, and the subsequent mortality rates at 30 days following culture collection. We based this analysis on two reported phenomena: (1) Inappropriate antibiotic treatment among ICU inpatients leads to higher mortality rate[36,37]. (2) Many negative blood cultures are false: As already mentioned, more than 80% of patients

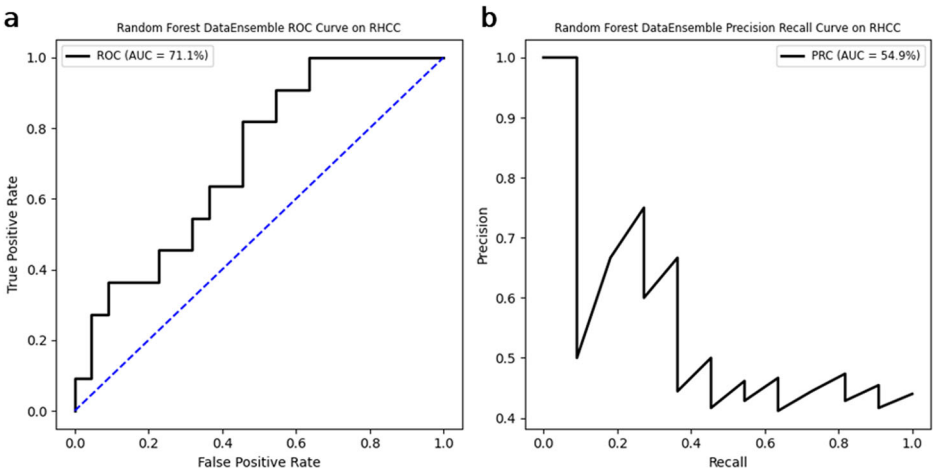

Fig. 6 | Performance of the Random Forest DataEnsemble model trained and evaluated on the RHCC cohort. a AUROC, **b** AUPRC.

Table 5 | The mortality rate of the top and bottom risk score quintiles of the patient in different datasets

	Top risk score quintile mortality rate (*p*-value)	Bottom risk score quintile mortality rate (*p*-value)
MIMIC-III positive cultures	59.3% (0.002)	14.8% (0.036)
MIMIC-III negative cultures	30% (0.001)	16.8% (0.001)
eICU	24% (0.555)	7.7% (0.015)

The *p*-values of the permutation test are written in parentheses.

with bacteremia have a negative blood culture[35]. Therefore, although our predictor was evaluated only on positive cultures (that had antibiogram), we hypothesized that our model could act as a predictor of mortality in patients who had negative cultures.

We first wished to assess whether the risk score generated by our model has the potential to predict mortality. To achieve this, we conducted an analysis on our complete cohort (training and validation set), consisting of 135 patients. We divided our patients into five quintiles based on the risk scores calculated by our model and computed the average 30-day mortality rate within each quintile. We then employed a permutation test to evaluate the significance of the mortality rates in the top and bottom groups. The mortality rate in the lowest risk quintile was significantly lower than what would be expected from random patient binning, with mortality rate of 14.8% (*p*-value 0.036). The highest risk score quintile exhibited a mortality rate of 59.3%, significantly higher than chance (*p*-value 0.002) (Table 5).

Subsequently, we conducted the same analysis on 4319 patients extracted from the MIMIC-III who had negative culture results. These patients represent a novel group that the model had not been trained on, serving as an external validation. Reassuringly, the mortality rate in the top and bottom risk score quintiles significantly differed from what would be expected through random binning, with mortality rate of 30% and 16.8%, respectively (*p*-value 0.001 for both groups).

Finally, we conducted the same test on an additional external validation dataset. The eICU dataset encompasses de-identified health information from numerous critical care units across the United States, covering approximately 200,000 patient admissions to critical care units during the years 2014 and 2015[38]. We used our model that was trained for predicting antibiotic appropriateness on MIMIC-III to predict mortality on 126 patients who met our inclusion criteria (Fig. 1b, c). Here only the bottom risk score quintile was significant, with a mortality rate of 7.7% (*p*-value of 0.015). However, a discernible difference of 16.3% was observed between the mean mortality rates of the top and bottom risk score quintiles.

Discussion

Approximately 70% of patients admitted to the ICU receive antibiotic treatment[39]. However, the percentage of patients who do not receive adequate therapy within the first 24 h of a bloodstream infection (BSI) is alarmingly high, reaching 47%[40]. On the other hand, ill-advised and excessive antibiotic use can contribute to the global antibiotic resistance problem[3]. In this study, we propose a machine learning algorithm to predict inappropriate empiric antibiotic treatment in patients with ICU-acquired bacteremia.

Previous research has focused on the utilization of machine learning models for predicting general bacterial infections[41,42], early prediction of ICU-acquired BSI[43], outcomes of BSI[44], antibiotic resistance in BSI[30,45], urinary tract infections[21,46], and other common infections[22,47]. Unlike previous studies predicting the severity of blood infections[41,42], our focus was different, namely, to predict the suitability of antibiotic treatments. While the prediction goals differed, there are common modeling aspects to those studies and ours: Reference[41] utilized time-series features, and reference[42] used data until 24 h post-antibiotic administration. Studies[22,30,44] predicted antibiotic susceptibility by creating a specific model for each antibiotic type, while study[46] predicted antibiogram results for particular groups of antibiotics. Another study[21], which focused on urinary tract infections, calculated the likelihood of antibiotic resistance to first- and second-line therapies and translated these probabilities into recommendations aimed at choosing the narrowest effective antibiotic spectrum. In contrast, the problem of antibiotic treatment appropriateness is not concerned with the resistance to each type of antibiotic but evaluates whether the treatment administered was effective by assessing the patient's response to it. These predictions have the potential to mitigate uncontrolled AMR infections, unnecessary AMR selection pressure, and subsequent transmission[48]. Due to the limited size of the available cohort, the model described in this study was not specifically trained for individual antibiotic types. Consequently, we developed one general model for predicting the appropriateness of the antibiotic treatment, without the need to specify which antibiotic was administered to the patient. The purpose of this model is to discern the physiological response to an appropriate antibiotic treatment from that of an inappropriate treatment. To the best of our knowledge, no prior studies have addressed the problem of determining the appropriateness of antibiotic treatment.

Our algorithm demonstrated promising performance both in cross-validation and in validation of an independent sample of patients (with AUROC of 82.76% and 77.27%, and AUPRC of 60.61%, and 40.44%,

respectively). These results suggest that with the use of readily accessible EMR data, it is possible to predict the appropriateness of an antibiotic treatment 48 h before the full antibiogram results are available and assist in the clinical assessment of the patient. The substantial reduction in mismatched treatment facilitated by machine learning-based recommendations that take into account the patient's medical history and records can pave the way for a future framework in which clinicians will routinely consult such algorithms and adjust the antibiotic treatment of patients accordingly. Adoption of the model in clinical practice could lead to a machine learning-guided personalized antibiotic prescription and help reduce treatment failure and overall use of antibiotics, contributing to the global effort to combat antibiotic resistance.

