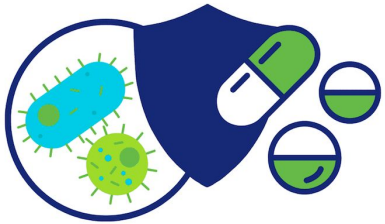
Predicting Antibiotic Resistance: Challenges, Findings, and Lessons Learned



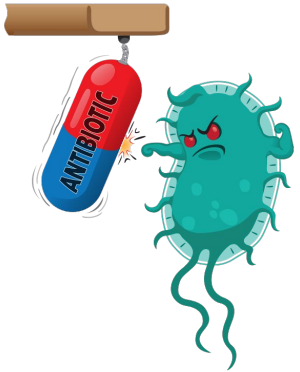
Mustafain Ali

Tinghao Huang

Dominique Hughes

Chiara Mattamira

Haejun (Stella) Oh



The Prescription Challenge

- When a patient presents with a bacterial infection, clinicians must choose among **several** possible antibiotics, often **before** lab results are available
- Resistance testing takes **24–72 hours** and may not be routinely available in all hospitals.
- Incorrect or delayed antibiotic choices can **worsen patient outcomes** and contribute to **rising resistance**.

Aims & Objectives

- Objective: Explore whether patient and microbiological data can **predict antibiotic resistance**.

- Impact:
  - Support clinicians in making **data-informed** antibiotic **prescriptions** while awaiting lab results.
  - Identify **key resistance risk factors** to guide more effective treatment decisions.

- Approach:
  - Train **machine learning models** that classify each antibiotic as **susceptible** or **resistant**.
  - Examine which patient and clinical **features** (e.g., age, ward type, prior antibiotic exposure) **most influence** resistance.

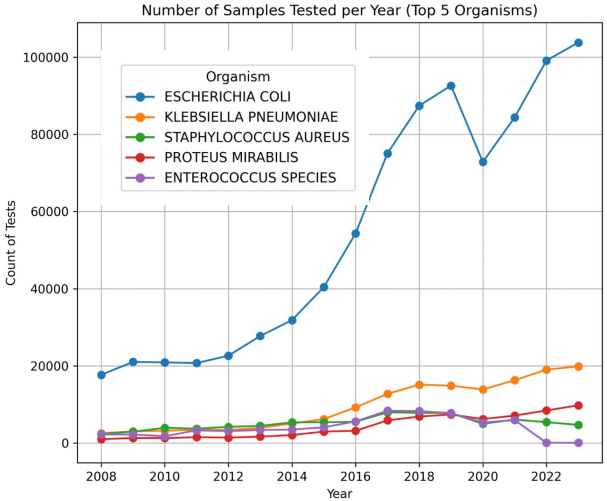
Dataset & Preliminary Preprocessing

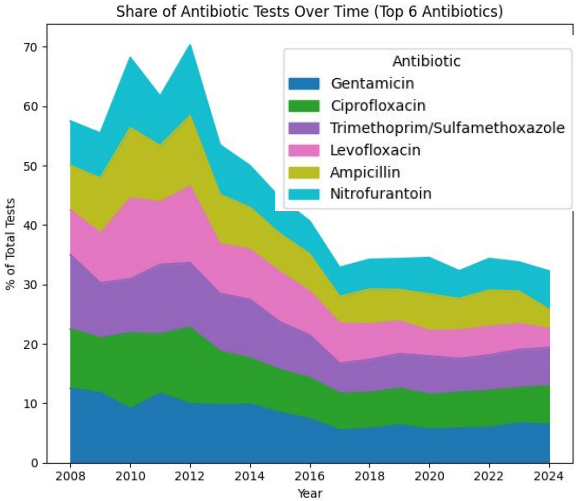
- We use the Dryad database (Fateme et. al, 2025):
  ~**750k** microbiological culture records from **~300k** unique patients (**1999-2024)**.
- **Features:** age, prior infections, prior antibiotics, adi score, nursing home visits, ward
- **Preliminary Preprocessing:**
  - Removed duplicates
  - Removed repeated visits → first visit only
  - One-hot encoded categorical features (e.g. ward info)

