Predicting Antimicrobial Resistance for Common Pathogen-Antibiotic Pairs using Early Clinical Data from MIMIC-III

Aayushee Agarwal
Sharang Agarwal

Abstract

Antimicrobial resistance (AMR) poses a significant threat to global health. Early prediction of AMR can guide appropriate empirical antibiotic therapy, improving patient outcomes and mitigating resistance spread. This project investigates the feasibility of predicting AMR for common pathogen-antibiotic pairs using readily available clinical data from the first 24 hours following a positive culture result. We utilized the MIMIC-III critical care database and developed a machine learning pipeline to process the top 10 most frequent pathogen-antibiotic pairs, extracting features related to patient demographics, admission details, comorbidities (diabetes), prior antibiotic exposure, and early laboratory results (Creatinine, WBC, Bicarbonate). A RandomForestClassifier was trained for each pair to perform multiclass classification of antibiotic susceptibility (Resistant, Susceptible, Intermediate). Evaluation using weighted F1-scores on a held-out test set showed variable performance across pairs, ranging from approximately 0.72 to 0.997, demonstrating the potential but also the complexity of predicting AMR from early, non-genomic data. Feature importance analysis highlighted the relevance of patient age, specific lab values, and prior antibiotic use, though the relative importance varied between pairs. Pairs with highly imbalanced class distributions or insufficient minority class samples presented challenges for modeling and cross-validation. This work provides a framework and baseline results for early AMR prediction using routinely collected clinical data.

CCS Concepts

• **Applied computing** → **Health informatics**; • **Computing methodologies** → *Machine learning*; *Classification and regression trees.*

Keywords

Antimicrobial Resistance, Machine Learning, MIMIC-III, Critical Care, Predictive Modeling, RandomForest, GCS

1 Introduction

Antimicrobial resistance (AMR) is a critical global health challenge, threatening the effectiveness of treatments for common infections and increasing morbidity, mortality, and healthcare costs [4]. In critical care settings, timely administration of appropriate antibiotics is crucial, but definitive susceptibility results from microbiology cultures often take 24-72 hours. During this period, clinicians rely on empirical therapy based on likely pathogens and local resistance patterns. Incorrect empirical choices can lead to treatment failure, prolonged illness, and contribute to further resistance development.

Machine learning (ML) offers potential solutions for predicting AMR earlier, using readily available patient data. By identifying patterns associated with resistance before culture results are finalized, ML models could support clinical decision-making for more targeted empirical therapy.

This project aims to develop and evaluate ML models for predicting antimicrobial susceptibility (Resistant, Susceptible, or Intermediate) for common pathogen-antibiotic combinations found in the MIMIC-III critical care database [3]. We focus on using clinical data available within the first 24 hours of a positive culture event, simulating a realistic scenario for early prediction. Specifically, we investigate the top 10 most frequent pathogen-antibiotic pairs, build a RandomForestClassifier for each, and evaluate their predictive performance using weighted F1-scores. We also analyze feature importances to understand factors driving the predictions. The data is sourced directly from MIMIC-III compressed csv files stored in Google Cloud Storage (GCS).

Related Work

Several studies have explored the use of ML for AMR prediction using diverse clinical datasets and computational approaches. The following key works provide important context and serve as foundational pre-work for research focused on early AMR prediction using critical care data:

Ghosh et al. (2019) [2] conducted a study highly pertinent to this research, as it also utilizes the MIMIC-III critical care database to predict antibiotic sensitivity. The authors developed an ensemble of machine learning algorithms and reported prediction accuracy (approximately 87% overall, with an average AUROC around 0.91) for antibiotic efficacy. They incorporated a comprehensive set of patient-specific attributes found in EHRs, such as gender, comorbidities, site of infection, history of past hospitalization, and previous antibiotic usage. This work highlights the significant potential of leveraging routinely collected patient data within a critical care database like MIMIC-III for developing personalized antibiotic susceptibility predictions, aligning directly with the aim of this project to use early clinical data for such predictions.