Moreover, the model exhibited good results in temporal validation on patients from the MIMIC-IV dataset, an extension of the MIMIC-III dataset from Beth Israel Deaconess Medical Center (Boston, MA, USA) that we used for model training and internal validation. The model achieved an AUROC of 73.2% and AUPRC of 42.3%, similar to the performance observed on the internal validation set. These results demonstrate that the model is robust enough to handle different cohorts within the same hospital. Theoretically, though, there is a potential for some data leakage between the training and temporal validation set, in case the MIMIC-IV cohort includes subsequent ICU admissions of patients included in the training cohort. While the odds for overlaps out of hundreds of thousands of admissions are very low, we have no way to identify such overlaps.

To further evaluate our method, we employed an additional external dataset from a different hospital, the Rambam Healthcare Campus (RHCC) in Haifa, Israel. This allowed us to evaluate our methodology on a completely different population with distinct data distributions, hospital practices, microbial ecology, and measurement technologies, thereby assessing its robustness. There are several major differences between these two cohorts. First, the two hospitals employ different sets of laboratory measurements and sampling frequencies due to variations in their protocols. Second, the microorganisms tested and the patterns of antibacterial resistance vary between the two hospitals, which in turn influences the types of antibiotics commonly prescribed. Third, policies are different, with patients at RHCC cohort staying on average 3.5 times longer in the hospital and ICU compared to the MIMIC-III cohort. Given these differences, we opted to use the same methodology (i.e., the same preprocessing, feature engineering, machine learning algorithm, and hyperparameters) but train a separate model for RHCC. The results of the model trained and validated on RHCC were comparable to those of the MIMIC-III model, achieving an AUROC of 71.1% and AUPRC of 54.9%. Notably, there are similarities in the top 30 features used by each model according to SHAP analysis. Both models rely heavily on time-series features derived from lab measurements and vital signs for their predictions, highlighting the consistent importance of these variables across different datasets.

As an additional support of the model, and as a way to apply it to patients with no positive culture results, we use the model's risk score to predict mortality. Our findings indicate that within our train and test MIMIC-III cohort, a higher risk score correlates with an increased mortality rate, with a significant difference in mortality rates between the highest and lowest risk quartiles. The same results were obtained on the much larger cohort of MIMIC-III patients who had negative cultures, who were not used in the model training. For the smaller eICU cohort, statistical significance was obtained only for the bottom risk score quartile. Still, mortality rates differed markedly between the highest and lowest risk quartiles. While our model was not specifically designed to predict mortality, the results indicate that the method exhibits adaptability and potential efficacy in predicting mortality in patients lacking a comprehensive antibiogram result or a positive culture result.

The models' prediction was primarily driven by the patterns in the time-series features of lab measurements and vital signs, such as the difference in the median measurement of WBC collected in the 5-day and 3-day windows prior to PT (Fig. 4c). Some of these features were recognized as significantly associated with bacterial infection in previous studies[32–34]. However, the temporal behavior of most of these features was not tested in relation to bacterial infection. Moreover, to the best of our knowledge, no study to date identified clinical measurements that are most relevant to predicting antibiotic treatment appropriateness. Notably, while time-series features of lab measurements and vital signs contributed greatly to our model, no raw measurement by itself was statistically significant for discriminating between appropriate and inappropriate treatments. Therefore, examining the features selected by our machine learning model can provide valuable insight into such discrimination. By identifying the predictors with the highest impact on the model's outcome, physicians can focus on those lab measurements and vital signs.

In addition to time-series features, our model utilized known risk factors for antibiotic-resistant infections as features, such as previous antibiotic-resistant infections, previously administered antibiotics, invasive procedures, and culture sample sites[49–51]. Many of these features were previously shown as predictive for antibiotic resistance in machine learning models that used EMR data[46,52,53]. Nonetheless, only a small fraction of the features chosen for the model (8 out of 86) were previous cultures, and they had lower importance compared to time-series features derived from lab measurements and vital signs (Fig. 4c).

In this study, we chose to set PT to 24 h after the blood culture was taken, as the results of the gram staining are typically retrieved at this time[8], and thus, at that time, clinicians could make adjustments to the patient's antibiotic treatment. Providing additional information at this time can improve decision-making by the doctors, potentially affecting the patient's outcome. Moreover, our cohort included only ICU inpatients with microbiological confirmation of a bacterial infection. However, studies have shown that only about 19.5% of inpatients with bacteremia have a positive blood culture[35]. Therefore, it is plausible that many of the patients with bacteremia will potentially not be considered for our model. Furthermore, our findings demonstrate that data collected during the additional 24 h lead to a significantly better prediction in comparison to a model trained solely on data obtained prior to the blood culture.

Our study has several limitations. First, in our study, we filtered out contaminants, while they might be considered eligible cultures for our prediction since no information is provided to classify them as contaminants at the time of gram stain. Additionally, while our methodology can be transferred to other hospitals, a specific model should be trained for each medical center. It is important to note that our model predicts antibiotic appropriateness and not resistance to specific antibiotics. In addition, the model predominantly relies on physiological data (i.e., lab measurements, vital signs, etc.), rather than relying solely on resistance patterns that might differ from hospital to hospital. Therefore, our model may be useful for detecting inappropriate treatment in other hospitals, without knowledge of specific antibiotic resistance patterns. Finally, conducting a prospective evaluation is necessary to assess the model's performance in real-world practical scenarios.

It is also important to note that the data collected pertained only to the hospitalization during which the blood culture was taken. Future studies could benefit from incorporating the complete medical history and previous hospitalization records of a patient[46]. In particular, the use of data on previous cultures can enhance the model's predictive ability, as previous instances of recurrent infections are associated with a higher risk of resistant infection in subsequent hospitalizations[54].

In addition to the contribution to predicting antibiotic resistance, this study also proposes a new pipeline for medical decision support. It outlines techniques to address challenges commonly encountered in EMRs, such as limited and imbalanced datasets and high rates of missing values. The key methods described here can serve as starting points for such an approach, but the specific model, parameters, and feature extraction process should be tailored to the medical question and the data.

Methods

Inclusion and exclusion criteria

All patients admitted directly to the emergency department or ICU who had blood cultures that were not contaminated (i.e., blood culture results

of Coagulase-negative *Staphylococcus, Diphtheroids*, *Bacillus, Aerococcus viridans, Aerococcus, Propionibacterium, Viridans streptococci, Lactobacillus*, and *Staphylococcus epidermidis*) or were not canceled during their hospitalizations were considered for this study's cohort. Out of those, to identify patients with hospital-acquired infection, we included only patients who satisfied at least one of the following conditions: (a) they were hospitalized for at least 48 h in the ICU and had their first blood culture in the ICU collected there after that time. Only the first culture collected in the ICU was used for labeling. (b) their first culture in the entire hospitalization was collected in the ICU and at least 24 h after hospital admission (Fig. 1b, c).