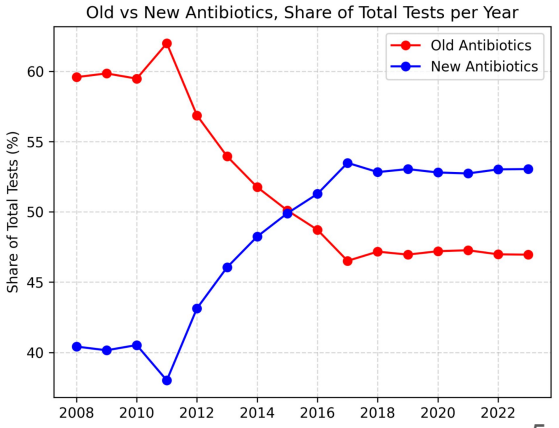
Anon ID	Prior Organism Exposure	Prior Antibiotics Used	Age of Subject	Gender	Area Deprivation Index (ADI Score)	ADI State Rank	Days between Nursing Home Visits & Blood Culture	Bacterial Culture from Inpatients	Bacterial Culture from Outpatients	Bacterial Culture from Emergency Room	Bacterial Culture from ICU
1164374	2	1	70	0	6	3	0	0	1	0	0
1064815	0	1	50	0	22	7	0	0	1	0	0

Additional Preprocessing Based on EDA

- Focused on ***E. coli*** isolates only (most frequently tested organism)
- Selected data from **2016 onward** to reflect current resistance patterns and antibiotic guidelines
- Kept only the **9 antibiotics** tested for ≥75% of *E. coli* patients







Modeling Approach

Goal: Predict if a patient's *E. coli* isolate is resistant (1) or susceptible (0).

Target: Each antibiotic susceptibility/resistance

Models Tested:

- Dummy Classifier - baseline
- Logistic Regression
- Random Forest
- XGBoost
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

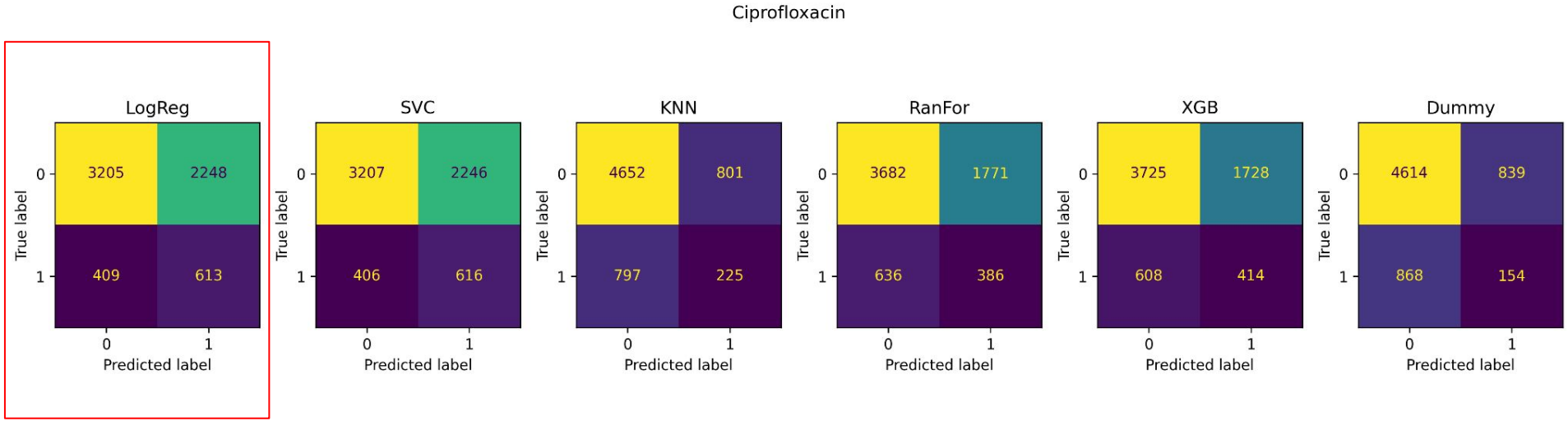
Pipeline Highlights:

