Personalized Antibiotic Regimen Prediction for Multidrug-Resistant Escherichia coli Infections Using Machine Learning

Chanbormei Suy*, Prompong Sugunnasil†
*CMKL University
Bangkok, Thailand
csuy@cmkl.ac.th
†Chiang Mai University
Chiang Mai, Thailand
prompong.sugunnasil@cmu.ac.th

***Abstract*—Multidrug-resistant (MDR) Escherichia coli has emerged as a leading cause of urinary tract infections (UTIs), complicating treatment strategies and contributing to increased recurrence rates, prolonged hospital stays, and higher healthcare costs. While previous studies have explored the impact of individual antibiotics, there remains a significant gap in optimizing combination regimens, dosages, and administration frequencies to improve clinical outcomes and reduce resistance development. This study proposes a data-driven framework using longitudinal patient data from the MIMIC-III database to identify effective antibiotic regimens for MDR E. coli UTIs. Employing machine learning models such as random forest, LightGBM, ensemble, and Pytorch, we aim to develop personalized antibiotic treatment recommendations based on patient demographics and treatment history. By integrating microbiology reports, medication administration records, and demographic data, this research facilitates the creation of interpretable and clinically actionable models. Preliminary work includes dataset integration and feature extraction. The expected outcome is a predictive toolkit that enhances clinical decision-making in UTI management, ultimately helping curb the spread of MDR pathogens through optimized, personalized antibiotic regimens.**

***Index Terms*—Multidrug resistance, Escherichia coli, urinary tract infection (UTI), antibiotic optimization, machine learning, LightGBM, random forest, Pytorch, ensemble model, treatment prediction, recurrence modeling, MIMIC-III, longitudinal healthcare data, antimicrobial resistance, personalized medicine.**

I. Introduction

Urinary tract infections (UTIs) caused by multidrug-resistant (MDR) strains of Escherichia coli have become a growing public health concern worldwide. These infections not only complicate treatment strategies but are also associated with higher recurrence rates, extended hospitalizations, and increased healthcare costs. The traditional approach to antibiotic selection often relies on empirical therapy, which may not account for patient-specific factors or evolving resistance patterns, thereby contributing to treatment failures and the further propagation of antimicrobial resistance.

Recent advances in machine learning (ML) offer promising avenues for developing personalized treatment recommendations based on patient demographics, clinical history, and microbiological profiles. Predictive models trained on large-scale electronic health records (EHRs) can assist clinicians in selecting effective antibiotic regimens tailored to individual patients, thereby improving clinical outcomes while mitigating the risk of resistance development.

Despite growing interest in ML-driven solutions for antimicrobial stewardship, existing studies often focus narrowly on predicting resistance to single antibiotics rather than optimizing complex treatment regimens. Furthermore, many models overlook the longitudinal aspects of treatment history and fail to incorporate recurrence risks into their predictions. There remains a critical need for interpretable, data-driven frameworks that integrate microbiological data, demographic variables, and treatment trajectories to inform personalized antibiotic recommendations.

This study addresses these gaps by leveraging the MIMIC-III clinical database to develop and evaluate machine learning models capable of predicting effective antibiotic combinations for MDR E. coli UTIs. Using demographic features such as gender and age group, combined with antibiotic susceptibility patterns, we apply a suite of models—including Random Forest, Stacked Ensemble, LightGBM, and a custom PyTorch neural network—to create clinically actionable decision-support tools. By systematically benchmarking these models and optimizing threshold settings for multi-label prediction, we aim to enhance the precision of empirical therapy and contribute to the broader effort of combating antimicrobial resistance.

II. Related Work

The growing threat of multidrug-resistant (MDR) Escherichia coli as a causative agent of urinary tract infections (UTIs) has prompted significant research interest in predictive modeling, resistance surveillance, and antibiotic regimen optimization. This section presents a review of relevant literature grouped into three key areas: (A) Machine learning for antibiotic resistance prediction, (B) Optimization of treatment

strategies for MDR infections, and (C) Personalized medicine and decision support systems for infectious diseases.