The cohort

We used the MIMIC-III database, containing data of 38,597 distinct adult patients[23]. Our exclusion criteria resulted in a total of 135 patients, who were split into training and validation sets. Our training set included EMRs of 105 inpatients, of whom 83 received appropriate antibiotic treatment and 22 received inappropriate antibiotic treatment. The validation set included 30 inpatients of whom 22 received appropriate treatment and 8 received inappropriate treatment.

Outcome definition

Microbiological cultures are routinely drawn in ICU. We defined the blood culture time (BCT) as the time of the culture sampling, and PT as 24 h after culture time. Only records charted before PT were used by the model.

Antibiogram results are usually available within 72 h of culture sampling[8], so prediction after 24 h may allow the physician to reconsider the antibiotic empirical treatment 48 h before the antibiogram results. The 24-h window enables one to obtain features that help assess the response of different clinical measures to the empiric antibiotic treatment (for example, the ratio between white blood cell levels before and after the empiric antibiotic treatment).

Patient treatments were designated as appropriate (negative class) or inappropriate (positive class) treatment based on the results of the culture, AST and the empirical antibiotic that was administered. The inappropriate class was defined as an antibiotic treatment where the pathogen was either not affected by the antibiotic or resistant to it. Appropriateness was decided by an internal medicine specialist and an infectious disease specialist who reviewed together the antibiotics administered and the antibiogram results for each patient.

Model's pipeline

In order to develop a robust model that will address the characteristics of our prediction objective, we constructed an extensive pipeline comprised of several steps (Fig. 7), and in each step, we evaluated a few alternative techniques. We tested each combination of techniques using five iterations of stratified 5-fold cross-validation over the training set and chose the combination that yielded the highest mean AUPRC. Below we describe each step briefly. Full details are provided in each respective section below.

The first step in the pipeline is the removal of values that were deemed outliers. We first excluded values that were not in the human range and then removed values based on two different metrics. Afterward, we filtered out features with missing rate ≥30% and removed features with variance ≤0.005.

In the next step, we created time-series features utilizing all data points available before PT. We calculated these features using two sets of timeframes, d and d+2 days before PT. Missing values in each timeframe were imputed using a linear regression model that was fitted per subject using all the feature values recorded within a larger timeframe, see "Data imputation" in "Methods". For these features, we evaluated different thresholds for the minimum number of values that are required for the fitting of a linear regression model (n) and we evaluated several timeframes (d = 2 and 3).

The next step was the normalization of the features. We evaluated two approaches: Min-Max scaling and standardization. Then we added a second imputation step to handle missing values that could not be imputed by the linear regression. We used K-Nearest Neighbors (*KNN*) algorithm[55] with k = 5 and tested several distance measures such as Sklearn's distance method (an Euclidean distance that accounts for missing coordinates), and two new distance measures.

As many of the features were highly correlated, particularly after the addition of the time-series features, we applied two steps of detecting and filtering correlated features. First, we kept only a small number of features derived from the same raw measurement by selecting those with the most significant p-value according to a two-sided t-test between the two classes. We tried several numbers of features. Following this step, we filtered highly correlated features based on hierarchical clustering.

After the removal of correlated features, we still had a high-dimensional feature space. Hence, we examined several feature selection methods: (1) Recursive Feature Elimination[56], (2) Taking the features with an importance score higher than the model's mean feature score (e.g., in the logistic regression model, taking the mean beta coefficient), (3) Taking the K features with the highest mutual information score[57] and (4) Taking the K features with highest SHAP values[31]. We also tested combinations of these four methods and several possible values of K.

Additionally, since our data was imbalanced (roughly 3/4 appropriate and 1/4 inappropriate) we tested the following approaches for oversampling of the training set: *ADASYN*[58], *SMOTENC*[59], and *BorderlineSMOTE*[60] with different balancing ratios, and also developed a novel ensemble method for data balancing which we named 'DataEnsemble'.

The last step was the prediction model selection. Here we evaluated eight different machine learning models: Random Forest[61], AdaBoost[62], Logistic Regression[63], SVM[64], SGDclassifier[57], LightGBM[65], Sklearn's Gradient Boosting Classifier[57,66] and Xgboost[67].

Every combination of techniques applied in each step above was tested in iterated cross-validation. The final prediction model chosen was Random Forest DataEnsemble. The combination and the parameter values chosen are described in each of the relevant method sections.

Outlier removal

To eliminate measurements that were grossly incorrect due to manual typos or technical errors, we manually defined with clinicians a range of possible values per each feature (including pathological values), and excluded values outside this range. A total of 711 values (0.4% of the values of all features) were excluded in this step, see Supplementary Table 7.

For the remaining values, we checked two approaches to removing extreme measurements. Both of these methods were calculated on the training set, and were later applied on the validation set. The first method used the IQR. Denote by $q_{0.75}$ ($q_{0.25}$) the value at the 75th (25th) percentile and set $q_{diff} = 1.5 \times (q_{0.75} - q_{0.25})$. Then only values in the range $q_{0.25} - q_{diff} < x < q_{0.75} + q_{diff}$ were kept.

The second approach used Z-scores, filtering out values that are more than two standard deviations from the mean of the feature.

We analyzed the percentage of values that were removed after applying both methods. Z-score discarded a mean of 4.12% of the feature values and a median of 3.69%, while IQR removed a mean of 5.44% and a median of 3.67% (Supplementary Table 7). Following these results, we used the Z-score approach.

Normalization

We evaluated two approaches for feature normalization. The first (Eq. (1)) is the normalization of all features to values between 0 and 1 according to the maximum ($X_{\max}$) and minimum ($X_{\min}$) values of each feature in the training dataset.

$$X^i_{scaled} = \frac{X^i - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

