Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Individualized antibiotic discontinuation support using machine learning: hospital- wide evaluation of gradient boosting and LSTM models

Ilesa Bose
Humanitas University
Anna Fancelli
IRCCS Humanitas Research Hospital
Giovanni Angelotti
Dalle Molle Institute for Artificial Intelligence Research, USI-SUPSI
Riccardo Barbieri
Politecnico di Milano
Davide Fiore Bavaro
Humanitas University
Linda Bussini
Humanitas University
Michele Bartoletti
Humanitas University
Elena Azzolini
Humanitas University
Maurizio Cecconi
Humanitas University
Massimiliano Greco
massimiliano.greco@hunimed.eu

Humanitas University

Research Article

Keywords: Antimicrobial stewardship, Antibiotic discontinuation, Machine learning, Clinical decision support, Electronic health records, XGBoost, Long Short-Term Memory (LSTM)

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11388287/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Individualized antibiotic discontinuation support using machine learning: hospital-wide evaluation of gradient boosting and LSTM models

Ilesa Bose[1], Anna Fancelli[2], Giovanni Angelotti[3], Riccardo Barbieri[4], Davide Fiore Bavaro[1,5], Linda Bussini[1,5], Michele Bartoletti[1,5], Elena Azzolini[1,6], Maurizio Cecconi[1,2], Massimiliano Greco[1,2]

Affiliations

1. Department of Biomedical Sciences, Humanitas University, Milan 20072, Italy.
2. Department of Anesthesia and Intensive Care, IRCCS Humanitas Research Hospital, Milan 20089, Italy.
3. Dalle Molle Institute for Artificial Intelligence, USI-SUPSI, Lugano, Switzerland
4. Department of Electronics, Informatics and Bioengineering, Politecnico di Milano, Milan, Italy
5. Infectious Diseases Unit - IRCCS Humanitas Research Hospital, Milan, Italy
6. IRCCS Humanitas Research Hospital, Via Manzoni 56, 20089 Milan, Italy

Corresponding

Massimiliano Greco

Department of Biomedical Sciences, Humanitas University, Milan 20072, Italy.

massimiliano.greco@hunimed.eu

ABSTRACT

Background: Prolonged antibiotic therapy contributes to antimicrobial resistance and may increase the risk of adverse effects in hospitalized patients. Timely antibiotic discontinuation is an important quality and safety goal; however, identifying the optimal time to stop treatment remains challenging. Machine learning models trained on routinely collected clinical data may provide data-driven decision support to identify patients in whom antibiotic discontinuation could be considered. This study developed and evaluated two machine learning approaches to predict antibiotic discontinuation in hospitalised, non-critically ill patients.

Methods: A retrospective cohort study was conducted using daily laboratory results and vital signs from hospitalized patients receiving antibiotic therapy at a large academic centre. Two models were developed: an XGBoost model operating on daily patient snapshots enriched with temporal features; and a Long Short-Term Memory (LSTM) neural network processing raw daily clinical sequence. Model performance was evaluated primarily using the Area Under the Precision-Recall Curve (AUPRC), alongside other metrics. Predictions were evaluated using a two-day anticipation window, with predictions considered to be correct when generated on the day of clinician-initiated antibiotic discontinuation or up to two days earlier. Model interpretability was assessed using SHAP (XGBoost) and Integrated Gradients (LSTM).

Results: The XGBoost model achieved an AUPRC of 0.579, while the LSTM achieved an AUPRC of 0.480 (prevalence-based baseline 0.385) Sensitivity was 41.2% for XGBoost and 19.6% for the LSTM, while precision was 58.5% and 55.5%, respectively. Feature importance analyses showed that infection-related features, such as C-reactive protein and temperature,

were among the most influential predictors in both models. Oxygen saturation was also highly influential, alongside data-availability patterns and other clinical variables.

Conclusions: Both models demonstrated the feasibility of identifying clinical patterns preceding clinician-initiated antibiotic discontinuation from routinely collected clinical data. The two models showed different performances across the evaluated metrics, with XGBoost achieving better performance than the LSTM. By identifying a subset of patients up to two days before clinician-initiated discontinuation, these models could support earlier clinical reassessment of ongoing antibiotic therapy. The models were designed as clinician-in-the-loop decision-support tools rather than autonomous systems. Prospective validation is needed to determine whether model-guided reassessment can safely facilitate earlier antibiotic discontinuation.

KEYWORDS: Antimicrobial stewardship; Antibiotic discontinuation; Machine learning; Clinical decision support; Electronic health records; XGBoost; Long Short-Term Memory (LSTM)

INTRODUCTION

Antibiotics are among the most frequently prescribed medications in hospitalized patients across both medical and surgical settings.[1] While their use is essential for the treatment of bacterial infections, determining the optimal duration of therapy remains a clinical challenge. Prolonged antibiotic administration contributes to antimicrobial resistance (AMR) and is associated with drug-related adverse events and prolonged hospitalization, with consequences for both patients and healthcare systems.[2] To promote appropriate antibiotic use, major international organizations, including the World Health Organization (WHO), the U.S. Center for Disease Control and Prevention (CDC), and the European Centre for Disease Prevention and Control (ECDC), have published antimicrobial stewardship (AMS) frameworks aimed at optimizing antibiotic prescribing.[3–5]

Despite these recommendations, substantial variability in antibiotic use across centres and regions remains. Bedside decisions on treatment durations are heavily influenced by multiple patient- and infection-specific factors, resulting in significant variations in antibiotic management.[6]

Machine learning (ML) has emerged as a promising approach for leveraging the rich information contained in Electronic Healthcare Records (EHR) data across a wide range of clinical prediction tasks, including antibiotic therapy optimization. A previous study showed that a LightGBM model applied to EHR data could support antibiotic de-escalation decisions.[7] Another study has explored supervised learning for antibiotic discontinuation, intravenous-to-oral transition, and early and late de-escalation. [8] These models, however, rely primarily on the stationarity of features and may not fully capture the dynamic nature of laboratory and vital data time trends.

Explicitly modelling these longitudinal trajectories may therefore provide additional information for identifying opportunities for earlier treatment reassessment.