A. Machine Learning for Antibiotic Resistance Prediction

Numerous studies have employed machine learning (ML) approaches to predict antimicrobial resistance (AMR), leveraging electronic health record (EHR) data and microbiological profiles. Huang et al. [1] applied logistic regression and random forest classifiers to predict antibiotic resistance in hospitalized patients using demographic, laboratory, and treatment-related features. Their findings demonstrated that ensemble methods significantly outperformed baseline models, particularly in sensitivity-specificity trade-offs. However, their analysis was constrained to a binary resistance outcome (resistant vs. sensitive), and it lacked temporal context regarding antibiotic exposure and treatment history. Similarly, Su et al. [2] explored the use of support vector machines (SVMs) and gradient boosting algorithms to predict the presence of extended-spectrum -lactamase (ESBL)-producing E. coli strains. While their models achieved high classification accuracy, the study focused exclusively on microbiological phenotypes and did not incorporate treatment administration data or patient outcomes. Consequently, while informative for diagnostic microbiology, the study offered limited insight into how specific antibiotic regimens influence clinical recovery or recurrence. Nguyen et al. [3] extended this work by leveraging the MIMIC-III database to develop predictive models for antibiotic resistance in critically ill patients. Using gradient boosting and logistic regression, they aimed to identify resistance risks based on microbiological data and antimicrobial prescriptions. Although their integration of real-world EHR data marked a notable advancement, the scope of their model was limited to infection prediction rather than recovery or recurrence outcomes. Additionally, their work did not evaluate the role of antibiotic combinations or treatment timing in shaping patient outcomes. These studies underscore the growing interest in ML-based resistance prediction but reveal limitations in their temporal scope, clinical applicability, and regimen-level detail. To advance clinical utility, there is a need for models that capture real-time treatment dynamics and personalize recommendations based on patient and pathogen characteristics.

B. Optimization of Treatment Strategies for MDR Infections

Optimizing treatment strategies for MDR infections, particularly those caused by E. coli, remains a complex task. Tamma et al. [4] reviewed the efficacy of combination therapy for MDR gram-negative infections and concluded that dual or triple antibiotic regimens may enhance bacterial eradication and reduce resistance emergence. However, their review was primarily narrative and lacked empirical modeling or data-driven regimen optimization. Furthermore, the study did not address how demographic or longitudinal factors influence treatment efficacy, nor did it explore recurrence trends post-therapy. Rawson et al. [5] discussed the integration of artificial intelligence (AI) into antimicrobial stewardship programs. They highlighted how machine learning can identify patterns in large-scale hospital data to guide empirical therapy. Importantly, they noted the challenge of balancing model accuracy with interpretability—an essential requirement in high-stakes clinical decision-making. Despite offering promising conceptual frameworks, the study emphasized that few models are deployed in real-world practice due to the lack of transparency, explainability, and validated outcomes across diverse patient populations. In a related effort, Aldeyab et al. [6] proposed a real-time forecasting model to predict antimicrobial use and resistance trends in hospital settings. Their work employed autoregressive integrated moving average (ARIMA) models and regression techniques to track resistance emergence. While useful for institutional policy planning, their model did not provide actionable treatment recommendations at the individual patient level and did not integrate patient-specific treatment responses or recurrence events. These findings suggest that while data-driven approaches to treatment optimization are being explored, most existing works do not account for patient-level treatment trajectories, nor do they provide granular guidance on optimal dosing regimens, antibiotic combinations, or recurrence mitigation strategies.

C. Personalized Medicine and Clinical Decision Support Systems

Efforts to develop personalized treatment models for infectious diseases are gaining momentum. Berner and La Lande [7] offered a comprehensive overview of clinical decision support systems (CDSS), highlighting how they aid clinicians in making evidence-based decisions using structured rules or predictive algorithms. While CDSS has proven effective in reducing diagnostic errors and improving treatment adherence, many systems rely on rule-based engines that cannot adapt dynamically to patient-specific variations in treatment response or resistance profiles. Foxman [8] conducted a seminal epidemiological study on UTI recurrence, identifying demographic factors such as age, sex, and sexual activity as strong predictors of reinfection. Although her study did not involve predictive modeling or machine learning, it emphasized the importance of demographic and behavioral variables—features that are often overlooked in data-driven infection models. More recent studies have advocated for integrating such variables into personalized risk assessments, but few have extended this to regimen-level optimization. Moreover, Giannini et al. [9] proposed a probabilistic framework for modeling UTI treatment outcomes using Bayesian inference. Their model incorporated patient history, comorbidities, and drug resistance patterns, demonstrating strong performance in retrospective analyses. However, the model's complexity limited its real-time applicability, and its reliance on prior probabilities constrained its adaptability in dynamic treatment environments. These works demonstrate a growing awareness of the need for personalization in infectious disease treatment. However, significant challenges remain, particularly in integrating time-sensitive, regimen-level data with patient-specific features in an interpretable and clinically actionable way.