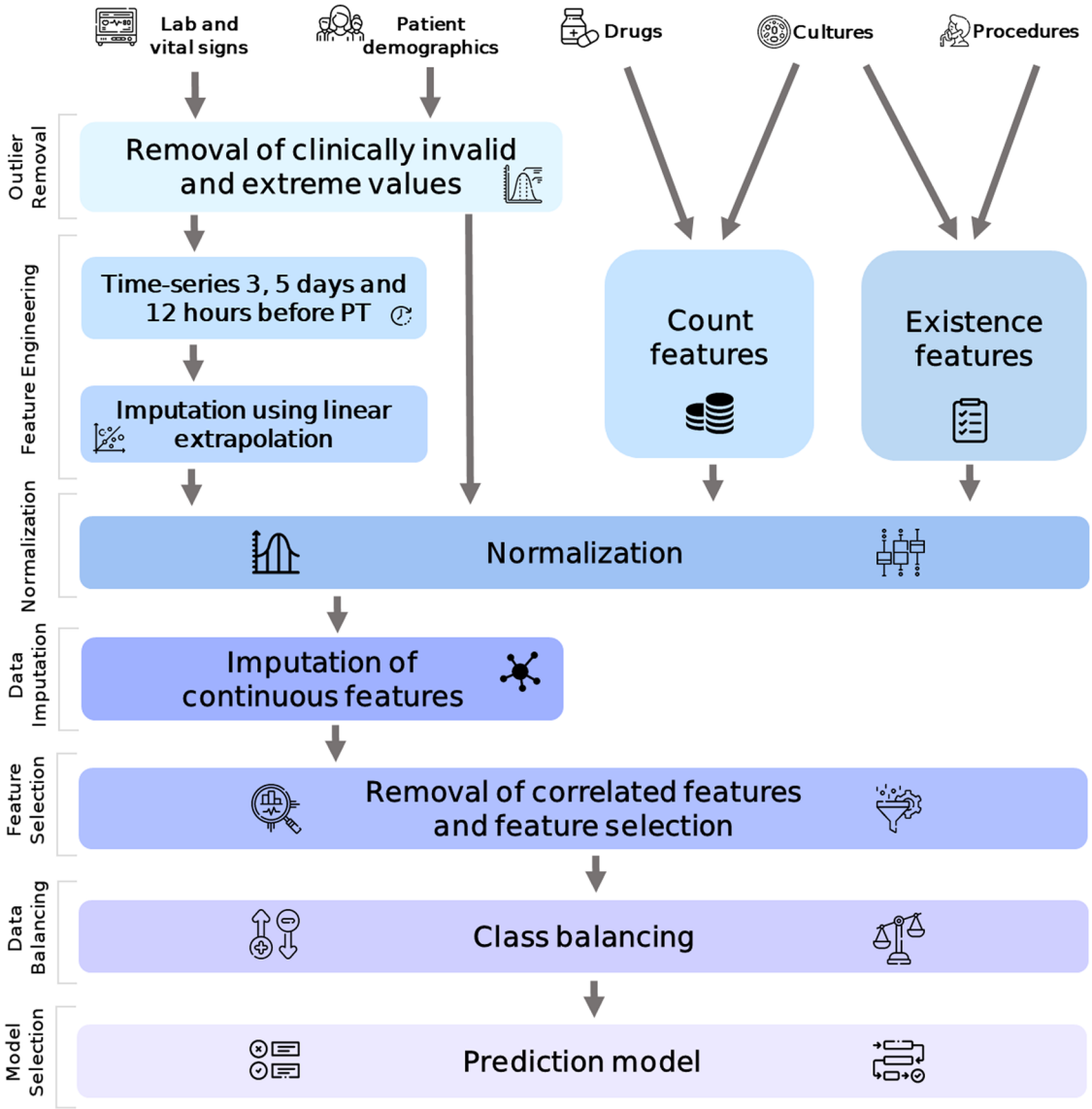

Fig. 7 | Model pipeline. The steps included in the model pipeline presented in this study. Existence features are binary (e.g., existence of a culture resistant to penicillin); Count features are numerical and ordinal (e.g., number of antibiotic drugs administered to the patient). PT prediction time.

The second (Eq. (2)) was standardization to a normal distribution with a mean of zero and a standard deviation equal to one.

$$X^i_{scaled} = \frac{(X^i - X_{mean})}{X_{std}} \quad (2)$$

Both normalizations were fitted on the training set data, and later applied on the validation set as well. The normalization method chosen was standardization as it yielded better results.

Feature engineering

The features created for the model are composed of six main categories: (1) patient demographics, (2) lab measurements, (3) vital signs, (4) drug administration, (5) previous lab cultures and (6) medical procedures. We tested removal of features with high missing rates, for rates 20%, 30%, 40%, and 50%, and chose to exclude features with missing rate >30%. Moreover, after the feature engineering process (see below), features with variance <0.005 were excluded as well.

Demographics. The demographic features included, among others, age, gender, and ethnicity, as well as time since admission to the hospital and to the ICU, and measurements such as weight and BMI.

Lab measurements and vital signs. We used as features the median, standard deviation, minimal value (min), maximal value (max), and their difference (min-max diff) per each timeframe described above. See Supplementary Note 1 for more details.

Drugs. We mapped all the drugs into 11 clinically relevant groups (Supplementary Table 2) with the help of a general physician. For each drug group, and each of the timeframes described above, we collected the total number of drugs from the group that the patient received.

Cultures. We extracted binary features indicating the properties of previous culture taken from the patient, when available. See Supplementary Note 1 for more details.

Medical procedures. Finally, we added binary features for four categories of invasive procedures that frequently cause infection:

Arterial Line, Catheter, Ventilation, and Tubes (Supplementary Table 8), and indicated if the patient has undergone a procedure from each category.

The mean time from the first lab or vital sign measurement to PT was 7.15 ± 4.65 days (Supplementary Fig. 4). Therefore, we generated time-series features for lab measurements, vital signs, and drugs for two time-frames: d and d+2 days before PT. We tested d = 3 and 4, and 3 yielded better results. Additionally, for lab measurements and vital signs, we also used a timeframe of the entire hospitalization period up to PT. See Supplementary Note 1 for more details.

Data imputation

Missing values were observed mainly in lab measurements and vital signs. For repeatedly measured values, a linear regression model was fitted (see "Feature engineering"). We imputed the missing values of features in a certain timeframe based on those linear regression models. This strategy assumed that missing values are more accurately imputed using patient-specific measurements rather than values of all patients. Regression was performed per 3 or 5-day timeframe. If a patient was missing max, min, median, or min-max-diff time features in a certain timeframe, we extended the timeframe used to impute these values to 5 and 10 days, respectively. Moreover, the feature value 12 h before PT was imputed using the 3-day linear regression, and if a regression model was not available for this timeframe, 5-day linear regression was used. Since large regression coefficients can lead to extreme imputed values, all the values produced by this extrapolation method underwent extreme and non-human values removal (Fig. 8).

The rest of the time-series features, other continuous features (e.g., last lab measurement recorded), and instances where there were not enough values for the fitting of a linear regression model (see "Feature engineering"), were imputed based on the *KNN* algorithm[55] with k = 5. In order to prevent vectors with high missing rate from being considered "closer" to all the other vectors, we developed two distance methods in addition to Sklearn's weighted distance metric[57,68] (Eqs. (3) and (4)) and evaluated them to choose the best one.

$$SharedFeatures(X,Y) = NotNullFeatures(X) \cap NotNullFeatures(Y) \quad (3)$$

$$dist(X,Y) = \sqrt{m} \cdot \frac{||X[SharedFeatures(X,Y)] - Y[SharedFeatures(X,Y)]||}{\sqrt{|SharedFeatures(X,Y)|}} \quad (4)$$

The imputed value of the feature l in a patient with feature vector X is $\hat{x}_l = \frac{1}{k}\sum_{j=1}^{k} y_l^{j^*}$, where Y^{j^*} is the feature vector of its j-th nearest neighbor.

