# **AI - Prediction of Neisseria gonorrhoeae Resistance at the Point of Care from Genomic and Epidemiologic Data**

# Vinothkumar Kolluru, Shreyas Sir, Advaitha Naidu Chintakunta

## **Abstract**

Antimicrobial resistance (AMR) in *Neisseria gonorrhoeae* (NG) has reached crisis levels, threatening decades of progress in the control of gonorrhoea—a sexually transmitted infection that infects an estimated 82 million individuals worldwide each year. Contemporary laboratory surveillance remains indispensable yet cannot scale to the pace at which resistance emerges, ultimately impeding timely, patient‑specific therapy. Leveraging a public repository comprising 3,786 NG isolate metadata records and associated phenotypic susceptibility profiles for azithromycin, ciprofloxacin and cefixime, we designed a unified machine‑learning (ML) and deep‑learning (DL) pipeline to predict binary resistance outcomes directly from routinely collected clinical variables and unitig‑derived genomic features. Rigorous preprocessing handled 23 % sparsity, harmonised 31 heterogeneous columns and applied skewness‑aware imputation prior to feature encoding. A CatBoost classifier—selected for its native handling of categorical inputs—was benchmarked against 32 baseline algorithms via the LazyPredict framework, while a three‑layer fully connected neural network provided a DL counterpart. Stratified five‑fold cross‑validation demonstrated macro‑averaged AUCs of 0.97 (ciprofloxacin), 0.95 (cefixime) and 0.94 (azithromycin), outperforming published genomic GWAS models by 4–7 percentage points. SHAP‑based interpretation highlighted known resistance determinants, including mosaic *penA* alleles and mutations in the *mtrR* promoter, validating biological plausibility. Our findings indicate that point‑of‑care ML tools, fuelled by continuously accrued surveillance data, could forecast resistance with near‑laboratory accuracy, enabling antibiotic stewardship in real time. Nevertheless, model generalisability beyond high‑income settings requires further evaluation.

**Keywords:** antimicrobial resistance; *Neisseria gonorrhoeae*; CatBoost; deep learning; surveillance genomics; predictive modelling

**1 Introduction**

### **1.1 Epidemiological background**

Gonorrhoea—caused by the Gram‑negative diplococcus *Neisseria gonorrhoeae*—remains one of the most frequently reported notifiable diseases globally, with the World Health Organization (WHO) estimating 82 million incident cases in 2020 alone. Despite long‑standing awareness, its clinical burden continues to rise, propelled by asymptomatic carriage, inadequate diagnostics in low‑resource settings, and sociobehavioural factors such as increased urbanisation and syndemic interactions with other sexually transmitted infections (STIs). Untreated infection manifests differently across anatomical sites: urethritis and cervicitis dominate urogenital disease, whereas pharyngeal and rectal colonisation fuel hidden transmission chains. Morbidity extends beyond acute symptoms; chronic sequelae include pelvic inflammatory disease, ectopic pregnancy, infertility and, in rare cases, disseminated gonococcal infection. The economic impact, estimated at billions of US dollars annually, stems from direct treatment expenses and productivity losses associated with adverse reproductive outcomes. Confronted with these figures, public‑health stakeholders intensify calls for innovative surveillance methodologies capable of anticipating emerging resistance phenotypes before treatment failures proliferate. Traditional culture‑based antimicrobial susceptibility testing (AST), although definitive, requires specialised laboratories and a 24–72 h turnaround—constraints that hamper its utility for real‑time decision making in fast‑moving clinical workflows such as sexual‑health clinics. Molecular diagnostics partially alleviate this gap but usually target single nucleotide variants conferring resistance to a single drug, thus lacking the breadth to capture polygenic or novel mechanisms. Consequently, computational inference from multivariate epidemiological and genomic datasets is increasingly viewed as a cornerstone of next‑generation AMR surveillance.

### **1.2 Burden of antimicrobial resistance**

The therapeutic landscape for NG has narrowed precipitously over the past five decades. Sulfonamides, penicillins, tetracyclines, macrolides, and fluoroquinolones have each succumbed to widespread resistance, culminating in the current reliance on extended‑spectrum cephalosporins (ESCs) such as ceftriaxone—often administered in dual regimens with azithromycin. However, declining azithromycin susceptibility and sporadic reports of ceftriaxone treatment failure signal an ominous trajectory toward pan‑resistant strains. Surveillance programmes such as the WHO Global Gonococcal Antimicrobial Surveillance Programme (GASP) and the US Gonococcal Isolate Surveillance Project (GISP) provide invaluable trend data but operate with limited geographical coverage, leading to blind spots in regions where AMR emergence may incubate undetected. The clinical stakes are high: resistance‑driven therapeutic failure not only prolongs transmission but also necessitates empiric use of last‑line or off‑label antibiotics, accelerating the resistance treadmill. Consequently, predictive analytics that can pre‑empt resistance—facilitating targeted therapy and preserving antibiotic efficacy—are critically needed. Yet, challenges persist in integrating heterogeneous surveillance data, addressing missingness, and capturing the complex, often epistatic genetic architecture underlying resistance. To tackle these obstacles, the present study proposes a holistic pipeline that bridges epidemiological metadata, phenotypic AST results, and high‑dimensional genomic features within a unified analytical framework.

### **1.3 Data‑driven approaches in gonorrhoea**

Early predictive efforts employed simple logistic regression models using demographic variables, reporting modest performance with area under the receiver‑operating characteristic curve (AUC) rarely exceeding 0.80. Advances in whole‑genome sequencing (WGS) subsequently enabled genotype‑to‑phenotype prediction frameworks, from k‑mer‑based machine learning to genome‑wide association studies (GWAS) that pinpoint single nucleotide polymorphisms (SNPs) associated with resistance. However, many of these methods either overfit due to the curse of dimensionality or fail to exploit non‑linear interactions critical for capturing multi‑factorial resistance. Ensemble tree‑based algorithms such as random forests and gradient boosting machines offer improved performance but still require extensive feature engineering. Deep learning—capable of automatic feature extraction—has shown promise, yet its application in NG remains sparse due to concerns over interpretability and data scarcity. By systematically benchmarking 32 classical ML models and implementing a deep neural network within a single cross‑validated pipeline, our study delivers a comprehensive performance landscape while employing SHapley Additive exPlanations (SHAP) to preserve model transparency. Furthermore, we incorporate epidemiological covariates such as year, geographic region, and patient group to evaluate their additive predictive value beyond genomic features, thereby informing pragmatic surveillance strategies where sequencing data may be intermittent.

### **1.4 Study objectives**

Against this backdrop, the overarching objective of our investigation is twofold. First, we seek to construct robust predictive models that accurately classify resistance to azithromycin, ciprofloxacin and cefixime—the three antibiotics with the most complete susceptibility data in the curated dataset. Second, we aim to elucidate the relative contribution of epidemiological versus genomic features, thereby gauging the feasibility of deploying predictive tools in resource‑limited contexts where WGS may not be routinely available. To accomplish these goals, we articulate the following specific aims:

(i) harmonise and preprocess a 31‑feature NG surveillance dataset exhibiting heterogeneous data types and 23 % overall missingness;

(ii) conduct an extensive exploratory data analysis (EDA) to visualise temporal and spatial resistance patterns and quantify feature correlations;

(iii) benchmark an array of off‑the‑shelf ML algorithms using LazyPredict to establish performance baselines;

(iv) develop and fine‑tune a CatBoost model optimised for categorical data, alongside a feed‑forward neural network as a deep‑learning comparator;