Tree-based and sequential models offer complementary approaches to incorporating this temporal information. Extreme Gradient Boosting (XGBoost) is well suited to structured clinical data and can incorporate longitudinal information through engineered features, such as previous measurements and changes over time.[9]

In contrast, Long Short-Term Memory (LSTM) networks are designed to learn temporal dependencies directly from sequential data, maintaining a stable memory over extended periods, [10,11] LSTM architectures have been applied to several longitudinal clinical prediction tasks, including septic shock, surgical site infection, and inpatient mortality, [12–14] supporting their potential application to longitudinal antibiotic treatment data.

In this study, we developed and evaluated two supervised machine learning models to identify patients approaching antibiotic discontinuation and potentially enable earlier clinical reassessment in hospitalized non-critically ill patients. The models represent complementary strategies for modelling longitudinal clinical information: XGBoost, which incorporates temporal information through engineered features, and LSTM, which learns temporal dependencies directly from clinical sequences.

METHODS

We conducted a retrospective cohort study using EHR data from IRCCS Humanitas Research Hospital. The study received institutional review board approval number n. 47/24. In

accordance with local regulations and given the anonymized nature of the data, the requirement for informed consent was waived. This study has not been registered.

The TRIPOD guidelines were followed for reporting the study.

Study population

We included adult patients (≥18 years) admitted to Humanitas Research Hospital between January 2016 and December 2025, who received at least one antibiotic administration during their hospital stay.

Data processing

Patient data, including laboratory measurements and vital signs recorded during periods of antibiotic administration, were extracted from the EHR system. Only patients admitted to general wards were included, as the Emergency Department and the Intensive Care Unit (ICU) use a different EHR system, and their data were not available for this analysis. We defined an antibiotic course as a period of consecutive antibiotic administration exceeding 48 hours, regardless of antibiotic type.

Courses lasting less than 48 hours were excluded because they were considered prophylactic rather than therapeutic antibiotic use. To limit heterogeneity from unusually prolonged courses, courses lasting more than 90 days were excluded. Courses ending in ICU admission or death were retained until the occurrence of these events but were considered ongoing rather than completed, because antibiotic discontinuation could not be attributed to a clinical decision to stop therapy. Different antibiotic courses from the same patient were treated as independent observations.

125 Antibiotic courses for which no vital signs or laboratory measurement were recorded

126 throughout the entire course were excluded.

FEATURES	REFERENCE (min-max)	Median-IQR First 24h	Median-IQR Last 24h	p-value
Total Bilirubin (mg/dL)	0.05 -1.1	0.80 [0.50-1.30]	0.70 [0.50-1.10]	<0.001
Creatinine (mg/dL)	0.6 -1.4	0.90 [0.68-1.33]	0.86 [0.65-1.24]	<0.001
Haematocrit (%)	39 -50.5	32.80 [28.40-37.60]	30.90 [27.050-35.20]	<0.001
Haemoglobin (g/L)	132.5-162.5	108.00 [94.00-125.00]	102.00 [91.00-117.00]	<0.001
Erythrocytes (10^{12} /L)	4.15-5.54	3.72 [3.19-4.27]	3.50 [3.07-4.02]	<0.001
Leukocytes (10^{9} /L)	4-10	9.54 [6.66-13.18]	7.73 [5.65-10.31]	<0.001
Absolute Lymphocytes (10^{9} /L)	1-4	1.00 [0.60-1.50]	1.30 [0.80-1.80]	<0.001
Absolute Neutrophiles (10^{9} /L)	2-7	7.40 [4.80-10.80]	5.30 [3.50-7.60]	<0.001
Platelets (10^{9} /L)	150-400	210.00 [148.00-286.00]	241.00 [163.00-334.00]	<0.001
Procalcitonin (ng/ml)	0.01-0.5	0.58 [0.17-2.66]	0.26 [0.11-0.74]	<0.001
C Reactive Protein (mg/L)	0-5	100.60 [41.70-169.60]	36.20 [14.50-84.90]	<0.001
Sodium (mmol/L)	135-145	139.00 [136.00-141.00]	140.00 [138.00-142.00]	<0.001

Cardiac Frequency (bpm)	60-100	79.00 [64.00-97.00]	74.00 [65.00-89.00]	<0.001
Respiratory Frequency (bpm)	12-20	18.00 [18.00-19.00]	18.00 [18.00-19.00]	<0.001
SpO2 (%)	90-100	96.00 [94.00-97.00]	96.00 [95.00-98.00]	<0.001
PAD (mmHg)	60-80	69.00 [56.00-83.00]	70.00 [60.00-80.00]	0.015
PAS (mmHg)	90-140	132.00 [106.00-149.00]	130.00 [110.00-143.00]	<0.001
Temperature (°C)	35-38	36.20 [36.00-37.30]	36.10 [36.00-36.60]	<0.001
Diuresis (ml)	500-2500	1300.00 [700.00-2200.00]	1500.00 [850.00-2300.00]	<0.001
Oxygen (L/min)	0-0	3.00 [2.00-6.00]	2.00 [2.00-4.00]	<0.001
Pain	0-10	0.00 [0.00–0.0]	0.00 [0.00–0.00]	<0.001

Table 1: Input features used for model training and their associated reference minimum and maximum values applied during preprocessing. The features ranges were selected to reflect clinically acceptable values.

A longitudinal dataset was constructed for both models. For each patient admission and antibiotic course, daily laboratory measurements and vital signs were summarized using the value with greatest deviation from the midpoint of the corresponding reference range. The clinical features included in the study, along with their corresponding reference values, are

listed in **Table 1**. Binary missingness indicators were added as features for both models to capture potentially informative patterns of data availability. The dataset also included demographic characteristics (age and sex), antibiotic escalation/de-escalation status, and hospital admission unit.