In the first distance metric we developed, "Mean Distance Penalty" (Eq. (6)), we added a penalty to the distance calculation for each feature that is missing in either vector (Eq. (5)). Define $penalty_f$ as the mean square distance calculated between non-missing values of feature f. For efficiency, we used in the computation 10% of the non-missing values of the feature, sampled from evenly-spaced quantiles of the feature.

$$g_f(X,Y) = \begin{cases} (x_f - y_f)^2 & if\ x\ is\ not\ null\ AND\ y\ is\ not\ null \\ penalty_f & else \end{cases} \quad (5)$$

$$dist_{pen}(X,Y) = \sqrt{\sum_{l=1}^{m} g_l(X,Y)} \quad (6)$$

In the second method introduced, named "Normalization by Count of Shared Features" (Eq. (7)), we normalized the default distance method by the number of not-null feature values shared by the two vectors instead of normalizing by the squared root of this number, as follows:

$$dist_{norm}(X,Y) = \frac{||X[SharedFeatures(X,Y)] - Y[SharedFeatures(X,Y)]||}{|SharedFeatures(X,Y)|} \quad (7)$$

This gives more weight to the number of not-null values than the default method.

We then evaluated the effect on the model performance and the running time of each distance method (Supplementary Table 9). Based on these results, we chose the third distance function.

Removal of correlated features

The creation of multiple time-series features in different timeframes, as well as the collection of a variety of lab measurements and vital signs that reflect the same trends in patients' medical condition, created feature redundancy. Two different methods were developed to deal with this problem, using clustering.

In the first method, we clustered features based on the type of original measurement they were derived from (e.g., all time-series features derived from heart rate measurements) and filtered only n_{keep} features from each cluster that had the best p-value for association with the target ($n_{keep} \in [1,2,3,4,5]$).

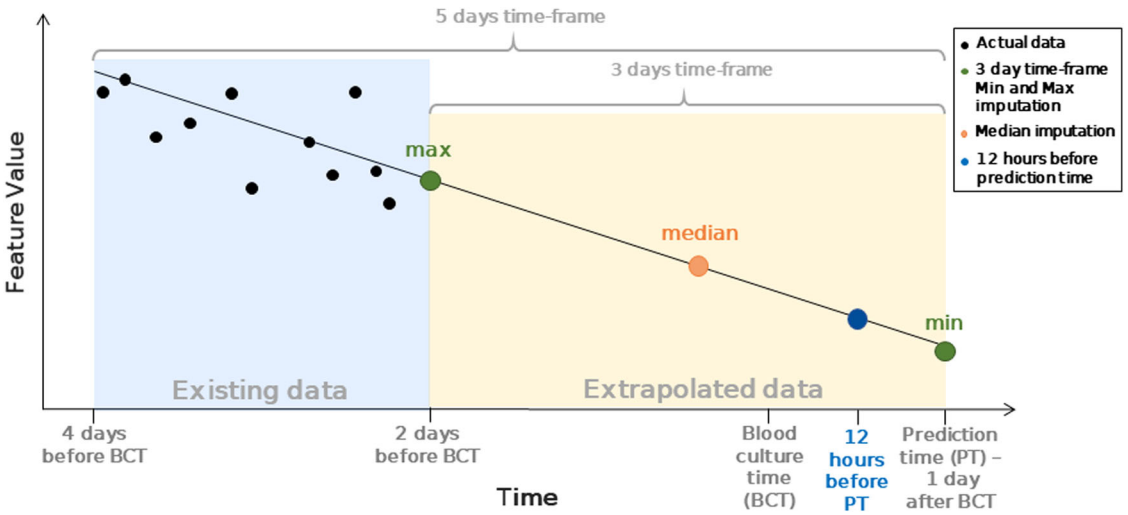

Fig. 8 | Illustration of the imputation scheme for time-series features. An example of existing feature values for a patient with missing data in the 3-day timeframe (yellow), the values of max, median, min, and at 12 h before prediction time are imputed using linear regression calculated based on existing values (black dots) in the 2 days before the beginning of the timeframe (light blue).

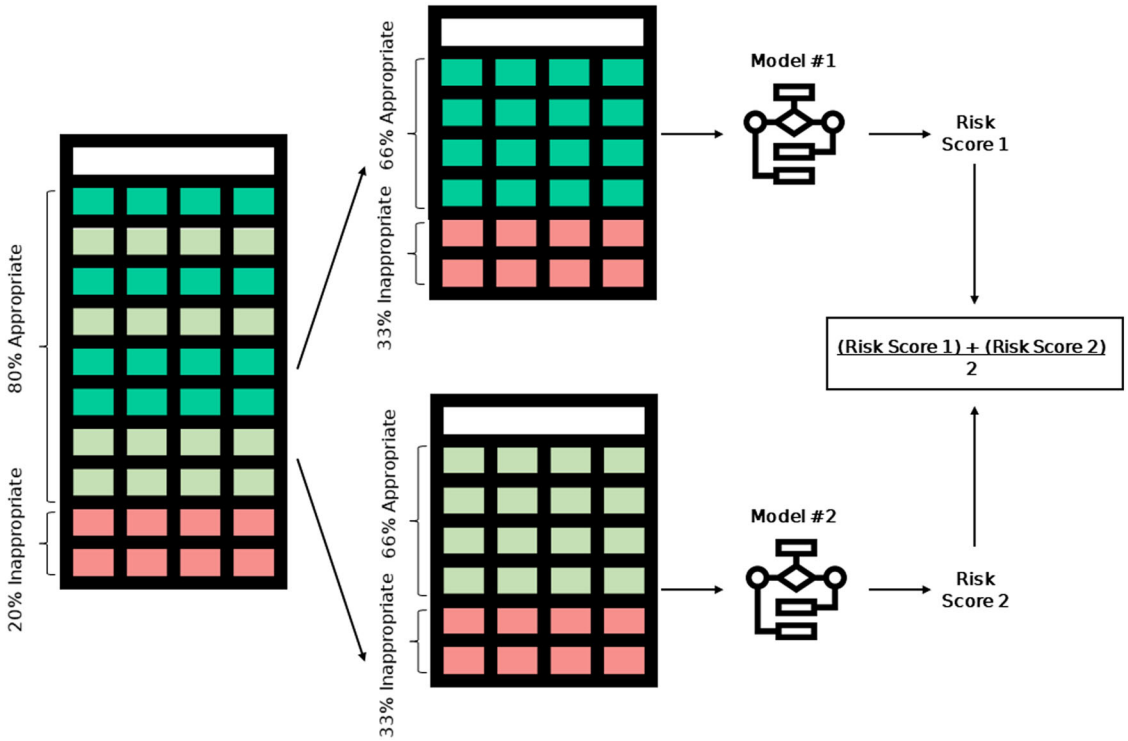

Fig. 9 | Illustration of the "DataEnsemble" model. On the left is the original dataset, rows of inpatients who received an appropriate antibiotic treatment (negatives) are colored in shades of green, and rows of inpatients who had received inappropriate treatment (positives) are colored in red. On the right are two subsets of the data, each containing all the positive patients, and a random, disjoint subset of the negative patients. The "DataEnsemble" is composed of identical models, each trained on a different subset of the data.