- Data split **80%** train / **20%** test
- Outer **5-fold** CV on training set for model selection
- Inner **3-fold** CV for hyper-parameter tuning with Randomized Search CV
- **StandardScaler** applied for distance-based models (SVM, KNN, LogReg)
- **SMOTE** - enhance learning performance over highly imbalance data
- Downsample on larger class to balance the dataset 50/50

Evaluation Metrics

- **F1 Weighted Score:** Harmonic mean of precision & recall
- **False Negative Rate (FNR):** Critical metric representing the % of resistant cases misclassified as susceptible. (Should be minimized, since missing a resistant infection can lead to ineffective treatment.)
- **Precision:** Correctly predicted resistant out of all predicted resistant
- **Recall / Sensitivity:** Correctly identified resistant cases
- **Accuracy:** Overall correctness
- **RR AUC**: Correctly distinguishing between positive/negative cases

Confusion Matrices

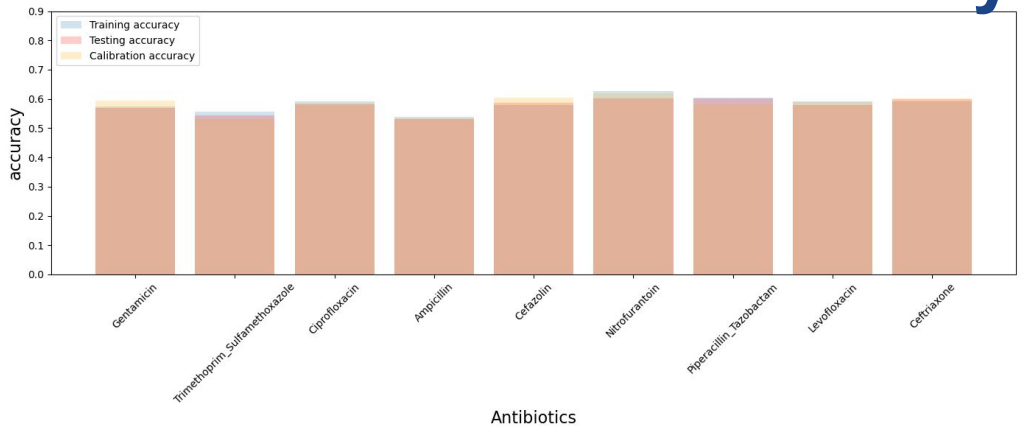


Chosen model: logistic regression

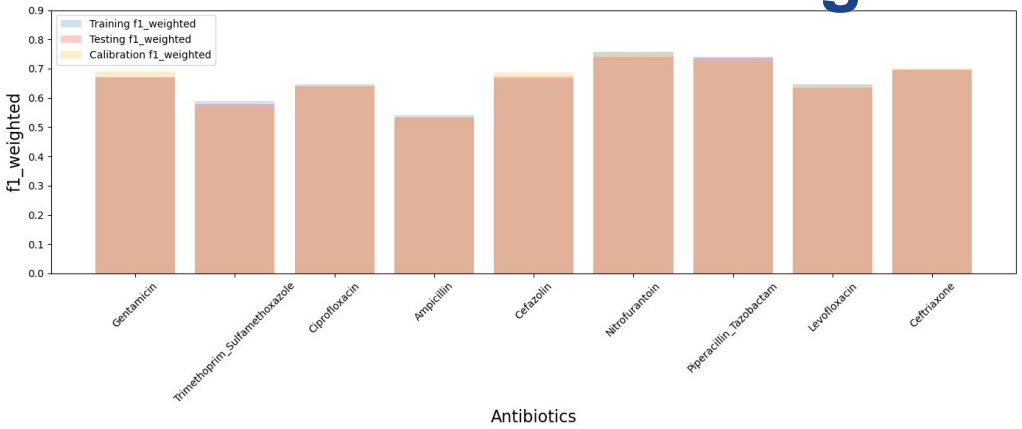
Results

- Looked at evaluation metrics between training, testing, and calibration dataset for any sign of over or under training.
- **Stable** accuracy and **small variability** in F1 weighted across antibiotics.