D. Research Gap

The current body of literature demonstrates promising progress in the application of machine learning to infectious disease modeling. However, three critical gaps remain: (1) the lack of longitudinal models that align antibiotic administration with microbiology outcomes over time; (2) limited personalization of antibiotic regimens based on individual patient history and demographics; and (3) minimal integration of recurrence prediction into treatment optimization frameworks. To address these challenges, this study proposes a machine learning–based approach using longitudinal data from MIMIC-III to optimize antibiotic regimens for MDR E. coli UTIs. By incorporating patient demographics, treatment history, and resistance profiles, and applying models such as random forest, LightGBM, and ensemble model, and Pytorch, we aim to deliver interpretable, personalized treatment recommendations that reduce recurrence and combat resistance development.

III. Materials and Methods

A. Dataset Description

This study utilizes data extracted from the MIMIC-III (Medical Information Mart for Intensive Care III) database, a large, publicly available critical care dataset comprising over 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III includes detailed, de-identified clinical information across multiple domains, including patient demographics, medication administration, microbiological tests, laboratory measurements, and hospital admission details. For the purpose of modeling effective antibiotic treatment for MDR *Escherichia coli* urinary tract infections (UTIs), the following core datasets from MIMIC-III were used:

- Microbiology Events (microbiologyevents.csv): This dataset contains results of microbiological tests, including isolated organisms, antibiotic names (`ab_name`), test interpretations (interpretation: Sensitive [S], Intermediate [I], or Resistant [R]), sample types, and test names. This dataset is crucial for identifying patient-specific antibiotic susceptibility profiles.
- Patient Demographics (patients.csv): This table includes basic demographic data such as `subject_id`, gender, and `anchor_age` (approximate patient age at admission). Non-essential fields, including `anchor_year`, `anchor_year_group`, and `dod` (date of death), were excluded from the analysis.

The microbiology data was filtered to isolate records for *Escherichia coli*, the focus pathogen for this study. From this filtered set, antibiotic susceptibility patterns for each patient were aggregated and later merged with corresponding demographic information using the `subject_id` field as a unique identifier.

B. Data Preparation and Feature Engineering

1) Dataset Integration: To create a unified dataset suitable for machine learning, the `microbiologyevents` and `patients` tables were joined using the `subject_id` key. Prior to merging, the microbiology data was filtered to include only records where the organism (`org_name`) was *Escherichia coli*. The resulting filtered dataset was aggregated by `subject_id` to consolidate all antibiotic test results into single patient records.

Key microbiology features captured through aggregation included:

- ab_name: Lists of antibiotics tested per patient.
- interpretation: Corresponding susceptibility interpretations (Sensitive - S, Intermediate - I, Resistant - R).
- spec_type_desc: Specimen types from which organisms were isolated (e.g., urine).
- test_name: Names of the susceptibility tests administered.

Two additional features were engineered to categorize antibiotics based on effectiveness:

- Effective Antibiotics: Antibiotics interpreted as Sensitive (S).
- Ineffective Antibiotics: Antibiotics interpreted as Resistant (R) or Intermediate (I).

Another feature, num_antibiotics_tested, quantified the total number of antibiotics each patient was evaluated against.

Following initial aggregation, the dataset was further refined by filtering out antibiotics tested fewer than five times across all patients. This step significantly reduced class imbalance, ensuring each antibiotic had sufficient representation in the dataset for reliable model training.

The microbiology dataset, now refined, was merged with patient demographic data including `gender` and `anchor_age`. The final integrated dataset was exported as `merged_microbiology_admissions_final.csv` for subsequent analysis and modeling.

2) Feature Engineering: Additional preprocessing steps were performed to prepare the dataset for supervised machine learning:

- Age Group Binning: The continuous variable `anchor_age` was categorized into five discrete bins:
  - 0–20
  - 21–40
  - 41–60
  - 61–80
  - 80+

  This binning captures potential non-linear age effects and simplifies model interpretability.
- Final Feature Set: After preprocessing, the final dataset consisted of:
  - `gender` (categorical)
  - `age_group` (ordinal categorical)
  - `effective_antibiotics` (multi-label categorical list)

  This refined feature set supported multi-label classification modeling aimed at predicting effective antibiotics for patients based on their demographic characteristics.