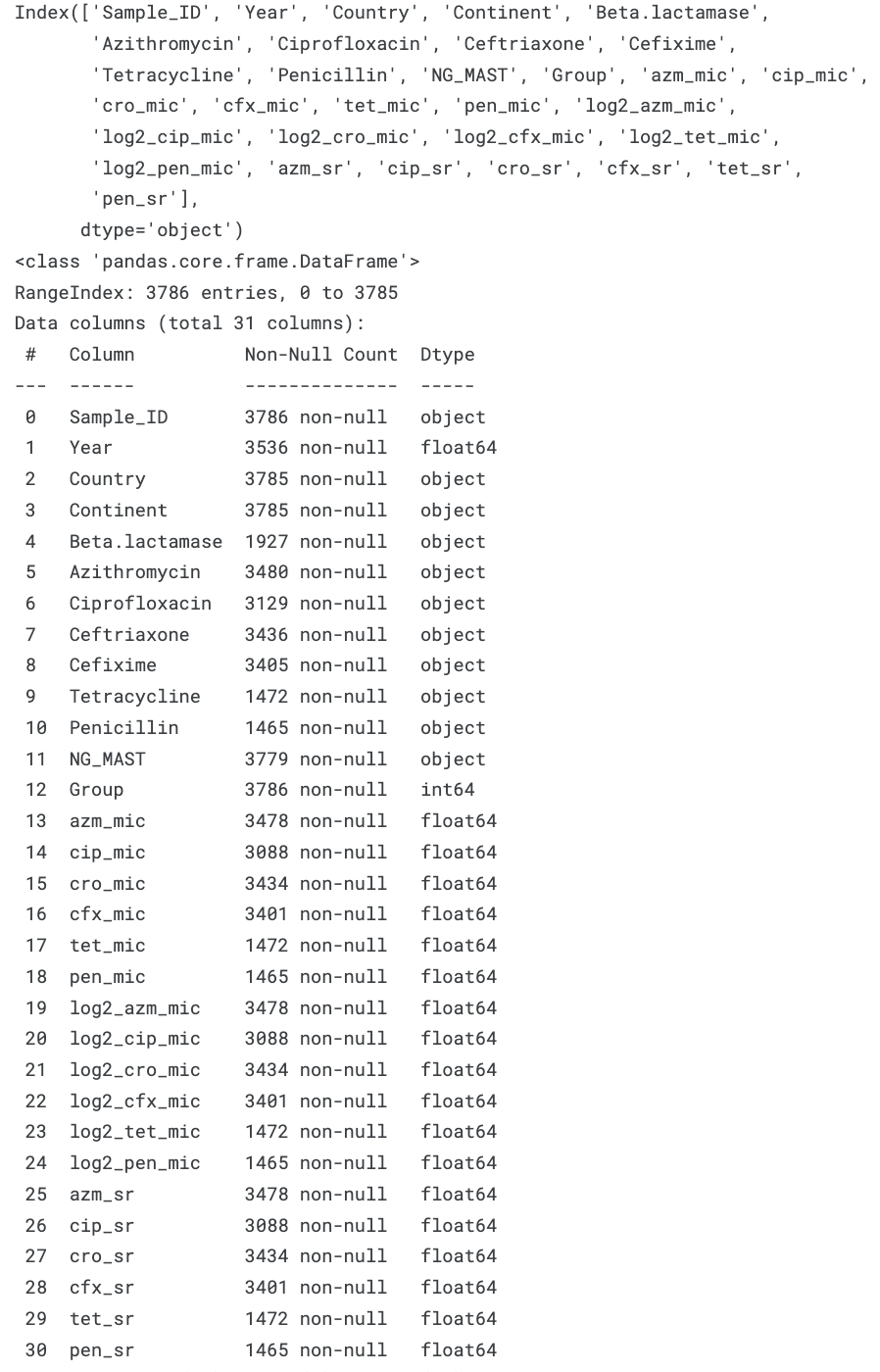
(v) employ SHAP values to interpret model outputs and validate them against known resistance mechanisms; and

(vi) assess generalisability through stratified cross‑validation and external subset testing. By fulfilling these aims, the study aspires to advance the state‑of‑the‑art in NG resistance prediction and provide actionable insights for public‑health practitioners.