Accuracy



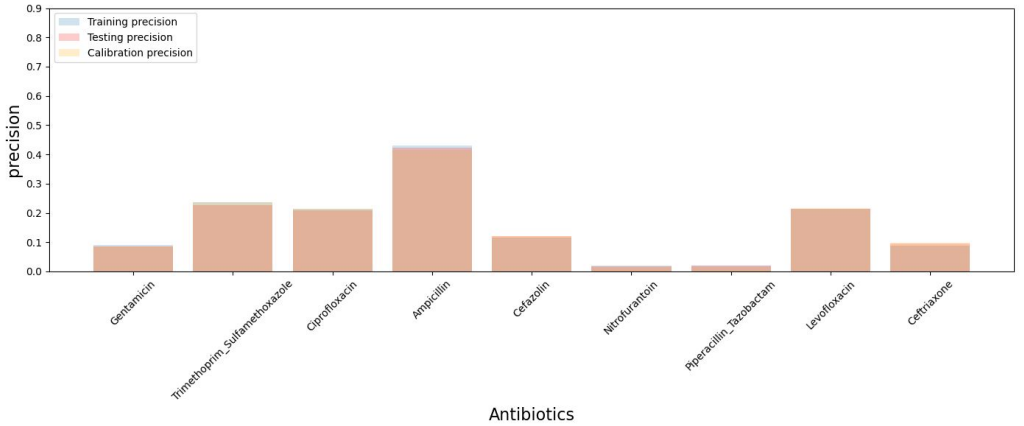
F1 Weighted



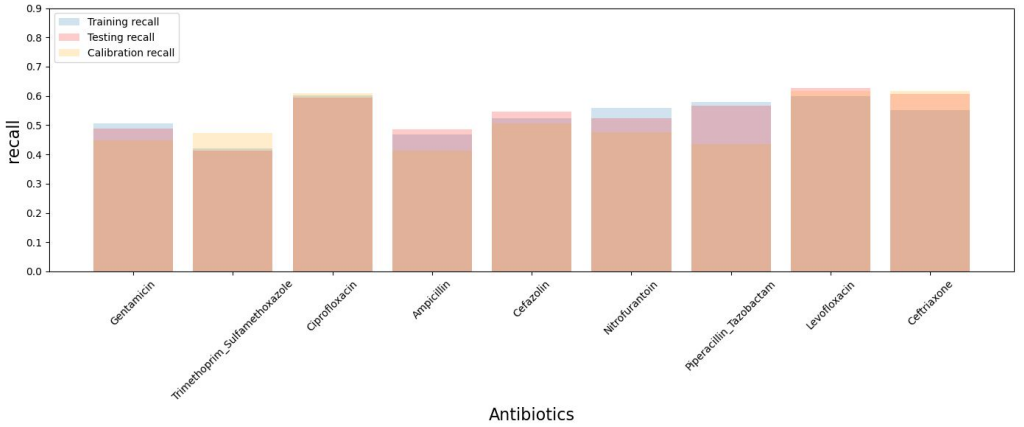
Results

- **Large** variability in Precision and **small** variability in recall across antibiotics