C. Modeling Approaches

In this project, several modeling strategies were implemented and evaluated for multi-label classification. These approaches range from tree-based ensemble methods to a custom deep learning model, each offering different strengths:

1) Random Forest (Baseline): As a baseline, a multi-output Random Forest classifier was trained to predict multiple labels simultaneously. The Random Forest serves as a robust starting point due to its ability to handle high-dimensional data and capture non-linear relationships [10]. It outputs a set of class probability estimates for each label, which can then be thresholded to obtain binary predictions. This baseline provided a reference performance level for subsequent, more complex models.

2) Stacked Ensemble Model: Beyond the single-model baseline, a more complex ensemble was developed using a stacking approach. In this stacked ensemble, two diverse tree-based classifiers (a Random Forest and an Extra Trees classifier) were used as base learners. Their predictions were combined and passed to a Logistic Regression meta-learner (final estimator), which learned how to best integrate the base learners' outputs. This StackingClassifier setup leverages the complementary strengths of the base learners and the meta-learner: the tree-based models reduce variance through bagging and randomization, while the Logistic Regression provides calibration by weighting each base model's contribution [11]. By aggregating the predictions of both base models, the ensemble produces multi-label probability outputs and was expected to improve performance over any individual classifier.

3) LightGBM Multi-Output Classifier: In addition to bagging-based ensembles, a gradient boosting approach was tested using LightGBM. Because LightGBM does not natively support multi-label classification, a MultiOutputClassifier wrapper was employed to train a separate LightGBM classifier for each label. LightGBM is known for its fast, memory-efficient training and ability to handle large datasets with many features [12]. Using one LightGBM model per label allows the system to capture complex patterns via boosting while still scaling efficiently. This approach yields probability predictions for each label independently (each LightGBM estimator outputs a probability for its specific label).

4) PyTorch Neural Network: Finally, a custom neural network model was implemented with PyTorch to explore a deep learning solution for the multi-label task. The neural network consists of multiple layers in a feed-forward architecture and produces an output neuron for each label with a sigmoid activation, outputting a probability between 0 and 1 for that label. The network was trained using a multi-label loss function (binary cross-entropy applied independently to each output) to optimize all labels simultaneously [13]. This model can potentially capture complex non-linear relationships and shared patterns across labels that may not be fully captured by the tree-based models.

D. Threshold Optimization and Model Evaluation

To ensure robust performance and fair comparisons, rigorous evaluation procedures were applied, including cross-validation, hyperparameter tuning, threshold optimization, and the use of multiple metrics. During model development, a 5-fold cross-validation (CV) strategy was used (where computationally feasible) to reliably estimate performance and guide hyperparameter tuning. For the tree-based models and LightGBM, `GridSearchCV` was employed over key hyperparameters (e.g., number of trees, maximum depth, learning rate) using the training folds, with model configurations selected based on the best cross-validated Macro F1 score. This process helped avoid overfitting to a single train–test split and ensured that each model's chosen parameters generalized well [**?**].

In addition to tuning model parameters, an optimal probability threshold for converting continuous outputs into binary predictions was determined for the probabilistic models. Rather than using the default 0.5 cutoff, threshold optimization was performed for the stacked ensemble and the neural network by analyzing precision–recall curves on a validation set. For each label, the probability threshold that maximized the F1 score (i.e., the point on the precision–recall curve where the F1 is highest) was selected. Using these per-label optimal thresholds improved the balance between precision and recall, thereby enhancing overall F1 performance. This step is particularly beneficial in an imbalanced multi-label scenario, as it customizes the decision boundary for each class to ensure no label is disproportionately favored or neglected.

Model performance was evaluated using several complementary metrics that capture different aspects of multi-label prediction quality:

- Macro F1 Score: This is the F1 score (harmonic mean of precision and recall) computed for each label independently and then averaged across all labels (treating each label equally). A higher macro F1 indicates better overall classification performance across all target antibiotics (labels).
- Jaccard Similarity: For multi-label outputs, the Jaccard similarity (also called Jaccard index) measures the overlap between the predicted label set and the true label set for each sample. It is defined as the size of the intersection divided by the size of the union of the two sets. We report the average Jaccard index across all samples, which reflects how closely, on average, the predicted label sets match the true label sets (with 1.0 indicating a perfect match).
- Hamming Loss: Hamming loss is the fraction of incorrect labels to the total number of labels. Equivalently, it is the proportion of labels that are misclassified (either a label predicted as positive when it is actually negative, or vice versa) out of all label instances. A lower Hamming loss (closer to 0) indicates better performance, meaning fewer errors on a per-label basis.