Ferrari et al. (2024) [1] focused on the critical challenge of predicting bloodstream infections (BSI) and associated AMR in ICU patients by utilizing clinical data available at the point of admission. They introduced an innovative Multi-Objective Symbolic Regression (MOSR) approach, which prioritizes the development of interpretable models—a key consideration for clinical adoption. Their research underscores the value of early prediction using readily accessible data to bolster antimicrobial stewardship (AMS) programs. The study also addresses common hurdles in AMR prediction, such as data imbalance, and strives for reliable and robust results. Although their work encompasses BSI prediction in conjunction with AMR and employs a different ML methodology, the emphasis on leveraging early ICU data to guide antimicrobial decision-making.

2 Methods

2.1 Data Source and Cohort

This study utilized the Medical Information Mart for Intensive Care III (MIMIC-III) v1.4 database [3], a large, publicly available dataset containing de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Data access was obtained under the specified credentialing requirements.

The raw data was accessed as compressed CSV files (.csv.gz) stored in Google Cloud Storage (GCS). Relevant tables included: MICROBIOLOGYEVENTS, PATIENTS, ADMISSIONS, LABEVENTS, PRESCRIPTIONS, and DIAGNOSES_ICD.

The primary cohort consisted of microbiology events from MICROBIOLOGYEVENTS.csv.gz where the interpretation was 'R' (Resistant), 'S' (Susceptible), or 'I' (Intermediate). We focused on the top 10 most frequently occurring organism (ORG_NAME) and antibiotic (AB_NAME) pairs within this filtered dataset. Each unique culture event (approximated by HADM_ID and CULTURE_CHARTTIME) served as a sample point.

2.2 Feature Engineering

For each sample (culture event), we extracted features based on data available within a defined time window relative to the culture time (CULTURE_CHARTTIME):

- **Demographics:** Age at admission (calculated from ADMITTIME and DOB, capped at 90 years), Gender (PATIENTS).
- **Admission Info:** Admission type, Ethnicity (ADMISSIONS).
- **Prior Antibiotic Exposure:** A binary flag (HAD_PRIOR _ANTIBIOTICS) indicating if any antibiotic prescription (PRESCRIPTIONS) started within 30 days prior to the culture time.
- **Comorbidities:** A binary flag (HAS_DIABETES) derived from ICD9 codes starting with '250' in DIAGNOSES_ICD associated with the hospital admission (HADM_ID).
- **Early Lab Results:** Average values of Creatinine (ITEMIDs 50912, 50811), White Blood Cell Count (WBC, ITEMID 51301), and Bicarbonate (ITEMID 50882) recorded in LABEVENTS within the first 24 hours following the CULTURE _CHARTTIME.
- **Specimen Type:** The type of specimen collected (SPEC_TYPE _DESC from MICROBIOLOGYEVENTS).

Vital signs from CHARTEVENTS were initially considered but excluded due to the significant processing time and memory requirements associated with loading and filtering this large table.

2.3 Data Processing

The data processing pipeline involved the following steps, implemented primarily using the pandas library:

1. Loading required CSVs from GCS using specified data types (dtypes) for memory optimization. Date columns were parsed, coercing errors to NaT (Not a Time).
2. Identifying the top 10 organism-antibiotic pairs based on frequency in the filtered MICROBIOLOGYEVENTS data.
3. Iterating through each of the top 10 pairs:
   - Filtering MICROBIOLOGYEVENTS for the current pair.
   - Merging with ADMISSIONS and PATIENTS to get base demographic and admission data.
   - Calculating age robustly, handling date shifts for older patients.
   - Defining time windows for feature extraction.
   - Filtering LABEVENTS, PRESCRIPTIONS, and DIAGNOSES _ICD for relevant HADM_IDs and time windows.
   - Calculating aggregated lab features, prior antibiotic flag, and diabetes flag using grouping and merging operations.
   - Assembling the final feature set for the pair.
   - Dropping duplicate culture events and rows with missing target values.
4. **Outlier Handling:** Numerical features (excluding binary flags) were capped at 1.5 times the Interquartile Range (IQR) below the first quartile (Q1) and above the third quartile (Q3) to mitigate the influence of extreme values.
5. **Target Encoding:** The categorical target variable (INTERPRETATION with values 'R', 'S', 'I') was encoded into numerical labels using sklearn.preprocessing.LabelEncoder, fitted globally on the relevant interpretations.