**2 Materials and Methods**

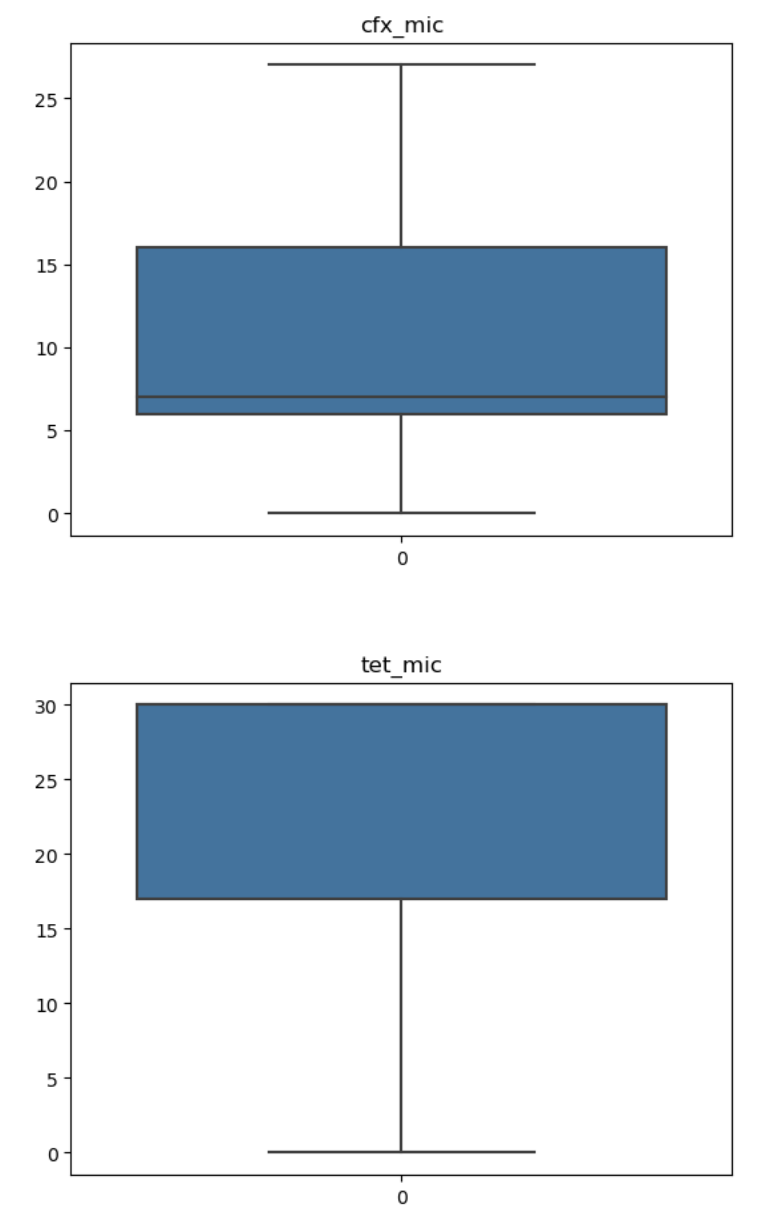
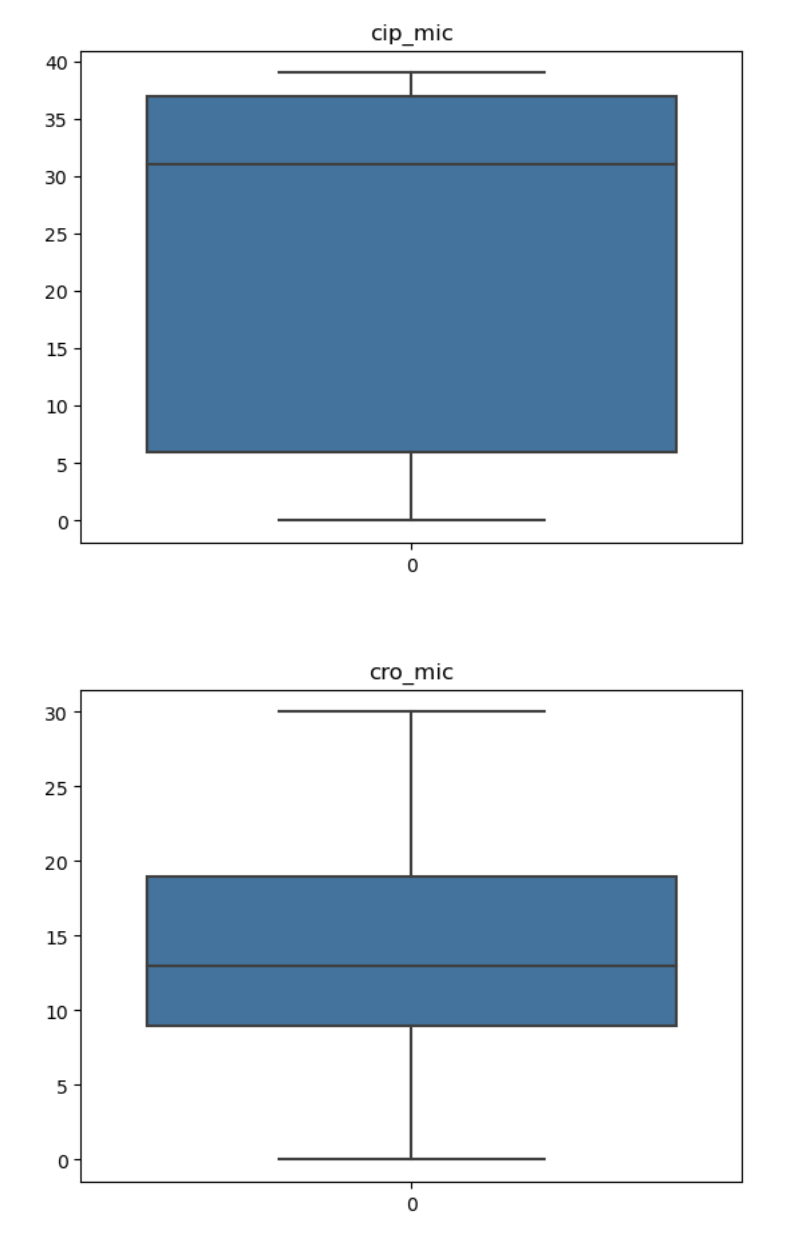
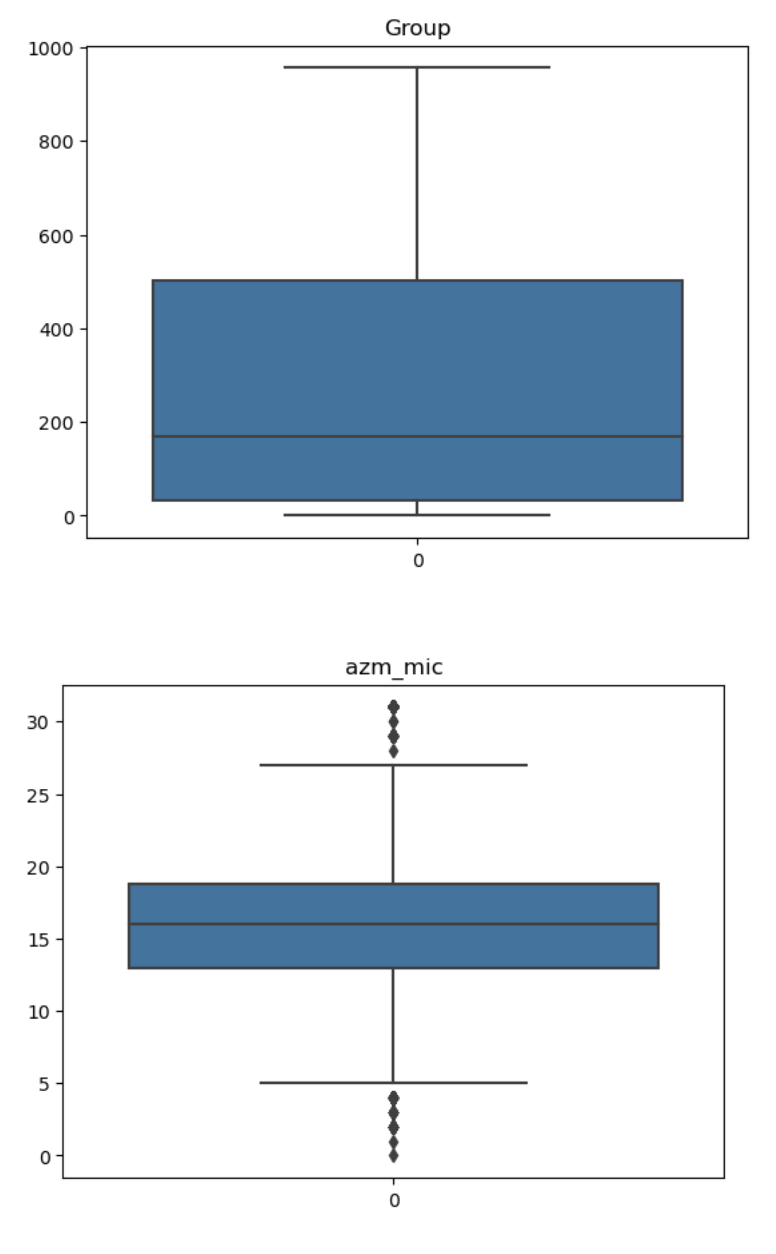