Together, these evaluation metrics provide a comprehensive assessment of model performance. By combining threshold

tuning with cross-validated hyperparameter optimization and multi-faceted metrics, the study ensured that the final models were both well-calibrated and rigorously evaluated.

1) Cross-Validation and Hyperparameter Tuning: To evaluate the generalization capability of the models and reduce the likelihood of overfitting, stratified 5-fold cross-validation was conducted. Hyperparameter optimization was performed using `GridSearchCV` over predefined parameter spaces:

- Random Forest: Number of trees, maximum depth, and minimum samples per split were tuned.
- Ensemble (Random Forest + ExtraTrees + Logistic Regression): CV and threshold tuning were used to maximize F1 score.
- LightGBM: Number of leaves, learning rate, maximum depth, and feature fraction were tuned. Early stopping based on validation loss prevented overfitting.
- PyTorch: Early stopping and class-weighted binary cross-entropy loss were applied to handle imbalance.

All models were evaluated on the same train–test splits for fair comparison.

2) Feature Importance and Model Explainability: Understanding the contribution of features is critical for clinical relevance:

- Random Forest: Feature importances based on impurity decrease.
- LightGBM: SHAP values and built-in feature importances.
- Ensemble: Indirect feature importance through meta-learner coefficients.
- PyTorch: Due to its black-box nature, feature importance was not directly interpretable but performance validation was emphasized.

Explainability techniques ensured that model predictions could be trusted and validated clinically.

E. Implementation Details

All experiments were conducted using Python (version 3.9) with libraries including Scikit-learn, LightGBM, PyTorch, and SHAP on APEX which is a supercomputer.

IV. Results and Comparative Analysis

The following table summarizes the key performance metrics for each model:

TABLE I
Model Performance Comparison

Model	Macro F1	Jaccard Similarity	Hamming Loss
Random Forest	0.540	0.830	0.100
Ensemble	0.568	0.646	0.281
LightGBM	0.540	0.830	0.100
PyTorch	0.568	0.646	0.285

Interpretation:

- Random Forest achieved strong baseline performance, especially with a high Jaccard Similarity (0.83) and low Hamming Loss (0.10), indicating good robustness for a relatively simple model.
- Ensemble model slightly improved Macro F1 (0.568) compared to Random Forest, suggesting that combining diverse tree-based models and Logistic Regression enhanced balanced performance across rare antibiotics. However, the Jaccard Similarity dropped to 0.646, indicating that while it predicted rare classes better, it was less precise at matching full label sets.
- LightGBM achieved identical performance to Random Forest (Macro F1: 0.54, Jaccard: 0.83, Hamming Loss: 0.10). Contrary to previous results, LightGBM did not outperform Random Forest, suggesting that on this specific dataset and feature set, complex boosting methods offered no additional gain.
- PyTorch achieved performance comparable to the Ensemble model, with slightly worse Hamming Loss (0.285 vs. 0.281), indicating that although it captured some rare labels better, it made more prediction errors overall. This suggests that the neural network was more sensitive to data noise or imbalance.

Overall, Random Forest and LightGBM emerged as the most balanced models with the highest Jaccard Similarity and lowest Hamming Loss, while Ensemble and PyTorch models achieved marginally better Macro F1 scores but with reduced precision in label prediction.

V. Discussion

The modeling experiments show that Random Forest and LightGBM provided the best overall balance between prediction quality and error rate, achieving identical performance across all evaluation metrics. Despite previous expectations, LightGBM did not outperform Random Forest. This suggests that for this relatively simple and small-featured dataset (only gender and age group), basic ensemble models are sufficient and that more sophisticated gradient boosting methods do not offer additional benefits. The Ensemble model (combining Random Forest, Extra Trees, and Logistic Regression) slightly improved the Macro F1 score by focusing better on rare antibiotics through threshold optimization, although at the cost of reduced Jaccard Similarity.The PyTorch neural network showed potential by achieving competitive Macro F1 scores but had higher Hamming Loss, indicating less stable performance. This highlights that deep learning models require richer feature sets and larger datasets to outperform simpler machine learning models.From a clinical perspective, the models demonstrate promise for aiding in personalized antibiotic selection based on basic demographics. Even with minimal features, the models produced meaningful recommendations, underlining the value of machine learning in clinical decision support.