2.4 Modeling

For each of the top 10 pairs (where sufficient data and class representation allowed), the following modeling steps were performed using scikit-learn:

1. **Preprocessing Pipeline:** A ColumnTransformer was used to apply different preprocessing steps to numerical and categorical features.
   - Numerical features: Missing values were imputed using the median, followed by standardization using StandardScaler.
   - Categorical features: Missing values were imputed using the most frequent value, followed by one-hot encoding using OneHotEncoder (handling unknown categories encountered during testing).
2. **Train/Test Split:** The dataset for each pair was split into training (75%) and testing (25%) sets using train_test_split, stratified by the encoded target variable to maintain class proportions. Pairs were skipped if the minimum class count was less than the number of CV splits (5) required for stratification.
3. **Model:** A RandomForestClassifier (with n_estimators=100, class_weight='balanced' to handle class imbalance, and random_state=42) was chosen as the classification model due to its robustness and ability to provide feature importances.
4. **Pipeline Integration:** The preprocessor and classifier were combined into a single Pipeline object.

2.5 Evaluation

Model performance was evaluated using:

1. **Cross-Validation (CV):** Stratified 5-fold cross-validation (StratifiedKFold) was performed on the training set for each pair. The mean weighted F1-score across the folds was calculated using cross_val_score. CV was skipped if the minimum class count in the training set was less than the number of splits (5).