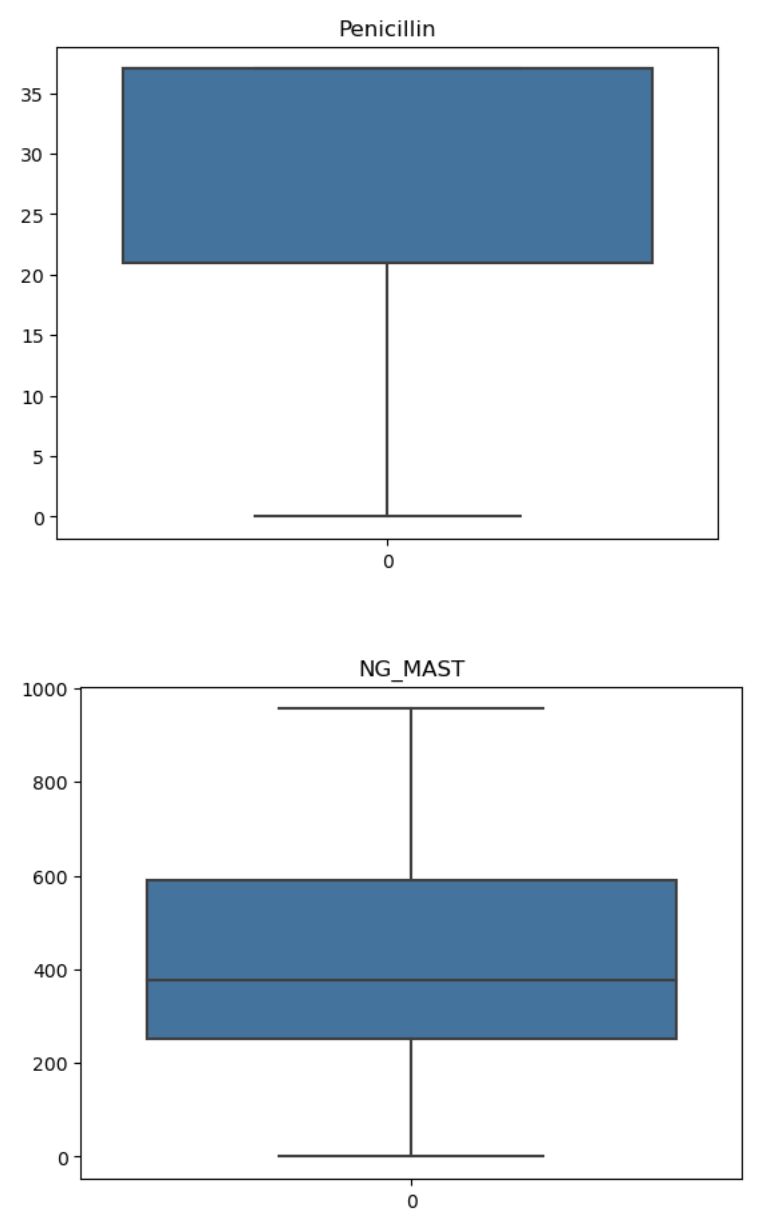
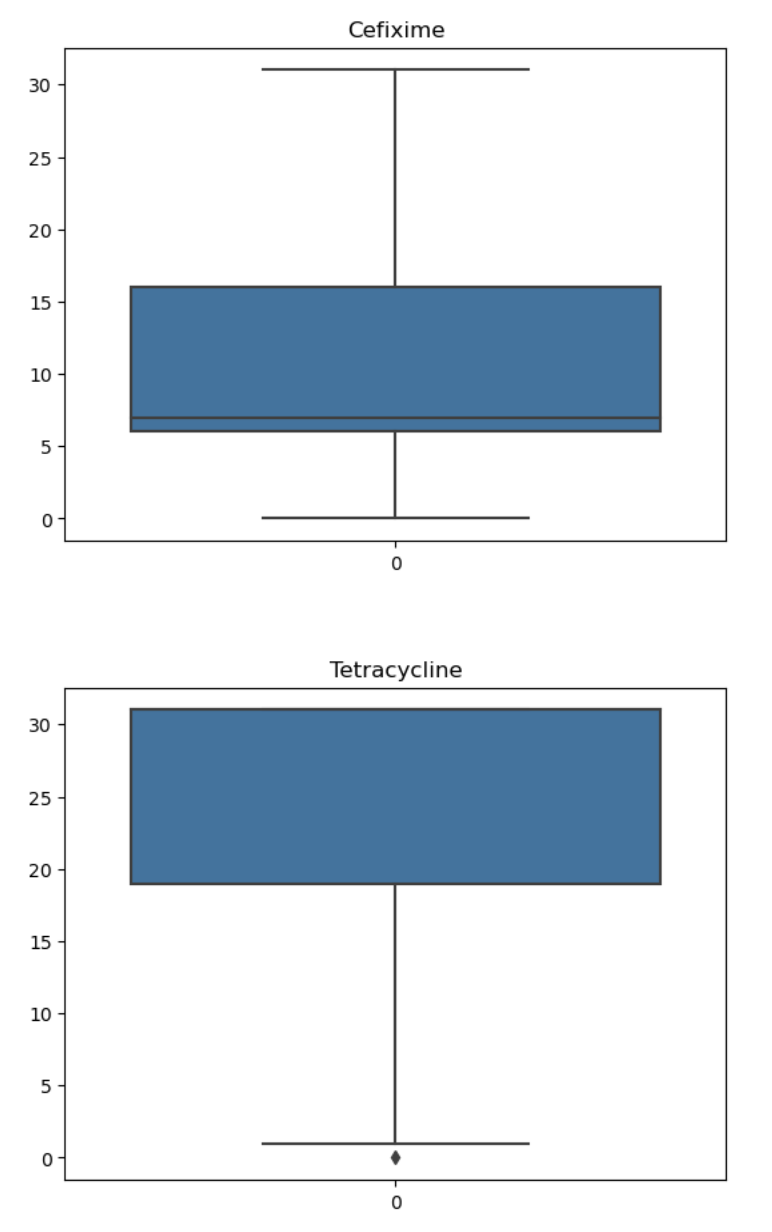
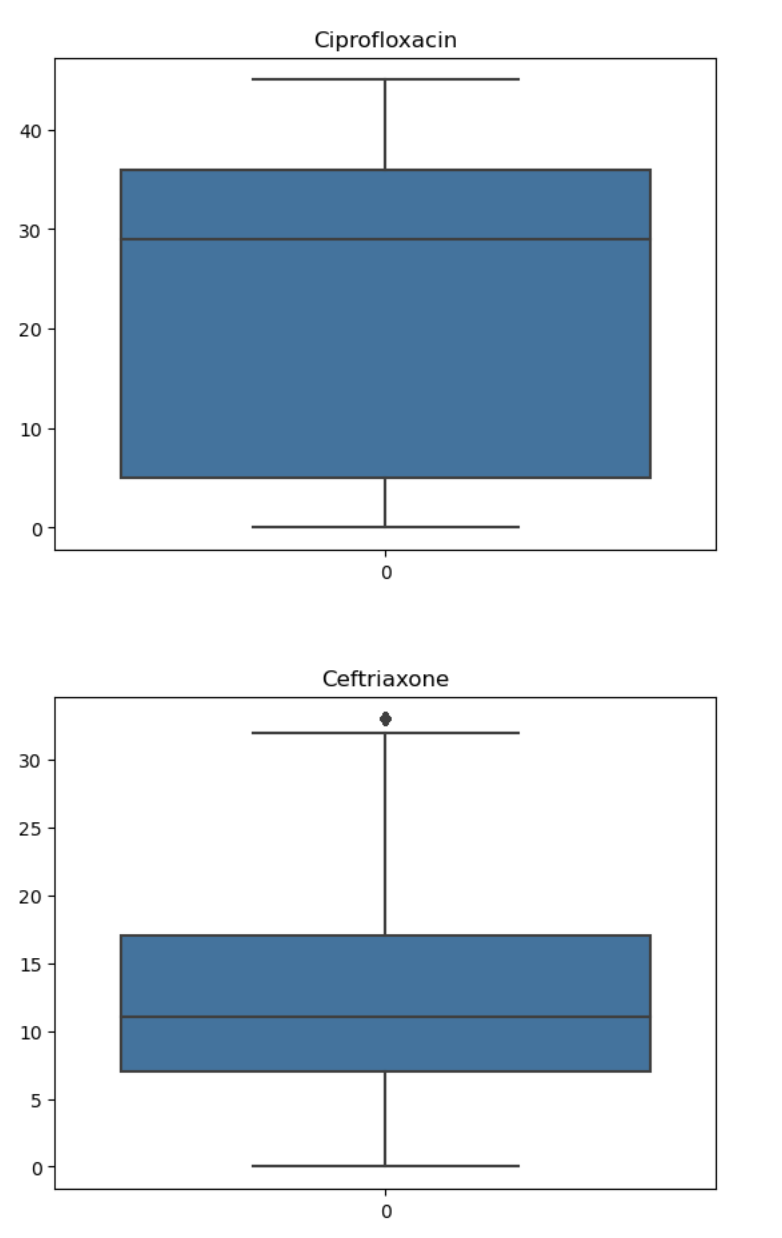
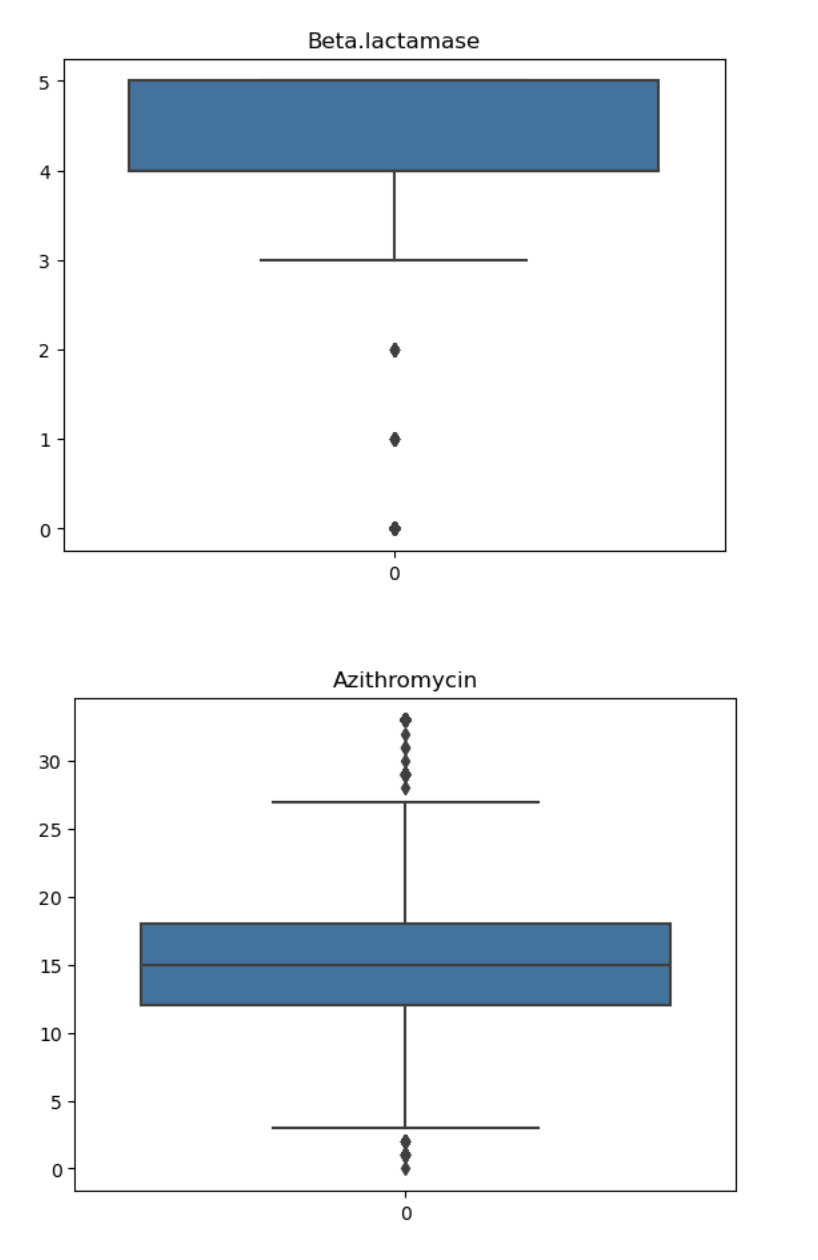