Before being used as inputs for the ML models, data were preprocessed using one-hot encoding for categorical variables and imputation of missing values. The full details of the data preprocessing, including imputation strategy, and the definitions of escalation and de-escalation are provided in the **Supplementary Material.**

Model input data

The XGBoost input comprised a three-day lag window and delta features representing changes between consecutive days. Specifically, for all time-dependent features, measurements from the current day (day *t*) and the two preceding days (days *t–1* and *t–2*) were included. Delta features were calculated as differences between consecutive observations within the lag window:

$$\Delta X_t = X_t - X_{t-1}.$$

In contrast, the LSTM model received the complete time-ordered sequence of daily features from the first day through the last available day of each antibiotic course. For both models, each day was labelled as 'continue antibiotic' when treatment continued on the following day $t + 1$ and "stop antibiotic" if no further antibiotic was administered after day t, except for the courses ending with an ICU admission or death. Thus, the "stop antibiotic" label identified the final day of a completed antibiotic course and corresponded to antibiotic discontinuation after day t. Because only the last day of each completed course was labelled "stop

antibiotic", the dataset was substantially imbalanced ('Continue Label': 89.2%; 'Stop Label': 10.8%).

Antibiotic courses from patients admitted before 2025 were divided at the patient level into training (90%) and validation (10%) sets. To limit further class imbalance arising from long antibiotic courses, the courses in these sets were limited to a maximum duration of 20 days. Antibiotic courses from patients admitted in 2025 were used as an independent temporal test. In the test set, courses lasting more than 20 days were truncated to the first 20 days to allow the same models to be applied consistently in a real-world setting.

Model development

Two supervised ML models were developed. The XGBoost model was implemented using scikit-learn. To address class imbalance, class weights were applied during training. The model used a three-day lag window (day t, day $t-1$, day $t-2$) as input to generate a probability of antibiotic discontinuation for day $t+1$. Hyperparameters were optimized to maximize the AUPRC; further details are provided in the **Supplementary Material**.

The LSTM model was implemented using TensorFlow. The model received the longitudinal sequence of clinical features and generated, at each time point, a probability of antibiotic discontinuation after the current day. A causal architecture ensured that predictions at day *t* depended only on information available up to and including that day, thereby preventing the use of future information. The LSTM model consisted of a single LSTM layer, with its hyperparameters optimized on the validation set to maximize the AUPRC. The model was trained using a weighted binary cross-entropy loss to address class imbalance. Further details on the model architecture and hyperparameter optimization are provided in the **Supplementary Material**.

Threshold optimization

Model predictions were expressed as probabilities of antibiotic discontinuation. To obtain the binary "continue antibiotic" and "stop antibiotic" labels, a probability threshold was applied. The thresholding strategy differed between the two models.

For XGBoost, a single global threshold was determined on the validation set by maximizing the F1 score. In contrast, a day-specific thresholding strategy was used for the LSTM. For each day t of antibiotic therapy, the optimal threshold θ_t was determined by maximizing the F1 score in the validation dataset, using all observations from patients who were still receiving antibiotic therapy on day t (i.e., courses lasting at least t days). This procedure was performed independently for each day, resulting in a set of day-specific thresholds.

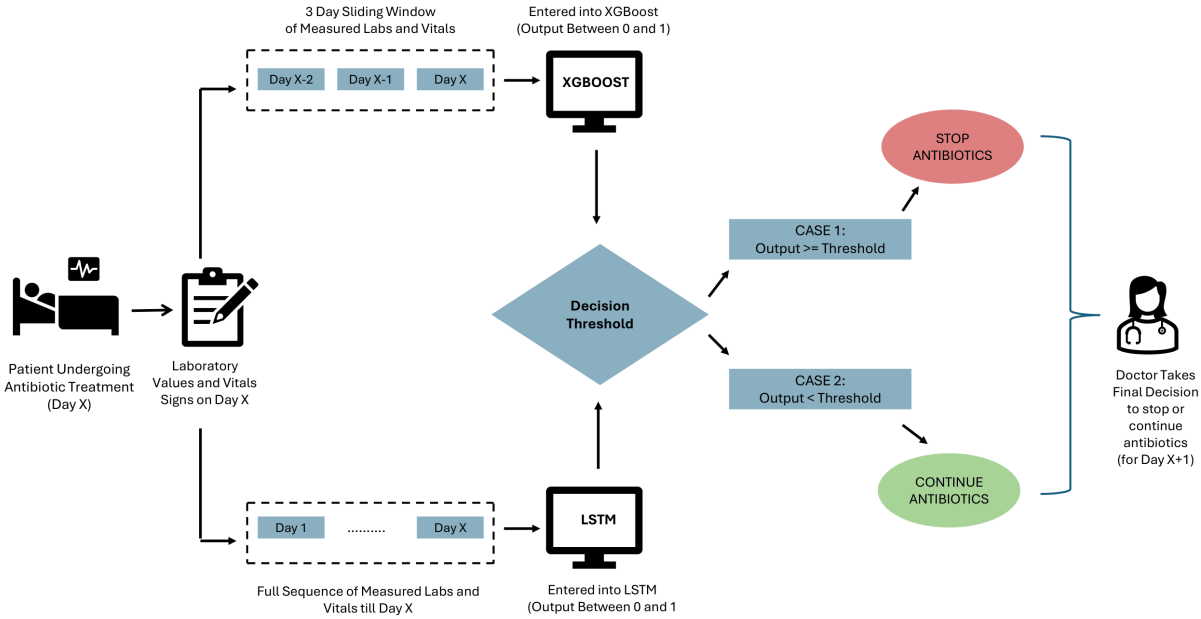


Figure 1: Overview of the study workflow. The schema illustrates the pipeline, including patient data collection and processing through the two models (XGBoost and LSTM) to generate daily prediction probabilities. These outputs are subsequently translated into

antibiotic discontinuation alerts intended to support clinical reassessment of ongoing therapy.

The optimized thresholds for each model are reported in the **Supplementary Material**. **Figure 1** illustrates the complete study pipeline, from EHR data extraction to model-generated predictions.

Model evaluation

Model performance was evaluated at two levels:

- Patient-day level: each day of an antibiotic course was treated as an independent observation.
- Antibiotic-course level: predictions were aggregated across an entire antibiotic course to produce a single outcome for each treatment episode. The first day on which the predicted probability exceeded the threshold was considered the model alert day for evaluation.

Because the intended use of the models was to support earlier reassessment of ongoing antibiotic therapy, an anticipation window was incorporated into the evaluation. A stop prediction was considered correct when generated on the day of clinician-initiated discontinuation or within the preceding two days.