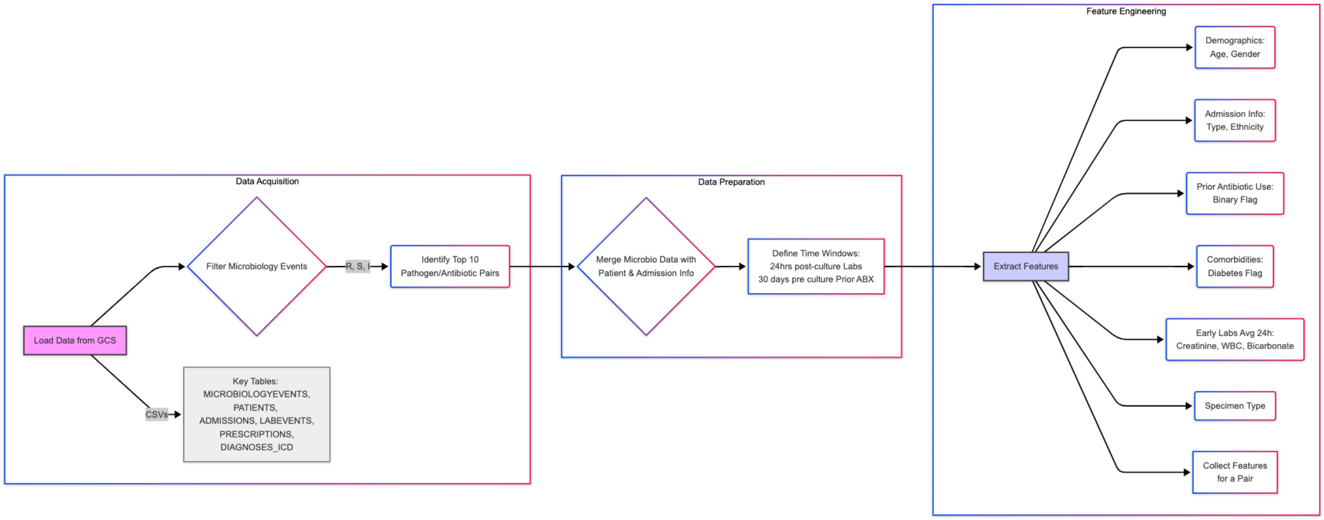

Data Acquisition
Load Data from GCS
Filter Microbiology Events
R, S, I
Identify Top 10 Pathogen/Antibiotic Pairs
CSVs
Key Tables: MICROBIOLOGYEVENTS, PATIENTS, ADMISSIONS, LABEVENTS, PRESCRIPTIONS, DIAGNOSES_ICD
Data Preparation
Merge Microbio Data with Patient & Admission Info
Define Time Windows: 24hrs post-culture Labs 30 days pre culture Prior ABX
Feature Engineering
Extract Features
Demographics: Age, Gender
Admission Info: Type, Ethnicity
Prior Antibiotic Use: Binary Flag
Comorbidities: Diabetes Flag
Early Labs Avg 24h: Creatinine, WBC, Bicarbonate
Specimen Type
Collect Features for a Pair

Figure 1: Diagram illustrating the data acquisition, preparation, and feature engineering steps.

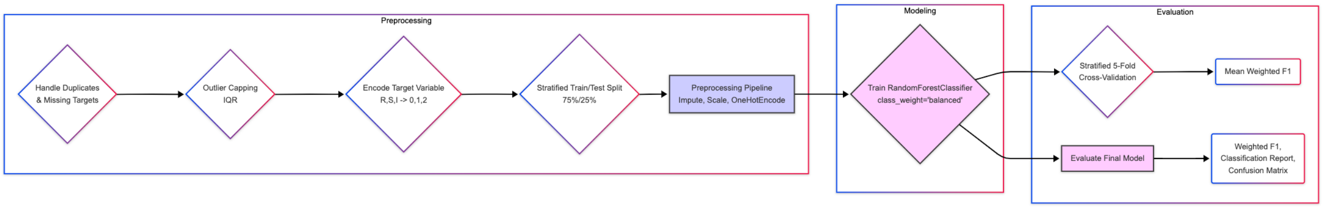

Preprocessing
Handle Duplicates & Missing Targets
Outlier Capping IQR
Encode Target Variable R,S,I -> 0,1,2
Stratified Train/Test Split 75%/25%
Preprocessing Pipeline Impute, Scale, OneHotEncode
Modeling
Train RandomForestClassifier class_weight='balanced'
Evaluation
Stratified 5-Fold Cross-Validation
Mean Weighted F1
Evaluate Final Model
Weighted F1, Classification Report, Confusion Matrix

Figure 2: Diagram illustrating the preprocessing, modeling, and evaluation steps.

(2) **Test Set Evaluation:** The trained pipeline was evaluated on the held-out test set.
   - A detailed classification_report was generated, showing precision, recall, and F1 score for each class present in the test set.
   - The overall weighted F1 score was calculated using f1_score (average = ’weighted’).
   - A confusion_matrix was generated and visualized using seaborn.heatmap to show misclassification patterns. Labels for the report and matrix were dynamically determined based on classes present in the test set.

(3) **Feature Importance:** Feature importances were extracted from the trained RandomForest model within the pipeline. The top 5 features for each pair were recorded.

3 Results

The analysis pipeline was executed for the top 10 most frequent organism-antibiotic pairs identified in the MIMIC-III microbiology data. Two pairs (’STAPH AUREUS COAG+ / CLINDAMYCIN’ and ’STAPH AUREUS COAG+ / TETRACYCLINE’) were skipped during preprocessing because the minimum count for the ’Intermediate’ (I) class was less than the required 5 samples for 5-fold stratified cross-validation.

The results for the remaining 8 pairs are summarized in Table 1. Note that feature names containing underscores have been escaped (e.g., AVG_WBC_FIRST24H).

The predictive performance, as measured by the weighted F1-score on the test set, varied considerably across the different pairs.