Precision



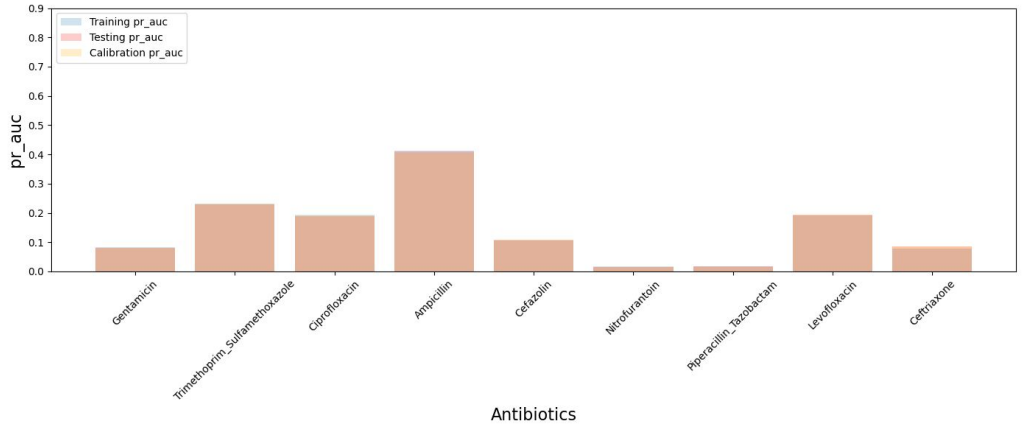
Recall



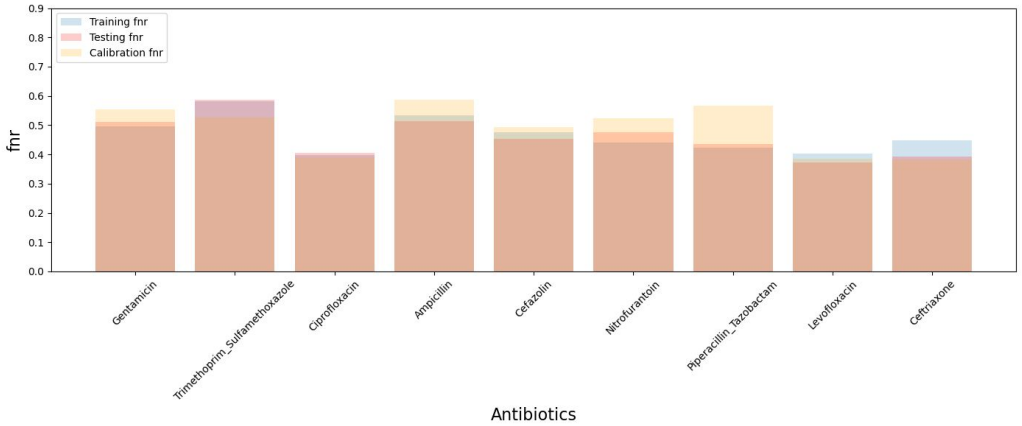
Results

- **Large** variability in PR-AUC and **small** variability in recall across antibiotics