Given the class imbalance, the primary performance metric was the AUPRC, which was compared with a prevalence-based baseline corresponding to the proportion of stop events in the test dataset. Additionally, recall, precision, and the area under the receiver operating characteristic curve (AUROC) were calculated**.**

For classification-based analyses, a true positive (TP) was defined as a stop prediction generated on the day of clinician-initiated discontinuation or one to two days earlier. A false negative (FN) occurred when no stop prediction was generated within this window. A false positive (FP) was defined as a stop prediction generated more than two days before clinician-initiated discontinuation. A true negative (TN) represented a correct continuation prediction when no discontinuation event occurred. Performance using exact-day matching (0-day window) and one- and two-day anticipation windows is reported in the **Supplementary Material**.

Feature importance and model interpretability

XGBoost feature importance was assessed using Shapley Additive exPlanations (SHAP), which quantifies the contribution of individual features to model predictions.

For LSTM, Integrated Gradients (IG) analysis was used to explain which input features influenced the prediction of the model over time. This method attributes changes in model output to individual features by comparing the observed input sequence with a baseline sequence in which each feature was set to its mean value. The global feature importance was obtained by averaging over 500 courses randomly sampled from the test set to reduce computational cost while ensuring stable estimates. Attribution scores were averaged across time steps to derive a global importance score for each feature.

Software and libraries

The predictive pipeline was implemented using Python (version 3.12.9). Data preprocessing was performed using Pandas (version 2.2.3) and NumPy (version 1.26.4). The LSTM was developed and trained using the Keras API integrated within TensorFlow (version 2.20.0).

Feature scaling was implemented using Scikit-Learn (version 1.6.1). Statistical analyses and performance metrics were conducted using Scikit-Learn and SciPy.

Data extraction, preprocessing, and model development were performed within the hospital's secure data analytics platform using Spark, SQL, and Python.

RESULTS

Cohort description

Following application of the inclusion criteria, 55,606 antibiotic courses from 39,268 hospitalized adult patients were included in the training and validation datasets. The population had a median age of 71 years (interquartile range [IQR], 60–80), and 41.82% were female. The median antibiotic course duration was 8 days (IQR, 5-11 days).

The test dataset, consisting exclusively of patients admitted in 2025, included a total of 5,347 patients contributing 6,984 antibiotic courses. The median age in the test population was 70 years (IQR, 60-79), 39.43% were female, and the median course duration was 7 days (IQR, 5 – 10 days).

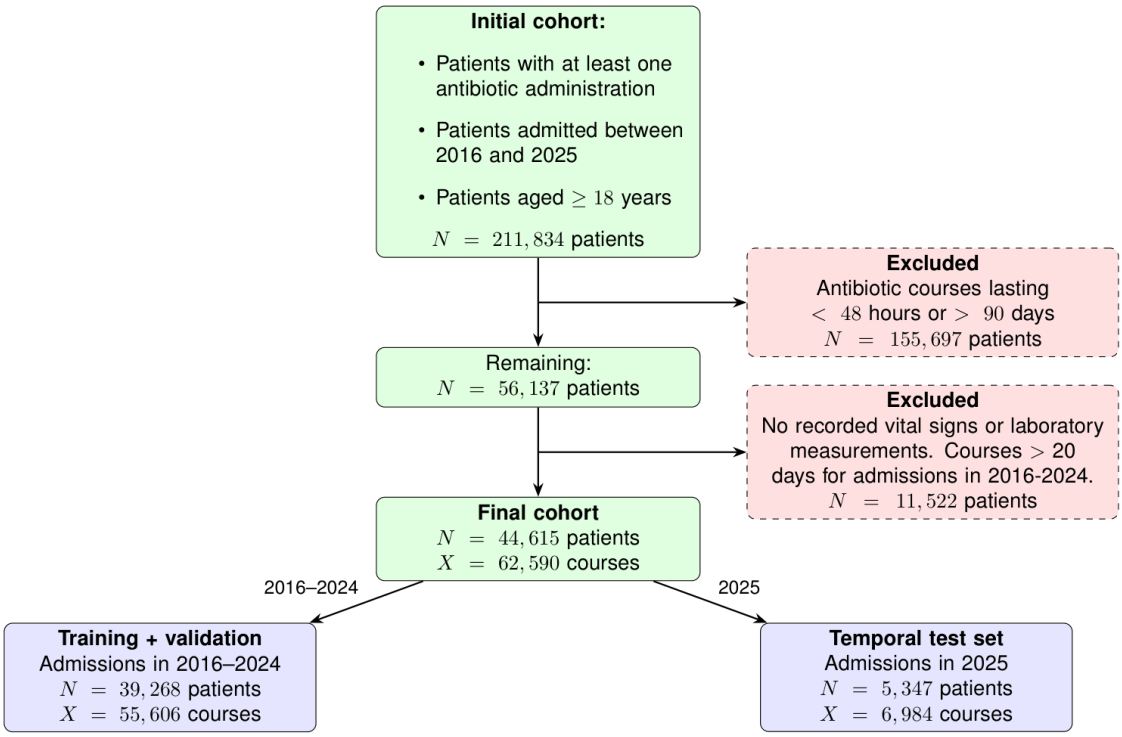


Figure 2: Cohort selection and dataset partitioning. The flow chart illustrates the patient selection process and the partitioning of the final cohort into training, validation and test datasets.

Figure 2 shows the cohort selection workflow, including patient exclusions at each stage and the final division into training/validation and test cohorts.