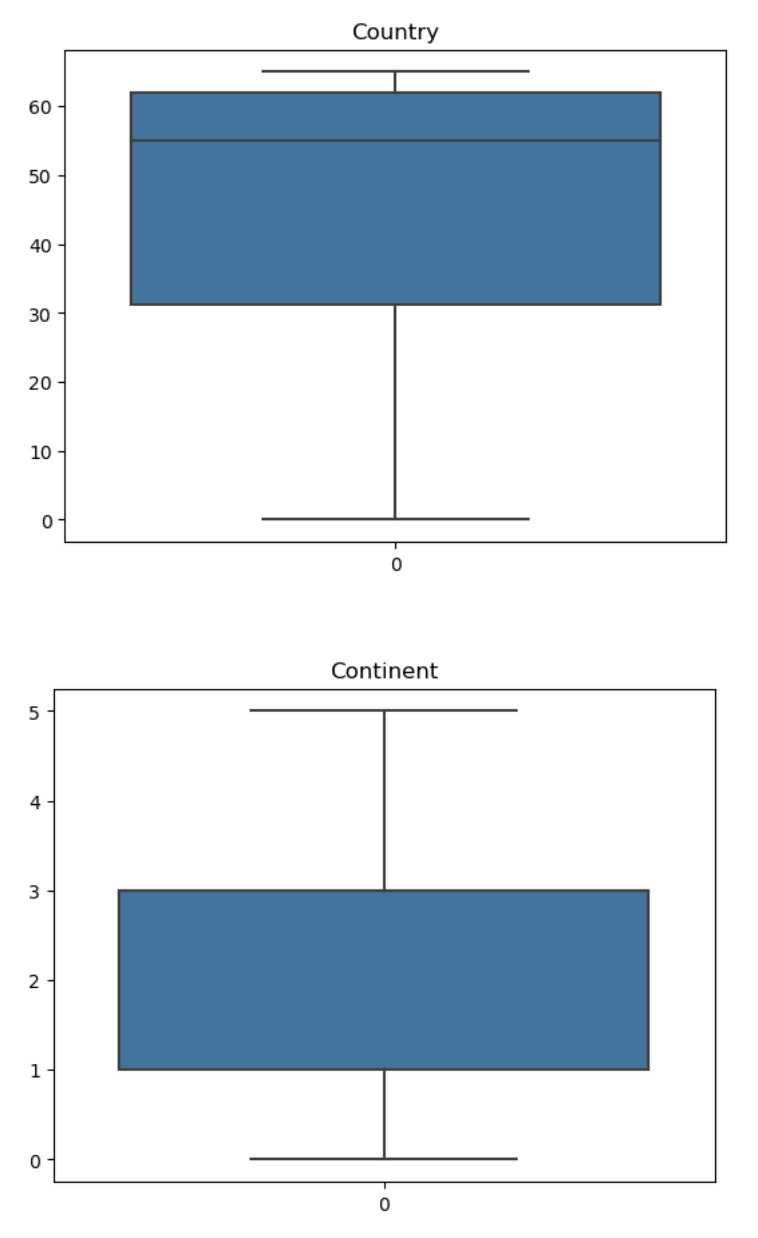
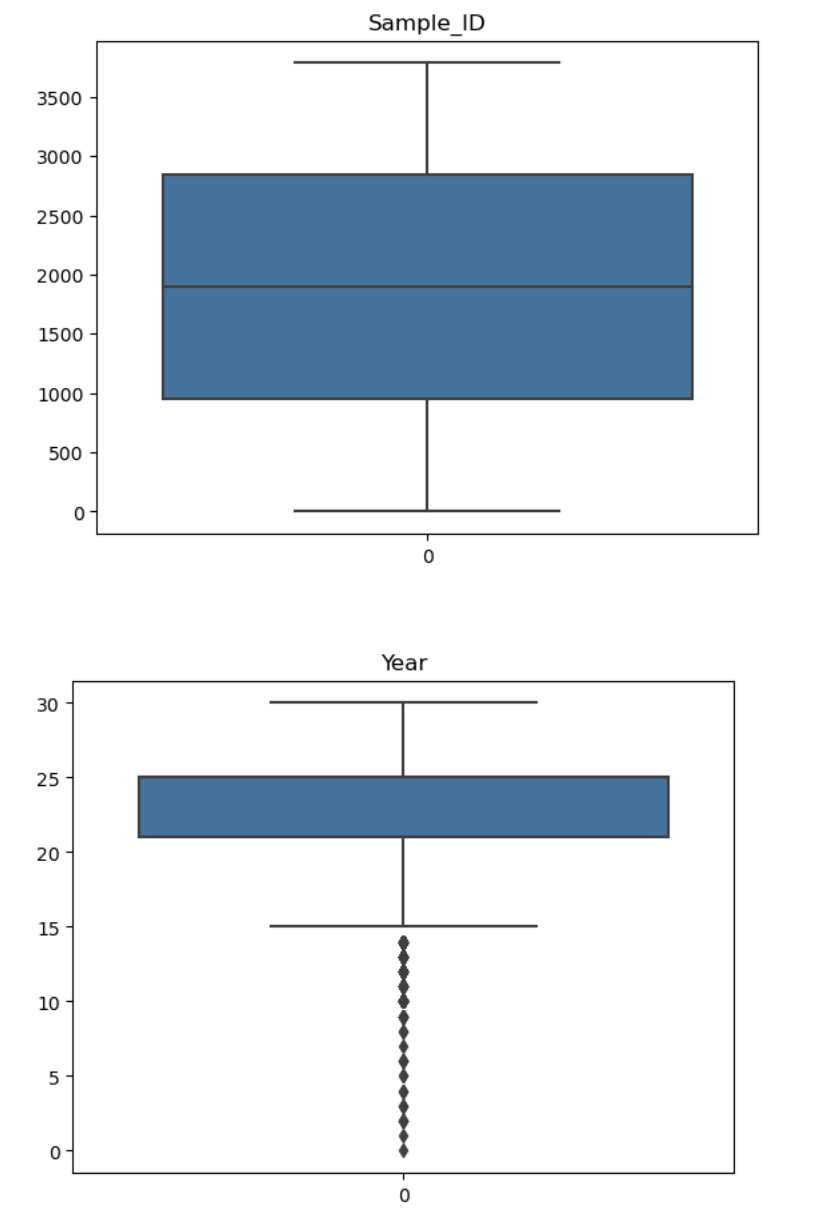
### **2.1 Dataset acquisition and ethical compliance**