- High performance (Test F1 > 0.95) was observed for pairs where one class was extremely dominant, such as Vancomycin, Penicillin, Rifampin, and Gentamicin resistance prediction for *Staphylococcus aureus*. In these cases, the model primarily learned to predict the majority class (’S’ or ’R’).
- Moderate performance (Test F1 ≈ 0.86) was achieved for *Escherichia coli* / Gentamicin, which had a more balanced distribution between ’S’ and ’R’ but still a small ’I’ class.
- Lower performance (Test F1 ≈ 0.72) was seen for *Staphylococcus aureus* against Erythromycin, Oxacillin, and Levofloxacin. These pairs had more balanced distributions between ’R’ and ’S’, suggesting greater difficulty in distinguishing resistant and susceptible cases based solely on the included early clinical features.

Confusion matrices (generated by the script, Figure 3) revealed common miss-classification patterns, particularly for the minority classes (’I’ and sometimes ’R’ or ’S’ depending on the pair).

Table 1: Summary of Multiclass Classification Results for Top Organism/Antibiotic Pairs

Organism / Antibiotic Pair	Status	N Samples	Target Distr.	Mean CV F1 (W)	Test F1 (W)	Top 5 Features (Approx.)
STAPH AUREUS COAG+ / VANCOMYCIN	Completed	4123	I: 0.2%, S: 99.8%	0.9970	0.9966	AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, AVG_WBC_FIRST24H, AGE_AT_ADMISSION, SPEC_TYPE_DESC_BLOOD CULTURE
STAPH AUREUS COAG+ / PENICILLIN	Completed	4069	R: 98.4%, S: 1.6%	0.9785	0.9833	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
STAPH AUREUS COAG+ / RIFAMPIN	Completed	3962	I: 0.7%, R: 2.2%, S: 97.2%	0.9647	0.9770	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / GENTAMICIN	Completed	6742	I: 0.3%, R: 3.0%, S: 96.7%	0.9666	0.9676	AVG_CREATININE_FIRST24H, AGE_AT_ADMISSION, AVG_WBC_FIRST24H, AVG_BICARBONATE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
ESCHERICHIA COLI / GENTAMICIN	Completed	3505	I: 0.9%, R: 12.2%, S: 86.9%	0.8565	0.8573	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, SPEC_TYPE_DESC_BLOOD CULTURE
STAPH AUREUS COAG+ / ERYTHROMYCIN	Completed	6479	I: 2.1%, R: 69.7%, S: 28.2%	0.6915	0.7262	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_BICARBONATE_FIRST24H, AVG_CREATININE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / OXACILLIN	Completed	7553	R: 59.6%, S: 40.4%	0.6938	0.7249	AGE_AT_ADMISSION, AVG_WBC_FIRST24H, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, HAS_DIABETES
STAPH AUREUS COAG+ / LEVOFLOXACIN	Completed	6717	I: 1.4%, R: 61.9%, S: 36.7%	0.7216	0.7177	AVG_WBC_FIRST24H, AGE_AT_ADMISSION, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H, HAD_PRIOR_ANTIBIOTICS
STAPH AUREUS COAG+ / CLINDAMYCIN	Skipped - Min Class Count < 5 (Full Set)	NaN	N/A	N/A	N/A	N/A
STAPH AUREUS COAG+ / TETRACYCLINE	Skipped - Min Class Count < 5 (Full Set)	NaN	N/A	N/A	N/A	N/A

Feature importance analysis consistently highlighted patient age (AGE_AT_ADMISSION) and the average values of early labs (AVG_WBC_FIRST24H, AVG_CREATININE_FIRST24H, AVG_BICARBONATE_FIRST24H) among the top predictors across most pairs. Prior antibiotic exposure (HAD_PRIOR_ANTIBIOTICS) and the presence of diabetes (HAS_DIABETES) also appeared in the top 5 for several pairs. The specific specimen type (e.g., SPEC_TYPE_DESC_BLOOD CULTURE) was important for some pairs like *Staph. aureus* / Vancomycin and *E. coli* / Gentamicin. The relative ranking of these features varied depending on the specific organism-antibiotic combination.