Predictive performance

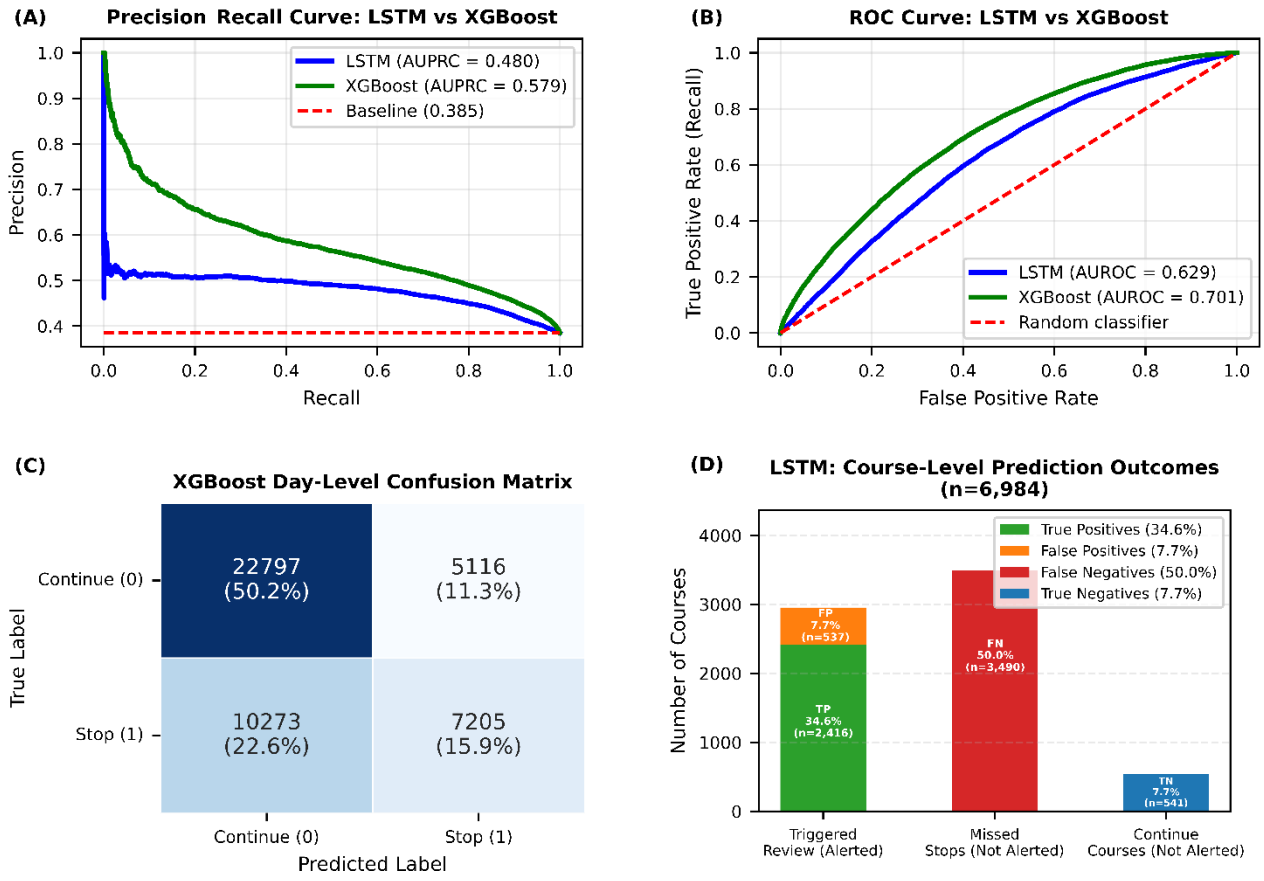


Figure 3 (A) Precision-Recall curves for XGBoost and LSTM models predicting clinician-initiated antibiotic discontinuation. The XGBoost model achieved an AUPRC of 0.579 while the LSTM model achieved 0.480; both exceeded the prevalence-based baseline of 0.385. (B) Receiver Operating Characteristic (ROC) curves for XGBoost and LSTM models predicting clinician-initiated antibiotic discontinuation. The XGBoost model achieved an AUROC of 0.701 while the LSTM model achieved an AUROC of 0.629; both exceeded the random classifier baseline (AUROC = 0.5). (C) Confusion matrix for the XGBoost model showing classification performance at the patient-day level. (D) LSTM model classification performance at the course level.

In the test set, the XGBoost model achieved an AUPRC of 0.579, compared with the prevalence-based baseline of 0.385. The LSTM achieved an AUPRC of 0.480, compared with

the same baseline (**Figure 3A)**. The corresponding AUROCs were 0.701 for XGBoost and 0.629 for LSTM (**Figure 3B**).

	XGBOOST			LSTM		
	TRAINING SET	**VALIDATION SET**	**TEST SET**	**TRAINING SET**	**VALIDATION SET**	**TEST SET**
Stop Prevalence	40.3%	40.0%	38.5%	40.3%	40.0%	38.5%
Sensitivity	43.3%	37.8%	41.2%	22.0%	21.9%	19.6%
Specificity	89.5%	88.3%	81.7%	95.7%	95.7%	90.1%
Precision	73.6%	68.4%	58.5%	77.5%	77.4%	55.5%
F1 Score	0.545	0.487	0.484	0.342	0.342	0.290
AUROC	0.783	0.747	0.701	0.763	0.756	0.629
AUPRC	0.709	0.654	0.579	0.683	0.673	0.480

Table 2: Performance of the XGBoost and LSTM models across training, validation, and test sets.

Model performance across the training, validation, and temporally held-out test cohorts is summarized in **Table 2.** Additionally, performance stratified by sex and age was comparable across the evaluated demographic groups for both models, with detailed results reported in the **Supplementary Material**.

Classification performance

Classification performance was assessed at the patient-day level for XGBoost and at the antibiotic-course level for the LSTM. Performance at the alternative evaluation level for each model is reported in the **Supplementary Material**.

The XGBoost model was evaluated at the patient-day level across 45,391 observations in the test set. The model correctly predicted continuation in 50.2% of patient-days (True Negatives, n=22,797) and generated a stop prediction within the two-day anticipation window in 15.9% (True Positives, n=7,205). Stop predictions generated more than two days before clinician-initiated discontinuation accounted for 11.3% of patient-days (False Positives, n=5,116). No stop prediction was generated within the anticipation window in the remaining 22.6% (False Negatives, n=10,273). The resulting confusion matrix is shown in **Figure 3C.**

The LSTM model was evaluated at the antibiotic-course level across 6,984 courses in the test set. Each course was classified according to the timing of the first generated stop alert relative to clinician-initiated discontinuation. The model generated at least one stop alert in 42.3% of courses (n=2,953). Of all courses, 34.6% (n=2,416) had an alert generated on the day of the clinician-initiated discontinuation or one to two days earlier (True Positives), while 7.7% (n=537) had an alert generated more than two days before discontinuation (False Positives). The model did not generate a stop alert in 57.7% of the courses (n = 4,031). Among these courses, 7.7% of the courses (n=541) had no clinician-initiated discontinuation and no stop alert (true negatives), while in the remaining 50.0% (n = 3,490) clinician-initiated discontinuation occurred without a preceding model alert within the anticipation window (False Negatives) **(Figure 3D)**.