PR-AUC

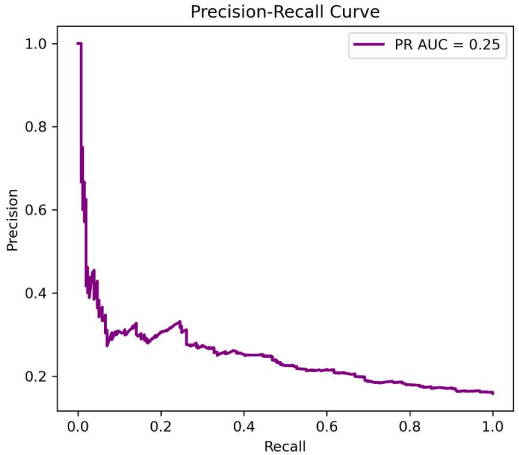


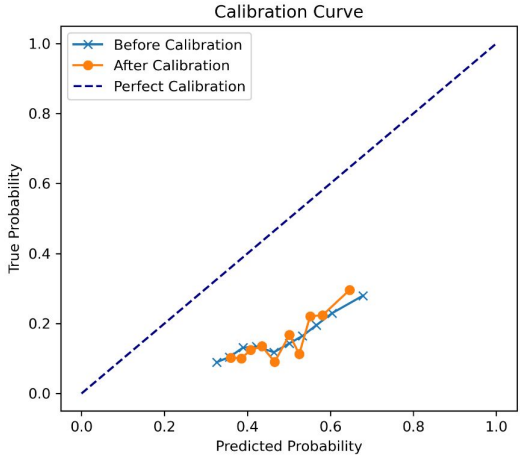
FNR

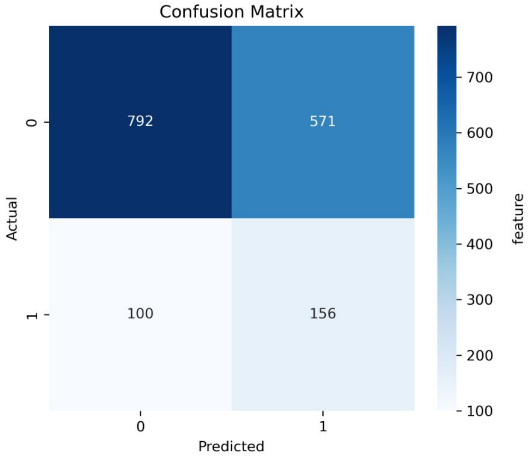


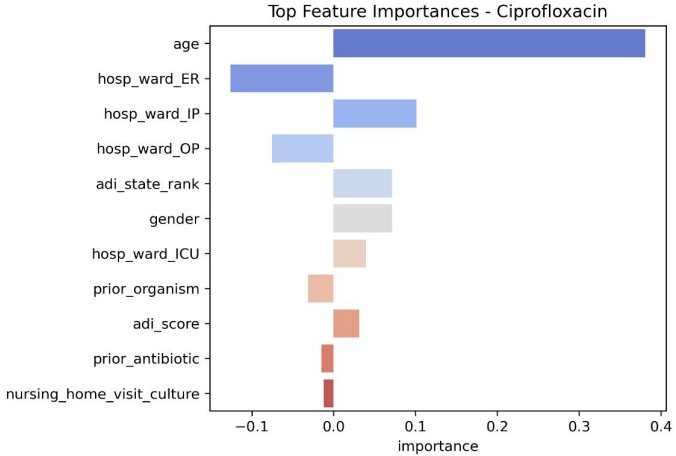
Results

Logistic Regression Model Evaluation for Ciprofloxacin

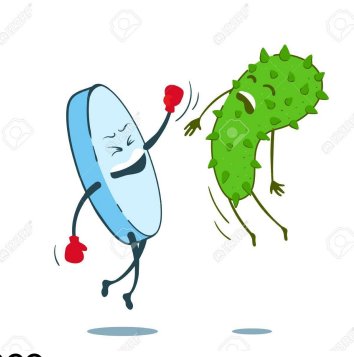








Conclusions



- Completed rigorous **data cleaning**, **preprocessing**, and **model calibration**.
- Achieved decent **accuracy** and **F1 weighted** score across most antibiotics.
- **False-negative rates** remain high, signaling limited sensitivity to predicting resistance.

1. Feature Limitations

- Datasets lack detailed biological or treatment features due to privacy, limiting model depth.

2. Data Imbalance

- Resistant cases are rare (e.g., *Ceftriaxone*: 7%). Even with SMOTE, models biased toward accuracy inflate false negatives.

3. Biological Complexity

- Resistance stems from diverse, nonlinear mechanisms—mutations, plasmids, efflux pumps—hard to capture computationally.

Future Directions

1. Time-Series Modeling

- Track monthly resistance trends and forecast future shifts.

2. Extended Feature Importance Exploration

- Use richer data to pinpoint key biological, clinical, and socioeconomic factors.

3. Cross-Antibiotic Correlation

- Analyze co-resistance patterns to reveal links across different drug classes.

Acknowledgements

- **Our mentor:** Adedolapo Ojoawo
- **Erdös Institute**
  - Steven Gubkin
  - Roman Holowinsky
  - Alec Clott