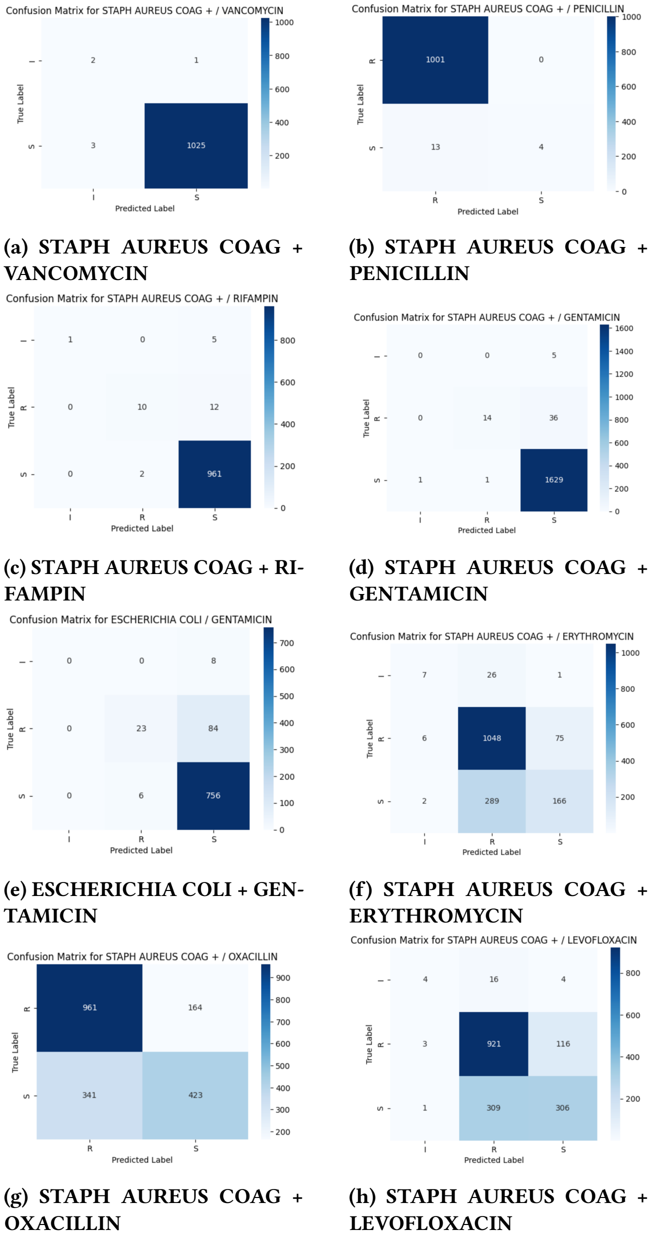
4 Discussion

This study demonstrated the application of a machine learning pipeline to predict antimicrobial susceptibility for the top 10 common organism-antibiotic pairs using early clinical data from MIMIC-III sourced from GCS. The results indicate that prediction performance is highly dependent on the specific pair being considered and the underlying distribution of resistance classes.

Pairs with highly skewed distributions (e.g., Vancomycin, Penicillin resistance for *Staph. aureus*) achieved high weighted F1-scores, but this performance is largely driven by correctly identifying the overwhelmingly majority class. The clinical utility for these specific pairs might be limited unless the model shows high precision/recall for the rare resistant/susceptible class. The confusion matrices generated by the script (visualized in the notebook output, see examples in Figure 3) are crucial for assessing this.

For pairs with more balanced distributions between resistant and susceptible classes (e.g., Oxacillin, Levofloxacin, Erythromycin for *Staph. aureus*), the lower F1-scores (≈0.72) suggest that the included early clinical features (demographics, basic labs, limited comorbidities/history) may not be sufficient to reliably distinguish resistance patterns. More informative features, potentially including vital signs (which were excluded here due to processing constraints), more detailed comorbidity information, prior hospitalization data, or specific procedure details, might be necessary to improve performance for these challenging cases.