Feature Importance

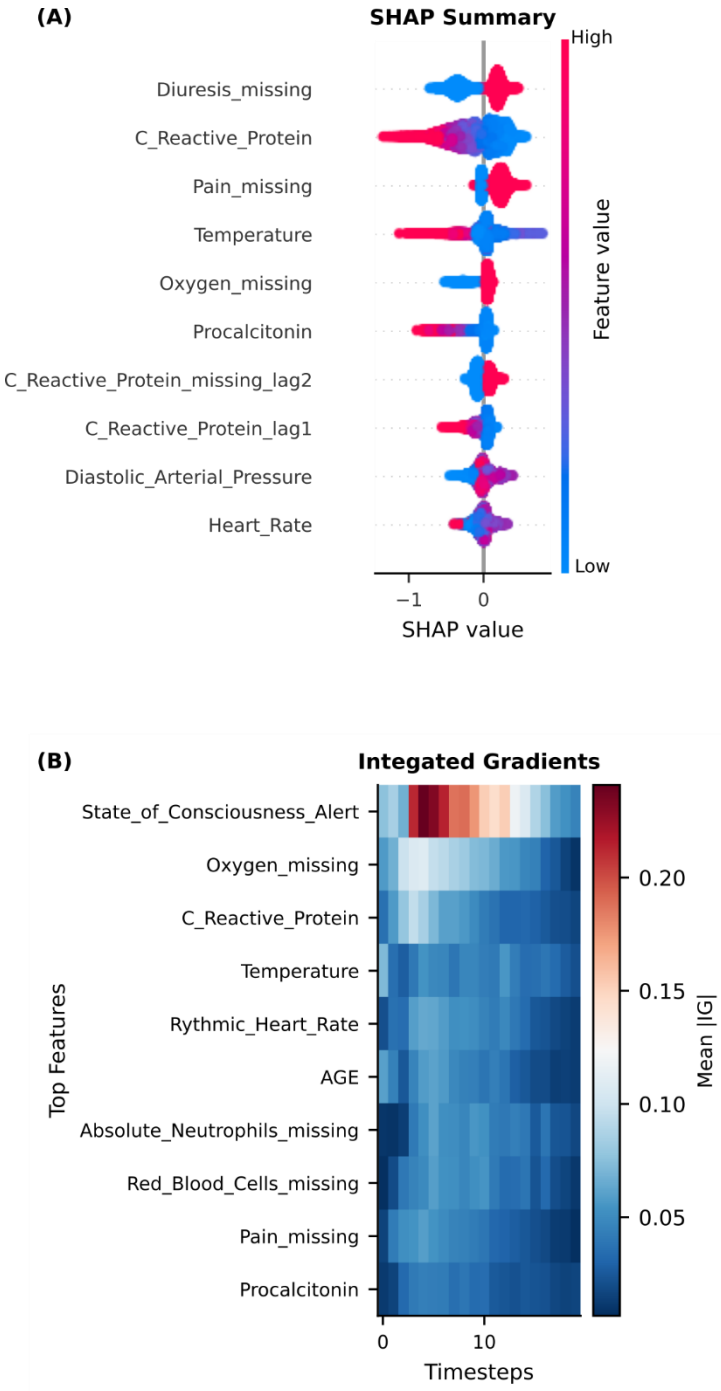


Figure 4 (A) SHAP-based feature importance for the XGBoost model. (B) Integrated Gradients temporal heatmap for the LSTM model. The 10 features with the highest overall attribution scores are shown for each model.

SHAP analysis was used to quantify the contribution of individual features to XGBoost predictions **(Figure 4A)**. The highest-ranking features included missingness indicators for pain and diuresis, as well as C-reactive protein (CRP) and body temperature. In particular, lower CRP and body temperature values, together with missing pain and diuresis measurements, contributed most strongly towards stop predictions.

For the LSTM model, Integrated Gradients analysis **(Figure 4B)** was used to quantify feature contributions across the antibiotic course. Alert neurological state was the highest-ranking feature, followed by the missingness indicator for oxygen measurements, CRP, and body temperature. Patient age was also among the features with the highest importance.

DISCUSSION

In this study, we developed and evaluated two machine learning approaches, XGBoost and LSTM, to identify patterns preceding antibiotic discontinuation in hospitalised, non-critically ill patients. The models were designed to identify potential discontinuation opportunities before the observed clinician-initiated stop and generate alerts that could support earlier clinical reassessment of ongoing antibiotic therapy. The two models processed longitudinal information differently. XGBoost used a three-day window with engineered temporal features, whereas the LSTM used the full treatment sequence. Given these differences, the results should not be interpreted as a strict head-to-head comparison, but rather as an evaluation of two complementary approaches to modelling longitudinal clinical data for antibiotic discontinuation.

Previous studies have demonstrated the potential of ML models for antimicrobial stewardship applications. Tabular machine learning models have been widely investigated

and integrated into clinical decision support systems (CDSS), including applications in antimicrobial resistance prediction, antimicrobial stewardship, and infection management.[15] Tam Tran-The et. al. used gradient boosting algorithms (XGBoost and LightGBM) to identify opportunities to reduce potentially inappropriate antibiotic use through intravenous-to-oral switching and early or late de-escalation. Their models incorporated clinical measurements, prescription records, and microbiological data as input features.[8] Our study extends this line of research by investigating early identification of antibiotic discontinuation using routinely collected EHR data from non-critically ill patients, including laboratory values and vital signs. Unlike stewardship strategies that rely on microbiological data or predefined intervention pathways, our approach focuses on identifying longitudinal clinical patterns from data available during routine clinical care. Although the exclusion of microbiological data may have reduced predictive performance, it may increase applicability to settings in which microbiological results are unavailable, or cultures have not been obtained.