VI. Limitations

Several limitations were identified during the course of this study:

- Feature Selection Constraint: Although the original `microbiologyevents` and `patients` datasets contained a wide variety of features (such as admission type, infection site, comorbidities, and laboratory results), the final modeling dataset was limited to only `gender` and `anchor_age` (transformed into `age_group`). This feature reduction was intentional for model simplicity and interpretability but significantly restricted the amount of information available to the machine learning models.
- Label Imbalance: Some antibiotics were rarely prescribed or rarely effective in the dataset, leading to highly imbalanced labels. This imbalance could bias models towards overpredicting more common antibiotics while ignoring rare but clinically important alternatives.
- Model Generalizability: Models were trained and tested on a single cohort filtered for *Escherichia coli* infections. As such, they may not generalize well to infections caused by other organisms or to populations in different hospital settings without retraining and external validation.
- Dataset Size for Deep Learning: Although tree-based models performed well, the PyTorch deep learning model likely underperformed due to the relatively small sample size and low feature dimensionality. Deep learning models generally require larger datasets and richer features to reach their full potential.
- Clinical Interpretability: While basic demographic features are easy to interpret, the absence of detailed clinical parameters means that model recommendations should be used cautiously and in conjunction with clinical judgment.

VII. Future Work

Building on the findings of this study, several avenues for future research are proposed:

- Expanded Feature Set: Future models should incorporate a wider range of clinical features available in the original datasets, such as comorbidities, laboratory results, prior antibiotic usage, and infection severity scores. This could improve predictive accuracy and clinical utility.
- Incorporation of Text Data: Natural Language Processing (NLP) techniques could be applied to unstructured notes or microbiology reports to extract additional predictive signals.
- Transfer Learning and Pretraining: Pretraining deep learning models on larger, related datasets and fine-tuning on the *Escherichia coli* subset could improve neural network performance.
- Multi-Center Validation: Expanding the dataset to include microbiology records from multiple hospitals would enhance model robustness and external validity.
- Real-Time Clinical Deployment: Development of a clinical decision support tool integrating the trained models could be piloted in hospital settings to assist infectious disease teams with early empirical therapy selection.
- Expert Clinical Review: Involving licensed physicians and infectious disease specialists to validate model recommendations in a clinical setting would be essential to ensure practical relevance, safety, and reliability before real-world deployment.

VIII. Conclusion

This study demonstrated that machine learning models, even when trained on minimal demographic features, can predict effective antibiotic regimens for Escherichia coli infections with high performance. Among the models tested, Random Forest and LightGBM exhibited the best overall metrics, followed by the stacking ensemble. Threshold optimization further improved multi-label performance, particularly for the ensemble and neural network models. While promising, the models are limited by the simplicity of their input features and require further validation and enhancement before clinical deployment. Nonetheless, the findings suggest that personalized, data-driven antibiotic recommendations are feasible and can play an important role in combating antibiotic resistance in hospital settings. Future work should aim to enrich input features, validate models externally, and explore real-time integration into clinical decision-making systems.

References

[1] Y. Huang et al., "Predicting antimicrobial resistance in hospitalized patients using machine learning techniques," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–12, 2021.

[2] Y.-C. Su et al., "Using machine learning to predict ESBL-producing E. coli infections: A comparative study," Comput. Biol. Med., vol. 126, p. 104043, 2020.

[3] H. Nguyen et al., "Predictive modeling for antibiotic resistance using MIMIC-III data," in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), 2020, pp. 2991–2996.

[4] P. D. Tamma et al., "Combination therapy for treatment of infections with gram-negative bacteria: What is the evidence?," JAMA, vol. 320, no. 9, pp. 979–981, 2018.

[5] T. M. Rawson et al., "Artificial intelligence in clinical decision support: Challenges for real-world implementation," Lancet Digit. Health, vol. 2, no. 10, pp. e489–e498, 2020.

[6] M. A. Aldeyab et al., "Modeling the impact of antibiotic use and infection control practices on hospital-associated infections and antimicrobial resistance," PLoS ONE, vol. 12, no. 3, e0173338, 2017.

[7] E. S. Berner and T. J. La Lande, "Overview of clinical decision support systems," in Clinical Decision Support Systems, Springer, 2016, pp. 1–17.

[8] B. Foxman, "Epidemiology of urinary tract infections: Incidence, morbidity, and economic costs," Am. J. Med., vol. 113, no. S1, pp. 5S–13S, 2002.

[9] H. Giannini et al., "A Bayesian framework for individualized prediction of UTI treatment response," Artif. Intell. Med., vol. 102, p. 101773, 2020.

[10] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[11] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992.

[12] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 3146–3154, 2017.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in IJCAI'95, pp. 1137–1145.