The consistent appearance of age, WBC, creatinine, and bicarbonate among the top features aligns with clinical intuition, as these relate to patient baseline health, immune response, and organ function, which can influence infection severity and potentially correlate with resistance mechanisms or prior exposures. The importance of prior antibiotic use and diabetes also highlights the role of patient history and comorbidities.



(a) STAPH AUREUS COAG + VANCOMYCIN

(b) STAPH AUREUS COAG + PENICILLIN

(c) STAPH AUREUS COAG + RIFAMPIN

(d) STAPH AUREUS COAG + GENTAMICIN

(e) ESCHERICHIA COLI + GENTAMICIN

(f) STAPH AUREUS COAG + ERYTHROMYCIN

(g) STAPH AUREUS COAG + OXACILLIN

(h) STAPH AUREUS COAG + LEVOFLOXACIN

Figure 3: Confusion Matrix

4.1 Limitations

This study has several limitations:

- **Excluded Features:** Vital signs (CHARTEVENTS) were excluded, potentially removing valuable predictive information.
- **Simplified Feature Engineering:** Only average lab values within the first 24 hours were used. Trends or more complex temporal patterns were not explored.
- **Single Dataset:** Results are based on the MIMIC-III dataset from a single center and may not generalize perfectly to other populations or healthcare systems.
- **Basic Model:** Only RandomForest was used. Other algorithms might yield different performance. Hyperparameter tuning was not performed.
- **Skipped Pairs:** Two of the top 10 pairs were skipped due to insufficient minority class samples for cross-validation, highlighting data limitations for less frequent resistance patterns.
- **Causality:** The model identifies correlations, not causal relationships, between features and resistance.

5 Conclusion and Future Work

We successfully developed and applied a pipeline to predict multi-class antimicrobial susceptibility for common pathogen-antibiotic pairs using early clinical data from MIMIC-III CSVs on GCS. The RandomForest models achieved variable performance, with weighted F1-scores ranging from ≈0.72 to ≈0.997 on the test set, heavily influenced by the class distribution of each pair. Key predictors consistently included age and early laboratory results (WBC, Creatinine, Bicarbonate).

This work establishes a baseline and highlights the challenges of predicting AMR using limited early clinical data, especially for pairs with balanced resistance profiles. Future work could focus on:

- Incorporating vital sign trends and other potentially relevant clinical features.
- Exploring more advanced feature engineering techniques (e.g., time-series analysis).
- Evaluating different machine learning models, including gradient boosting and potentially deep learning approaches.
- Performing systematic hyperparameter tuning.
- Validating the models on external datasets.
- Investigating methods to improve prediction for minority classes, such as specialized sampling techniques or cost-sensitive learning.

Developing reliable early AMR prediction models holds promise for optimizing antibiotic therapy and combating the rise of resistance.

Acknowledgments

This research utilized the MIMIC-III database collected by the MIT Laboratory for Computational Physiology, which is supported by NIH grant R01 EB017205.

References

[1] D. Ferrari, P. Arina, J. Edgeworth, V. Curcin, V. Guidetti, F. Mandreoli, and Y. Wang. 2024. Using interpretable machine learning to predict bloodstream infection and antimicrobial resistance in patients admitted to ICU: Early alert predictors based on EHR data to guide antimicrobial stewardship. *PLOS Digital Health* 3, 10 (2024), e0000641. doi:10.1371/journal.pdig.0000641

[2] P. Ghosh, S. Sharma, E. Hasan, S. Ashraf, V. Singh, D. Tewari, S. Singh, M. Kapoor, and D. Sengupta. 2019. Machine learning based prediction of antibiotic sensitivity in patients with critical illness. *medRxiv* (2019). doi:10.1101/19007153

[3] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. doi:10.1038/sdata.2016.35

[4] World Health Organization. 2021. Antimicrobial resistance.