In contrast to previous studies focusing primarily on tabular models, our work also evaluated an LSTM network, enabling assessment of a sequence-based approach to the same prediction task. Neural network applications in this context remain limited and are largely restricted to intensive care settings. For instance, Bolton et al. developed a recurrent neural autoencoder combined with a synthetic control approach using ICU data.[16] Their study estimated outcomes associated with different durations of antibiotic therapy and reported an association between earlier cessation and shorter hospital stay without increased mortality. The ICU setting is naturally well suited to sequence-based models such as LSTMs, given the high temporal resolution of clinical variables, which allows for explicit modelling of patient trajectories. Whether the same advantages extend to general wards, where measurements are less frequent and irregularly sampled, is less clear. In this study, the sequence-based

LSTM approach did not demonstrate superior performance across the evaluated metrics compared with the feature-engineered tabular XGBoost approach. Both models showed AUPRC values above the prevalence-based baseline of 0.385 (XGBoost: 0.579; LSTM: 0.480), indicating predictive information beyond the prevalence of discontinuation events alone. Because classification performance was assessed at different levels for the primary analyses of the two models, sensitivity and precision should not be interpreted as direct head-to-head comparisons. The higher discrimination achieved by XGBoost may indicate that relevant predictive information is concentrated in recent clinical measurements and their short-term changes, which is clinically plausible. Alternatively, the LSTM may have been unable to fully exploit longer temporal dependencies, or the relatively low and irregular temporal resolution of general-ward data may limit the benefit of sequence-based modelling. These models are intended as decision-support tools within a broader antimicrobial stewardship program, designed to prompt reassessment of ongoing therapy rather than recommend autonomous antibiotic discontinuation.

Feature importance analyses provided insight into the features contributing to the models' predictions. Several of the highly ranked features were clinically compatible with changes occurring during resolution of infection, although attribution methods identify associations with model predictions and should not be interpreted as evidence of causal or biological importance. For XGBoost, C-reactive protein (CRP) and body temperature were the most influential predictors. This is plausible as CRP typically rises during infection and declines with treatment. Lower CRP and temperature values contributed towards stop predictions, consistent with their established role as markers followed during clinical response to infection.[17] Procalcitonin (PCT), a well-established infection biomarker, emerged recently as one of the best daily recovery biomarkers, with a safe reduction in total antibiotic days for

daily PCT-guided protocol but not for daily CRP-guided protocol, according to prospective randomized evidence.[18,19]

Notably, missingness indicators were also highly influential, including those related to pain and diuresis measurements. Patterns of missingness in EHR data may encode aspects of clinical workflow and patient status; for example, measurement frequency may change as monitoring intensity changes during hospitalization. Consequently, these features may capture healthcare processes in addition to patient physiology.

For LSTM, the Integrated Gradients analysis identified partially overlapping predictors. CRP and temperature were again among the highly influential features, while oxygen-measurement missingness contributed substantially across the treatment sequence.[17] Alert neurological state was the highest-ranking feature, suggesting that neurological status contributed to the model's predictions. Altered mental status is incorporated into several clinical severity assessments, including those used in sepsis, although its attribution in our model should not be interpreted as evidence that improvement in consciousness independently determines readiness for antibiotic discontinuation.[20] Age also emerged as an important feature, consistent with its well-established association with disease severity and outcomes in infectious diseases. Overall, the importance of missingness indicators in both models also highlights a potential dependence on centre-specific documentation and monitoring practices, which may affect transportability to other institutions.

Limitations and Future work

This study has several limitations. As a single-centre EHR-based study, the findings may not fully generalise to other hospital settings or patient populations, where patient

406 characteristics, prescribing practices, monitoring frequency, and EHR documentation may
407 differ.

408 An important limitation is the definition of the prediction target. Antibiotic discontinuation
409 was derived from observed clinician behaviour rather than from an independent assessment
410 of the optimal duration of therapy. This distinction is important for antimicrobial stewardship,
411 where the goal is to support appropriate antibiotic use rather than reproduce current
412 prescribing practices. Consequently, a prediction occurring before the observed
413 discontinuation cannot establish that stopping antibiotics earlier would have been clinically
414 appropriate or safe. The two-day anticipation window should therefore be interpreted as a
415 window for prioritizing earlier clinical reassessment rather than as evidence supporting
416 earlier discontinuation itself.

417 Class imbalance represented an additional methodological challenge, because only the final
418 day of each completed antibiotic course was labelled as a stop event. In addition, long
419 antibiotic courses were underrepresented because courses exceeding 20 days were
420 excluded, limiting the models' ability to learn patterns associated with prolonged treatment
421 trajectories.

422 Additionally, unmeasured factors, including individual clinician prescribing preferences and
423 local stewardship practices, may influence discontinuation decisions but were not explicitly
424 captured by the models.

425 The development cohort spanned 2016–2024, while temporal testing was performed in
426 patients admitted in 2025. Although temporal testing provides a more stringent assessment
427 than a random split alone, clinical guidelines and prescribing behaviours may have evolved

throughout the development period, meaning that the models may partially reflect historical practices that differ from current decision-making patterns.

Finally, model explainability was assessed using different attribution methods: SHAP for XGBoost and Integrated Gradients for LSTM. While SHAP provides directionality along with magnitude for each feature, IG has some limitations for assessing feature importance, including its dependence on a chosen baseline and the fact that averaging attributions across time can hide temporal differences in feature contributions. Therefore, the LSTM feature-importance results should be interpreted with caution.

Future work should focus on external validation across different healthcare settings, incorporation of additional clinical and microbiological information, and prospective evaluation of model-generated alerts. Future model could also consider the appropriateness of antibiotic discontinuation in different way. Reinforcement learning, in which model training is guided by a reward or loss function, and causal inference approaches that incorporate expert knowledge or guideline-based recommendations into causal relationships could further enhance the capabilities of these models. Ultimately, interventional studies will be required to determine whether incorporating these predictions into antimicrobial stewardship workflows can safely prompt earlier reassessment and reduce unnecessary antibiotic exposure.

CONCLUSIONS

In summary, this study demonstrates that both XGBoost and LSTM models can identify clinical patterns preceding antibiotic discontinuation using routinely collected hospital data, with the potential to provide early signals up to 2 days before clinician-initiated

discontinuation. These models may support clinical decision-making by generating alerts that identify patients who may benefit from earlier reassessment of ongoing antibiotic therapy.