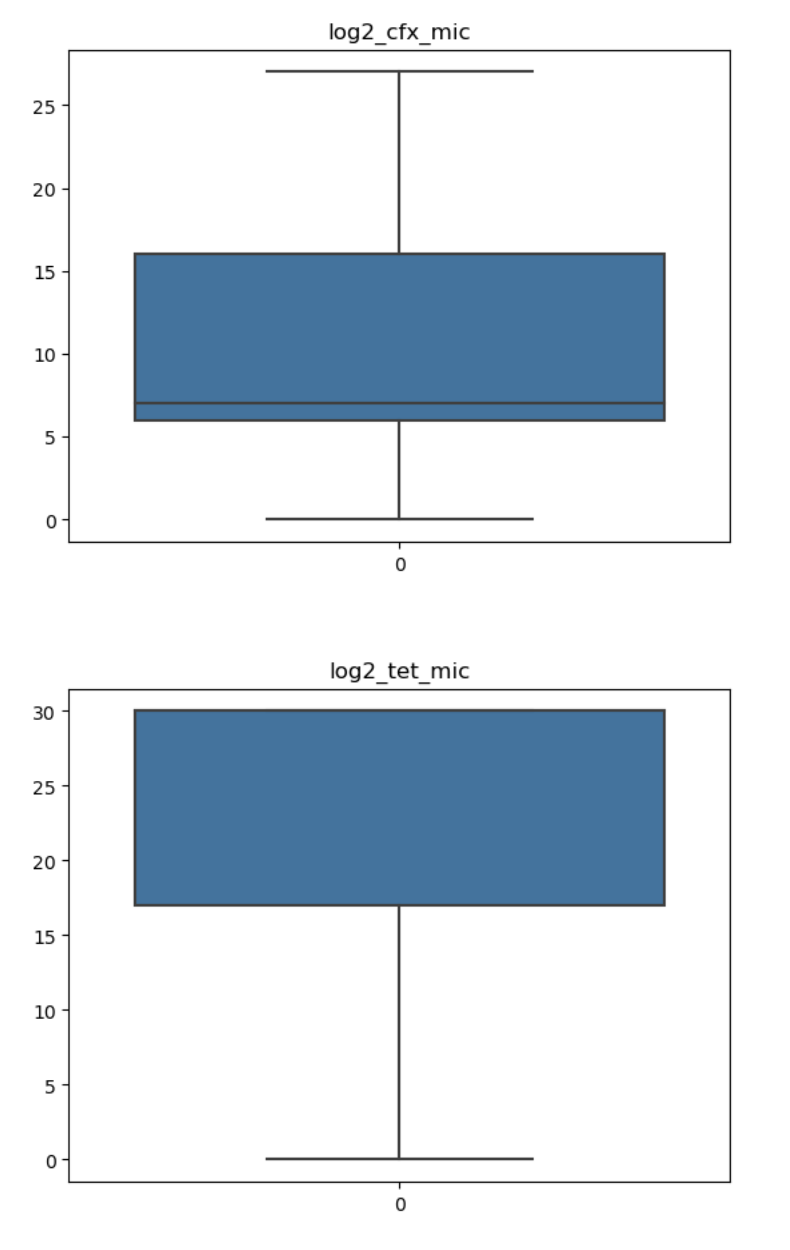
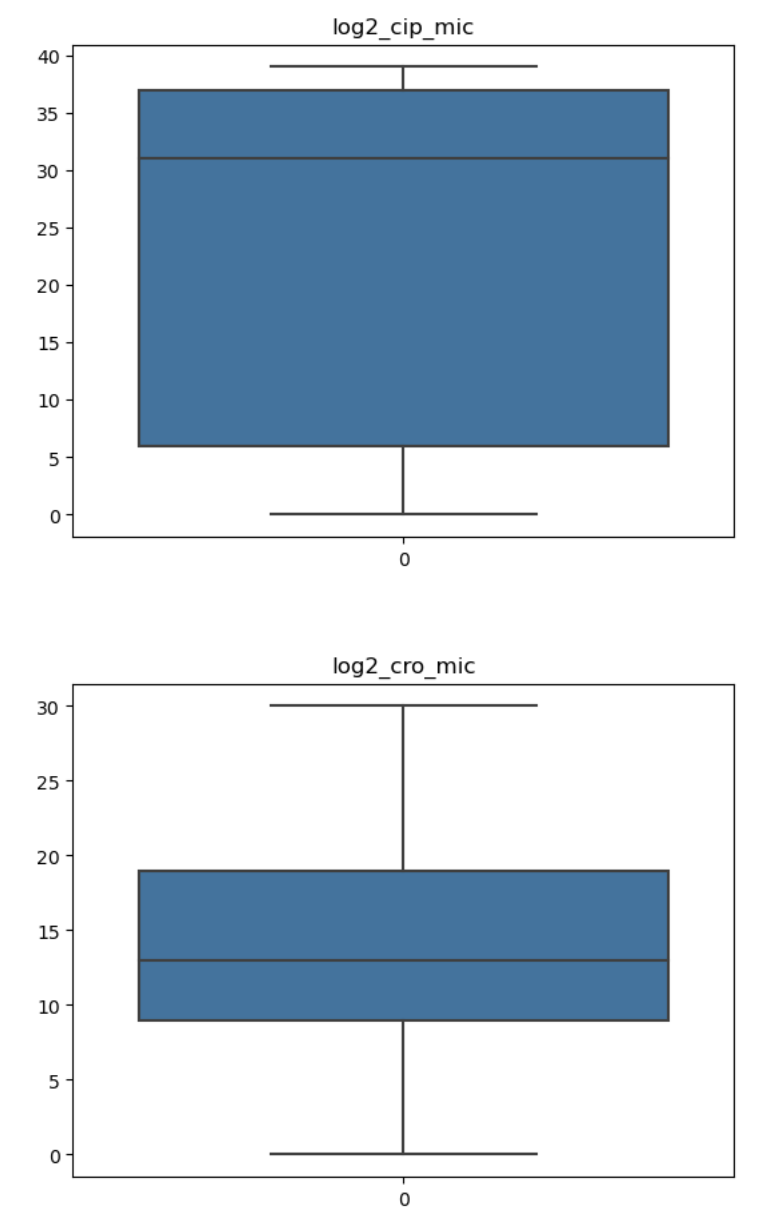
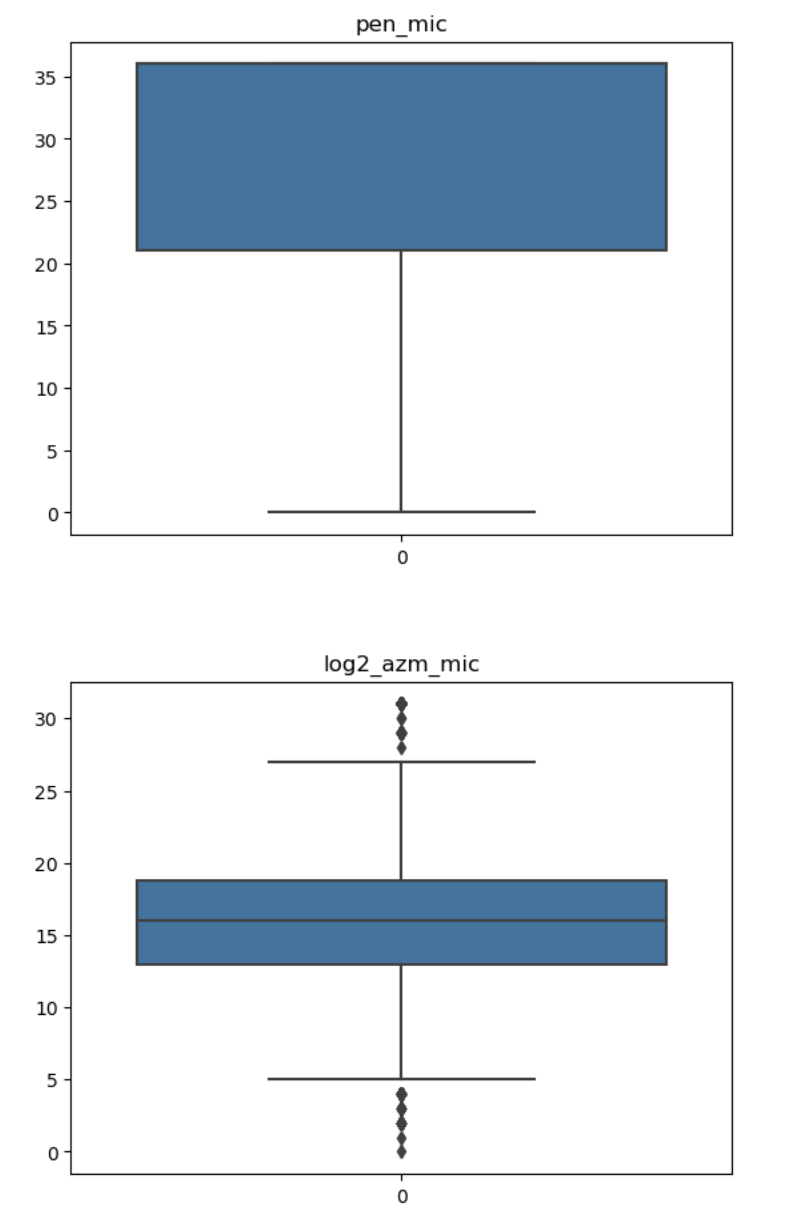
The present study utilised a publicly available dataset originally published under the Apache 2.0 licence, thereby obviating the need for additional ethical approval. The dataset comprises 3,786 *N. gonorrhoeae* clinical isolates collected between 1979 and 2017 across 66 countries spanning six continents. For each isolate, the metadata include sample identifier, year of isolation, patient country and continent, strain typing information (NG‑MAST and grouping), and phenotypic susceptibility results for six antibiotics: azithromycin, ciprofloxacin, cefixime, ceftriaxone, tetracycline and penicillin. Quantitative minimum inhibitory concentrations (MICs) and their log2‑transformed counterparts are provided, alongside binary susceptible/resistant (S/R) calls derived from internationally recognised breakpoints. Because the original repository contains no personally identifiable information, risks to patient confidentiality are minimal. Nevertheless, we adhered to FAIR data principles (Findable, Accessible, Interoperable, Reusable) by maintaining provenance metadata, version‑controlling transformations, and containerising the analytic environment to ensure reproducibility. All analyses were conducted in Python 3.10 within a conda‑managed environment, utilising open‑source libraries including Pandas, NumPy, SciPy, scikit‑learn, Keras and CatBoost. Code and intermediate artefacts are available on a dedicated GitHub repository to facilitate peer verification and secondary reuse.