In the second method, we filtered out features with high correlation to other features. A correlation matrix C of all the features was created and transformed into a distance matrix $M_{ij} = 1 - C_{ij}$. This matrix M was then used for hierarchical clustering in which the final clusters were formed such that no two features in the cluster had a cophenetic distance greater than 1 minus a correlation threshold. The correlation thresholds 0.55, 0.6, 0.65, 0.7, 0.75, 0.8 were tested and 0.7 was chosen. Out of each cluster, only the feature with the best p-value for association with the target was kept. After comparing the effect of those parameters on the model's performance, we kept only one feature per each of the raw features ($n_{keep} = 1$).

Feature selection

Four methods of feature selection were evaluated. The first method is Recursive Feature Elimination with Cross Validation (RFECV)[56,57]. The second method utilizes the model's default feature importance method and selects only features with importance higher than the mean importance of all features. The third method is filtration of K features with the best Shap values[31]. The fourth method selects the K features with the highest mutual information score with the target. The K values 20, 25, 30, 35, 40, 45, 50 were evaluated for those two latter methods, and the best value K = 45 was selected.

In order to increase the robustness of feature selection, we summarized the results of the four feature selection methods tested in two ways. First, we checked the union of all the features selected by the methods. Second, we chose only the features that were selected by at least two of the four methods. The union of the features selected by all four methods using yielded the best results.

Data balancing

Three different methods for oversampling the training set data were tested. The first two methods, *ADASYN*[58] and *BorderlineSMOTE*[60] generate synthetic data mostly based on the "most difficult" samples for learning, and they both assume that all features are continuous. The third method, *SMOTENC*[59] distinguishes between continuous and categorical features and samples those features accordingly. Moreover, for each method, we tested different balancing ratios for the inappropriate treatment class, which was the smaller class, taking ratios of 0.3, 0.35, 0.4, 0.45 and 0.5.

Furthermore, we developed an ensemble model (*DataEnsemble*) that is composed of two instances of the same model trained on all positive samples and a different, disjoint subset of negative samples. Therefore, each model in the ensemble is trained on a proportion of 1:2 for positive compared to negative patients. The risk score of this model is the average score of the two models in the ensemble (Fig. 9). After evaluating all those methods, *BorderlineSMOTE* with a balancing ratio of 0.3 and utilization of DataEnsemble model were chosen for data balancing.

Another possible way to handle class imbalance is using class weights. Class weights adjust the loss function of the model to penalize the misclassification of the minority more heavily than those of the majority class, thus improving the model's learning process on the minority class. We evaluated different forms to allocate a high weight to the positive class (Supplementary Table 11) but did not obtain any substantial enhancement in the performance of the model.

Model selection

In order to choose the best model possible for our data, eight different binary classification models were compared—Random Forest[61], AdaBoost[62], Logistic Regression[63], SVM[64], SGDclassifier[57], LightGBM[65], sklearn's Gradient Boosting Classifier[57,66] and Xgboost[67]. For each model we created a DataEnsemble model as described above.

Hyperparameter optimization

After choosing the best model and pipeline parameters using an exhaustive search over the parameter combinations (e.g., data normalization method, see "Model's pipeline"), we used grid search to evaluate the effect of different model hyperparameters (e.g., Random Forest's max depth) on the results of the model trained on each of the five iterations of 5-fold cross-validation on

the training set. We tried different parameter combinations (Supplementary Table 10) and chose the combination that yielded the best mean AUPRC results.

Temporal validation dataset

MIMIC-IV is a publicly available dataset of ICU patients from Beth Israel Deaconess Medical Center (Boston, MA, USA) spanning the years 2008 and 2019. Since MIMIC-IV is an extension of MIMIC-III, some patients may overlap and appear in both; however, due to the use of different patient identifiers between the two datasets, we have no way to identify such overlaps. Therefore, to collect likely new patients that fit the inclusion criteria for our study, we only utilized patients from MIMIC-IV whose "anchor year" group (a 3-year interval that represents the admission time of the patient) was not included in MIMIC-III (i.e., 2014–2016 and 2017–2019). To obtain additional patients from the anchor year group of 2011–2013, which partially overlaps with MIMIC-III (collected between 2001 and 2012), we extracted gender, age, and ethnicity information for each of our MIMIC-III patients and filtered out every patient from the 2011–2013 anchor year group that matched the same combination of gender, age, and ethnicity as a patient from our MIMIC-III cohort. The final MIMIC-IV cohort comprised 65 patients, with 55 receiving appropriate antibiotic treatment and 10 receiving inappropriate treatment.

External dataset

The RHCC dataset was extracted from two computerized database systems: iMDsoft Metavision, the electronic patient record system of the ICU, and Prometheus, the hospital's electronic patient record system. These databases encompass a wide range of information, including patient demographics, background conditions, chronic illnesses, medication details, vital signs, laboratory measurements, dosages and durations of all pharmacologic treatments, the timing of insertion and removal of various invasive devices, and a registry of all fatalities. It also includes records of any operations performed during the admission. From this cohort, 161 ICU patients met our inclusion criteria, with 106 receiving appropriate treatment and 55 receiving inappropriate treatment.

Mortality analysis

Mortality analysis was conducted on three distinct cohorts. The first, utilized for the training and validation of our predictive model, included 135 patients, all of whom presented with positive blood culture (BC) outcomes. The second cohort comprised 4319 patients from the MIMIC-III database who met the inclusion criteria of the study but had negative BC. The third cohort consisted of 126 patients from the publicly accessible eICU dataset who met the same criteria as the second cohort (i.e., the same inclusion criteria and negative BC results).