Model interpretability analyses also provided insight into the clinical features contributing to predictions, including markers related to infection resolution and patterns of clinical monitoring. Integrated within antimicrobial stewardship programmes and infectious disease services, these tools could support earlier treatment reassessment and, if prospectively validated, reduce unnecessary antibiotic exposure and its associated harms. External and prospective validation, followed by evaluation within clinical decision-support workflows, will be essential to determine their safety, generalisability, and clinical impact.

LIST OF ABBREVIATONS

XGBoost: *Extreme Gradient Boosting*

LSTM: *Long Short Trem Memory*

AUPRC: *Area Under Precision Recall Curve*

AMR: *Antimicrobial Resistance*

WHO: *World Health Organization*

CDC: *Center for Disease Control and Prevention*

ECDC: *European Center for Disease Control and Prevention*

AMS: *Antimicrobial Stewardship*

ML: *Machine Learning*

EHR: *Electronic Health Records*

ICU: *Intensive Care Unit*

AUROC: *Area Under Receiver Operating Characteristics Curve*

TP: *True Positives*

FP: *False Positives*

FN: *False Negatives*

TN: *True Negatives*

SHAP: *Shapley Additive exPlanations*

IG: *Integrated Gradients*

IQR: *Interquartile Range*

ROC: *Receiver Operative Curve*

CRP: *C-Reactive Protein*

CDSS: *Clinical Decision Support Systems*

PCT: *Procalcitonin*

DECLARATIONS

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study protocol was reviewed and approved by the IRB Comitato Etico Territoriale Lombardia 5 (CET 5), Humanitas Research Hospital, (Approval No. 47/24). The requirement for informed consent was waived by the committee due to the retrospective nature of the study.

CONSENT FOR PUBLICATION

Not Applicable.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

The code used in this study is available in the following GitHub repository: https://github.com/ilesabose/Antibiotic_Discontinuation

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was funded by 5x1000 Ministero della Salute AF2022, and by National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

AUTHOR CONTRIBUTING STATEMENT

I.B. curated and prepared the data, developed the models, and interpreted the results. A.F. and G.A. provided senior data analysis and modelling advise throughout the project. M.G. and

D.F.B. conceived the initial research protocol and supervised the project and the interpretation of results. D.F.B. contributed to the conception and design of the study. D.F.B., L.B. and M.B. additionally provided expertise in infectious disease and antimicrobial therapy. I.B., wrote the original manuscript draft and prepared the figures. A.F, G.A., D.F.B., E.A., L.B. contributed extensively to manuscript review and revision. R.B., M.B., M.C, M.G. reviewed and commented on the manuscript. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Pierandrea Morandini, Stefano Accornero, Edoardo Ottino, Marta Ranzini and Vignesh Kembu for their valuable contributions to data collection, curation, and clinical discussion throughout this study. Their expertise and input were instrumental in shaping the clinical framing of this work.

This work was funded by 5x1000 Ministero della Salute AF2022, and by the National Plan for NRRP Complementary Investments (PNC) in the call for the funding of research initiatives for technologies and innovative trajectories in the health - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM).

REFERENCE

1. Chen, R. *et al.* Global antibiotic prescription practices in hospitals and associated factors: a systematic review and meta-analysis. *J. Glob. Health* **15**, 04023 (2025).

2. Ryu, J. *et al.* Impact of antibiotic changes on hospital stay and treatment duration in community-acquired pneumonia. *Sci. Rep.* **14**, 22669 (2024).

3. WHO - AWaRe classification of antibiotics for evaluation and monitoring of use, 2023.

4. CDC - Core Elements of Antibiotic Stewardship.

5. ECDC - Antimicrobial Stewardship.

6. Jamil, E. *et al.* Global variation in antibiotic prescribing guidelines and the implications for decreasing AMR in the future. *Front. Pharmacol.* **16**, 1600787 (2025).

7. Haredasht, F. N. *et al.* Enhancing Antibiotic Stewardship: A Machine Learning Approach to Predicting Antibiotic Resistance in Inpatient Care. *AMIA Annu. Symp. Proc.* **2024**, 857–864 (2024).

8. Tran-The, T. *et al.* Development of machine learning algorithms for scaling-up antibiotic stewardship. *Int. J. Med. Inform.* **181**, 105300 (2024).

9. Chen, T. & Guestrin, C. XGBoost. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.

10. Elman, J. L. Finding Structure in Time. *Cogn. Sci.* **14**, 179–211 (1990).

11. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

12. Fagerström, J., Bång, M., Wilhelms, D. & Chew, M. S. LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock. *Sci. Rep.* **9**, 15132 (2019).

13. Kiser, A. C., Eilbeck, K. & Bucher, B. T. Developing an LSTM Model to Identify Surgical Site Infections using Electronic Healthcare Records. *AMIA Jt. Summits Transl. Sci. Proc.* **2023**, 330–339 (2023).

14. Wang, H.-Y. *et al.* Predicting Inpatient Deterioration in General Wards Using LSTM. in *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)* 52–58 (ACM, New York, NY, USA, 2024). doi:10.1145/3696271.3696280.

15. AlGain, S. *et al.* Can we rely on artificial intelligence to guide antimicrobial therapy? A systematic literature review. *Antimicrobial stewardship & healthcare epidemiology : ASHE* **5**, e90 (2025).

16. Bolton, W. J. *et al.* Machine learning and synthetic outcome estimation for individualised antimicrobial cessation. *Front. Digit. Health* **4**, 997219 (2022).

17. Póvoa, P. *et al.* C-reactive protein as a marker of infection in critically ill patients. *Clinical Microbiology and Infection* **11**, 101–108 (2005).

18. Lee, H. Procalcitonin as a biomarker of infectious diseases. *Korean J. Intern. Med.* **28**, 285–91 (2013).

19. Dark, P. *et al.* Biomarker-Guided Antibiotic Duration for Hospitalized Patients With Suspected Sepsis. *JAMA* **333**, 682 (2025).

20. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- XGBoostLSTMsupplementary.docx