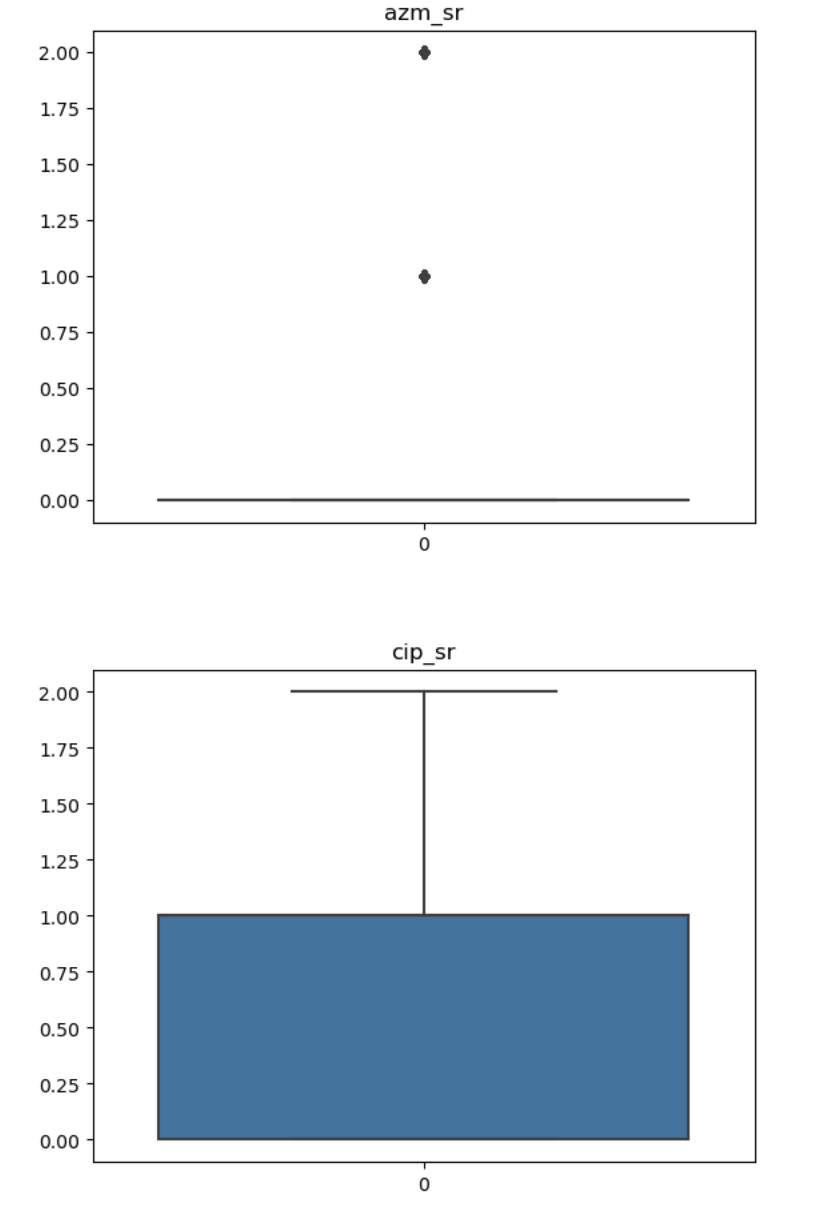
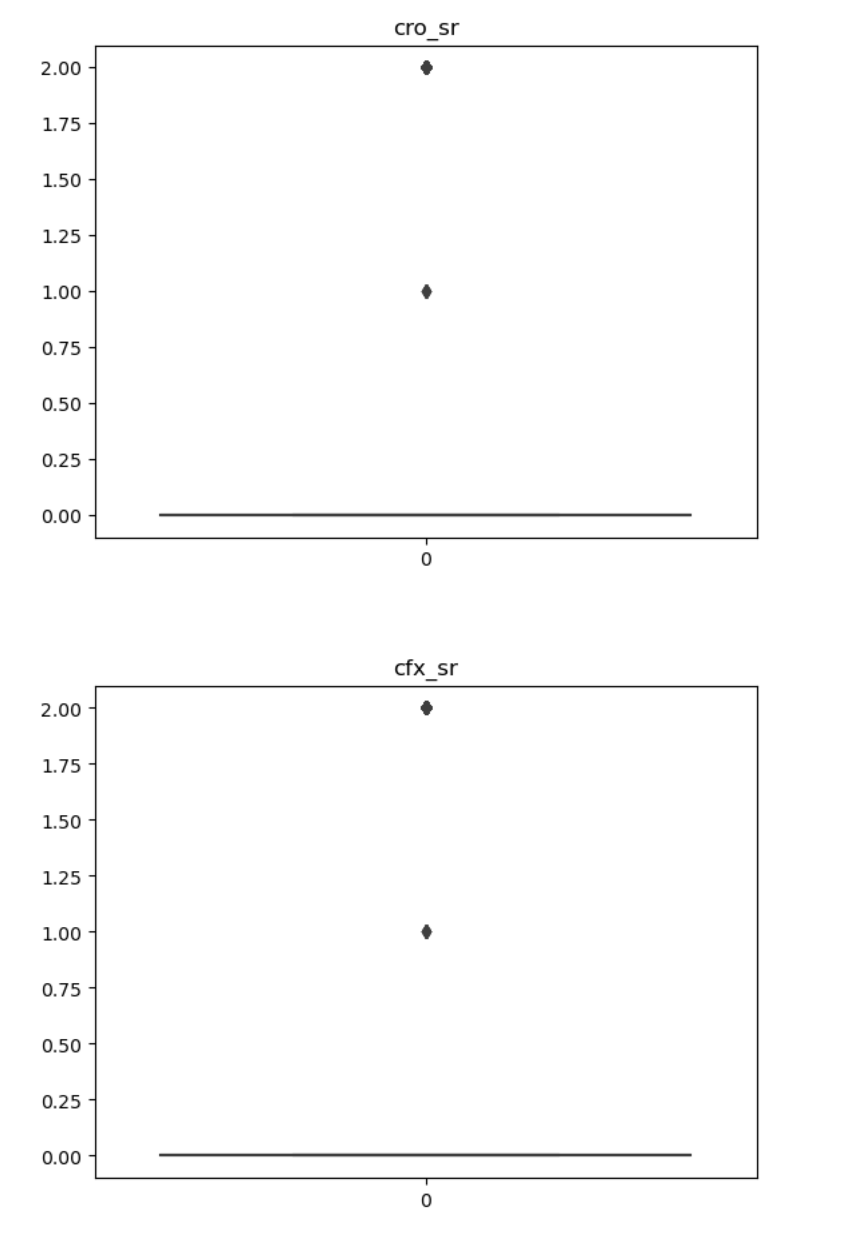
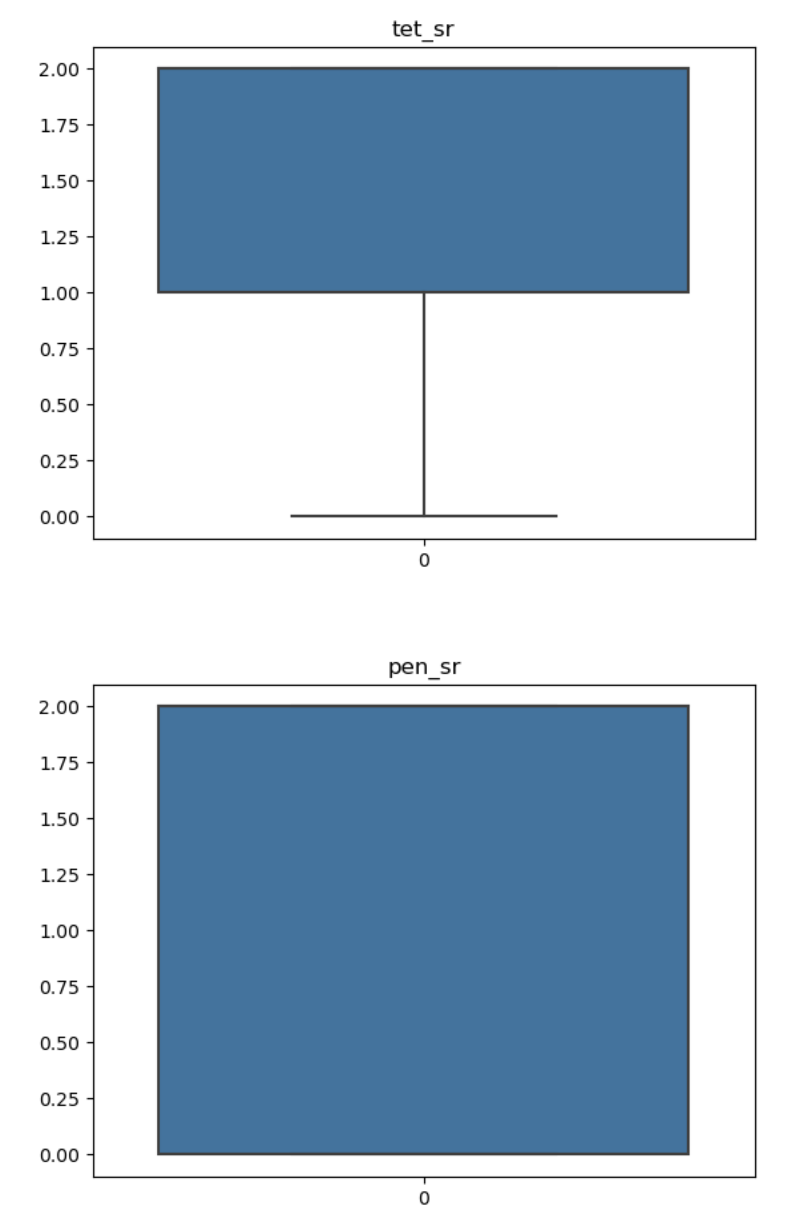