For the purpose of generating the necessary features for the eICU patient model, we identified and aligned each feature in MIMIC-III with its corresponding feature in the eICU database. To ascertain the accuracy of the feature matching across the datasets, we employed a two-sided *t*-test on all matched continuous features in order to check if they had dissimilar distributions. Any feature that had FDR-corrected p-value < 0.025 in the *t*-test was excluded, as the features were from two distinct distributions, and thus unsuitable for matching. Additionally, a manual assessment of the distribution of the retained features within the eICU and MIMIC-III datasets was performed to eliminate any incompatible pairs. Out of 86 features selected by the model trained on the MIMIC dataset, only the following seven did not have a match in eICU: 'Culture from MRSA Screen_existence', 'Enterococcus sp. - VANCOMYCIN - R_existence', 'Glucose, Urine_existence', 'Ketone existence', 'RBC, Urine_existence', 'pH, Urine_all_days_min_max_diff', 'pH, Urine_is_imputed'.

Subsequently, the final appropriateness model was applied to each cohort, giving a risk score for each patient. Patients were then divided into equal-size quintiles based on their risk scores, and the 30-day mortality rate was computed for each quintile. To empirically evaluate the significance of the mortality rate for the lowest quintile, we employed a permutation test: We counted the fraction of 1000 random permutations for which the average mortality obtained in the lowest quintile was equal to or lower than that in the real data. A similar test was done for the highest quintile.

Ethics

MIMIC-III and MIMIC-IV are free de-identified EHR datasets. MIMIC-III was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA)[23]. Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was de-identified. The collection of patient information and creation of the MIMIC-IV research resource was reviewed by the Institutional Review Board at the Beth Israel Deaconess Medical Center, which granted a waiver of informed consent and approved the data-sharing initiative[26].

The eICU Collaborative Research Database is a freely available multi-center database for critical care research[38]. The study is exempt from institutional review board approval regarding the use of this dataset due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2).

RHCC is a de-identified dataset of patients from the Rambam Health Care Campus. The study was approved by the Institutional Review Board of the RHCC, approval number 0092-20-RMB. All research procedures were followed by the ethical standards of the responsible committee for human experimentation and with the Helsinki Declaration. The need for informed consent was waived.

Data availability

The MIMIC-III, MIMIC-IV and eICU databases analyzed in this study are available on PhysioNet repository[69–71]. The RHCC datasets are available upon request from M.R. Due to regulatory and ethical restrictions aimed at protecting the privacy of the research population, that data is not publicly accessible.

Code availability

The code used for data processing and model development is available at https://github.com/Shamir-Lab/ABXAppropriatenessML.

Received: 22 August 2023; Accepted: 30 December 2024;
Published online: 06 February 2025

References

1. Bell, B. G., Schellevis, F., Stobberingh, E., Goossens, H. & Pringle, M. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infect. Dis.* **14**, 13 (2014).
2. Wall, S. Prevention of antibiotic resistance—an epidemiological scoping review to identify research categories and knowledge gaps. *Glob. Health Action* **12**, 1756191 (2019).
3. Laxminarayan, R. et al. Antibiotic resistance—the need for global solutions. *Lancet Infect. Dis.* **13**, 1057–1098 (2013).
4. Nathan, C. & Cars, O. Antibiotic resistance—problems, progress, and prospects. *N. Engl. J. Med.* **371**, 1761–1763 (2014).
5. Aslam, B. et al. Antibiotic resistance: a rundown of a global crisis. *Infect. Drug Resist.* **11**, 1645–1658 (2018).
6. Mendelson, M. Review: Role of antibiotic stewardship in extending the age of modern medicine. *South Afr. Med. J.* **105**, 414–419 (2015).
7. Niederman, M. S. Appropriate use of antimicrobial agents: challenges and strategies for improvement. *Crit. Care Med.* **31**, 608 (2003).
8. Thomson, R. B. & McElvania, E. Blood culture results reporting: how fast is your laboratory and is faster better? *J. Clin. Microbiol.* **56**, e01313–e01318 (2018).

9. Livermore, D. M. & Wain, J. Revolutionising bacteriology to improve treatment outcomes and antibiotic stewardship. *Infect. Chemother.* **45**, 1–10 (2013).
10. Kumar, A. Antimicrobial delay and outcome in severe sepsis. *Crit. Care Med.* **42**, e802 (2014).
11. Kumar, A. et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589 (2006).
12. Luyt, C.-E., Bréchot, N., Trouillet, J.-L. & Chastre, J. Antibiotic stewardship in the intensive care unit. *Crit. Care* **18**, 480 (2014).
13. Bassetti, M. et al. Systematic review of the impact of appropriate versus inappropriate initial antibiotic therapy on outcomes of patients with severe bacterial infections. *Int. J. Antimicrob. Agents* **56**, 106184 (2020).
14. Raman, G., Avendano, E., Berger, S. & Menon, V. Appropriate initial antibiotic therapy in hospitalized patients with gram-negative infections: systematic review and meta-analysis. *BMC Infect. Dis.* **15**, 395 (2015).
15. Vallés, J., Rello, J., Ochagavía, A., Garnacho, J. & Alcalá, M. A. Community-acquired bloodstream infection in critically ill adult patients: impact of shock and inappropriate antibiotic therapy on survival. *Chest* **123**, 1615–1624 (2003).
16. Tabak, Y. P. et al. Blood culture turnaround time in U.S. acute care hospitals and implications for laboratory process optimization. *J. Clin. Microbiol.* **56**, e00500–18 (2018).
17. American Thoracic Society; Infectious Diseases Society of America. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *Am. J. Respir. Crit. Care Med.* **171**, 388–416 (2005).
18. Rhodes, A. et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive Care Med.* **43**, 304–377 (2017).
19. McGuire, R. J. et al. A pragmatic machine learning model to predict carbapenem resistance. *Antimicrob. Agents Chemother.* **65**, e0006321 (2021).
20. Van Steenkiste, T. et al. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. *Artif. Intell. Med.* **97**, 38–43 (2019).
21. Kanjilal, S. et al. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Sci. Transl. Med.* **12**, eaay5067 (2020).
22. Corbin, C. K. et al. Personalized antibiograms for machine learning driven antibiotic selection. *Commun. Med.* **2**, 1–14 (2022).
23. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
24. Johnson, A., Pollard, T. & Mark, R. MIMIC-III clinical database (version 1.4). PhysioNet https://doi.org/10.13026/C2XW26 (2016).
25. Leekha, S., Terrell, C. L. & Edson, R. S. General principles of antimicrobial therapy. *Mayo Clin. Proc.* **86**, 156–167 (2011).
26. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
27. Gupta, M. et al. An extensive data processing pipeline for MIMIC-IV. *Proc. Mach. Learn. Res.* **193**, 311–325 (2022).
28. Bradley, J. et al. Predicting hospitalisation for heart failure and death in patients with, or at risk of, heart failure before first hospitalisation: a retrospective model development and external validation study. *Lancet Digit. Health* **4**, e445–e454 (2022).
29. Rasmy, L. et al. Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. *Lancet Digit. Health* **4**, e415–e425 (2022).
30. Goodman, K. E. et al. A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum β-lactamase–producing organism. *Clin. Infect. Dis.* **63**, 896–903 (2016).
31. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4766–4775 (2017).
32. Bilavsky, E., Yarden-Bilavsky, H., Ashkenazi, S. & Amir, J. C-reactive protein as a marker of serious bacterial infections in hospitalized febrile infants. *Acta Paediatr.* **98**, 1776–1780 (2009).
33. Rasmussen, N. H. & Rasmussen, L. N. Predictive value of white blood cell count and differential cell count to bacterial infections in children. *Acta Paediatr.* **71**, 775–778 (1982).
34. Brown, L., Shaw, T. & Wittlake, W. A. Does leucocytosis identify bacterial infections in febrile neonates presenting to the emergency department? *Emerg. Med. J.* **22**, 256–259 (2005).
35. Previsdomini, M., Gini, M., Cerutti, B., Dolina, M. & Perren, A. Predictors of positive blood cultures in critically ill patients: a retrospective evaluation. *Croat. Med. J.* **53**, 30–39 (2012).
36. Abdelkarim, O. A. et al. Impact of irrational use of antibiotics among patients in the intensive care unit on clinical outcomes in Sudan. *IDR* **16**, 7209–7217 (2023).
37. Ali, M. et al. Rational use of antibiotics in an intensive care unit: a retrospective study of the impact on clinical outcomes and mortality rate. *Infect. Drug Resist.* **12**, 493–499 (2019).
38. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
39. Vincent, J.-L. et al. International study of the prevalence and outcomes of infection in intensive care units. *JAMA* **302**, 2323–2329 (2009).
40. Tabah, A. et al. Characteristics and determinants of outcome of hospital-acquired bloodstream infections in intensive care units: the EUROBACT International Cohort Study. *Intensive Care Med.* **38**, 1930–1945 (2012).
41. Eickelberg, G., Sanchez-Pinto, L. N. & Luo, Y. Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults. *J. Biomed. Inf.* **109**, 103540 (2020).
42. Eickelberg, G., Sanchez-Pinto, L. N., Kline, A. S. & Luo, Y. Transportability of bacterial infection prediction models for critically ill patients. *J. Am. Med. Inform. Assoc.* **31**, 98–108 (2023).
43. Roimi, M. et al. Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. *Intensive Care Med.* **46**, 454–462 (2020).
44. Zoabi, Y. et al. Predicting bloodstream infection outcome using machine learning. *Sci. Rep.* **11**, 20101 (2021).
45. Oonsivilai, M. et al. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. *Wellcome Open Res.* **3**, 131 (2018).
46. Yelin, I. et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat. Med.* **25**, 1143–1152 (2019).
47. Müller, L. et al. A risk-based clinical decision support system for patient-specific antimicrobial therapy (iBiogram): design and retrospective analysis. *J. Med. Internet Res.* **23**, e23571 (2021).
48. Howard, A. et al. Antimicrobial learning systems: an implementation blueprint for artificial intelligence to tackle antimicrobial resistance. *Lancet Digit. Health* **6**, e79–e86 (2024).
49. MacFadden, D. R. et al. Utility of prior cultures in predicting antibiotic resistance of bloodstream infections due to gram-negative pathogens: a multicentre observational cohort study. *Clin. Microbiol. Infect.* **24**, 493–499 (2018).
50. Chatterjee, A. et al. Quantifying drivers of antibiotic resistance in humans: a systematic review. *Lancet Infect. Dis.* **18**, e368–e378 (2018).
51. Vazquez-Guillamet, M. C., Vazquez, R., Micek, S. T. & Kollef, M. H. Predicting resistance to piperacillin-tazobactam, cefepime and meropenem in septic patients with bloodstream infection due to gram-negative bacteria. *Clin. Infect. Dis.* **65**, 1607–1614 (2017).
52. Lewin-Epstein, O., Baruch, S., Hadany, L., Stein, G. Y. & Obolski, U. Predicting antibiotic resistance in hospitalized patients by applying

machine learning to electronic medical records. *Clin. Infect. Dis.* **72**, e848–e855 (2021).
53. Hernàndez-Carnerero, À. et al. Dimensionality reduction and ensemble of LSTMs for antimicrobial resistance prediction. *Artif. Intell. Med.* **138**, 102508 (2023).
54. Agarwal, M. & Larson, E. L. Risk of drug resistance in repeat gram-negative infections among patients with multiple hospitalizations. *J. Crit. Care* **43**, 260–264 (2018).
55. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
56. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
57. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. **12**, 2825–2830 (2011).
58. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *Proc. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328 (2008).
59. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
60. Han, H., Wang, W.-Y. & Mao, B.-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Proc. 2005 International Conference on Advances in Intelligent Computing* (eds. Huang, D.-S. et al.) 878–887 (Springer, 2005).
61. Breiman, L. *Random Forests*. http://Machinelearning202.Pbworks.Com (1999).
62. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
63. Cox, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B (Methodol.)* **20**, 215–242 (1958).
64. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
65. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In *Proc. 31st International Conference on Neural Information Processing Systems*, 3149–3157 (Curran Associates, Inc., 2017).
66. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
67. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (Association for Computing Machinery, 2016).
68. Dixon, J. K. Pattern recognition with partly missing data. *IEEE Trans. Syst. Man Cybern.* **9**, 617–621 (1979).
69. Johnson, A. et al. MIMIC-IV. PhysioNet https://doi.org/10.13026/6mm1-ek67 (2023).
70. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
71. Pollard, T. et al. eICU Collaborative Research Database. PhysioNet https://doi.org/10.13026/0PZC-DM64 (2019).
72. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).

Acknowledgements

We thank Sarah Amar, MD, for helpful inputs. The study was supported in part by the Israel Science Foundation (grant No. 3165/19, within the Israel Precision Medicine Partnership program, and grant No. 2206/22) and by the Tel Aviv University Center for AI and Data Science (TAD) to R.S. E.G., E.R. and D.C. are supported in part by fellowships from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. This work was carried out in partial fulfillment of the requirements for the Ph.D. degree of D.C. at the Blavatnik School of Computer Science, Tel Aviv University. All icons used in this paper are designed by Freepik and are available at https://www.flaticon.com/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

E.G., E.R., D.C., R.S. conceived and designed the analysis; E.G., E.R. performed the data analysis, model development and model evaluation; E.G., E.R., D.C., R.S. contributed to the study design; A.W., D.B. assisted in the evaluation of the clinical aspects and data interpretation; M.R., A.S. provided the external validation set and assisted in the model evaluation on it; E.G., E.R., D.C., R.S. wrote the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01426-9.

Correspondence and requests for materials should be addressed to Dan Coster or Ron Shamir.