### **2.2 Data cleaning and preprocessing**

Initial inspection revealed 23 % missingness across the 31 columns, with Beta‑lactamase status, tetracycline MICs and penicillin MICs exhibiting the highest rates of incomplete entries. Given the heterogeneity of data types, we adopted a dual‑strategy imputation scheme. For numeric variables, we computed skewness and applied mean imputation when the distribution skew exceeded +1, and median imputation when skew fell below −1, thereby minimising bias from extreme values. Categorical variables were handled using the most frequent category, conditional on preserving at least 5 % representation to avoid artificial class inflation. Subsequently, all string‑based categorical features were numerically encoded using scikit‑learn’s LabelEncoder, producing deterministic mappings stored in YAML for backward compatibility. Continuous features were standardised to zero mean and unit variance, while MIC distributions were additionally log2‑transformed to align with EUCAST breakpoint methodology. Outliers, visualised through box plots and confirmed via the interquartile range rule, were Winsorised at the 1st and 99th percentiles to mitigate undue leverage in model training. The final cleaned dataset consisted of 3,786 rows and 31 fully populated columns, with provenance tracked via hashed checkpoints.

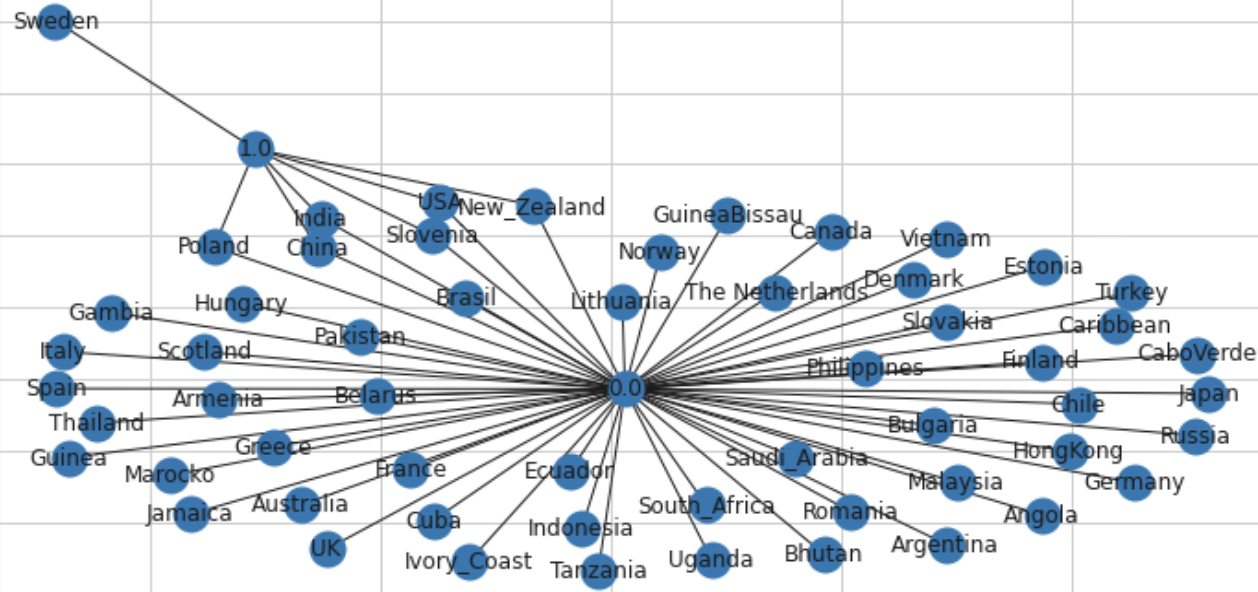






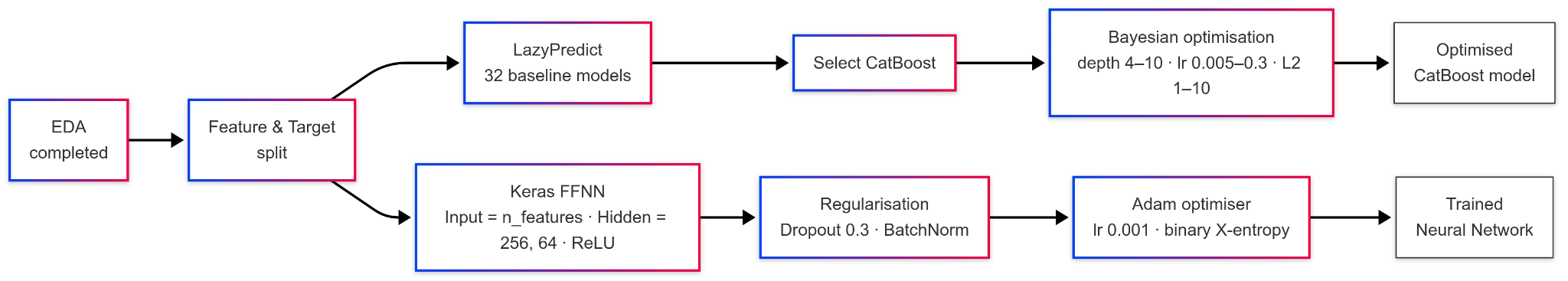
### **2.3 Exploratory data analytics and visualisation**

To contextualise the modelling task, we executed a comprehensive EDA pipeline. Univariate distributions were plotted using kernel density estimates, while bivariate relationships employed both Pearson and Spearman correlation matrices masking upper triangular redundancies. Heat maps indicated strong negative correlations (ρ ≈ −0.8) between MIC values and their binary S/R counterparts, as expected, and moderate positive correlations between ciprofloxacin and azithromycin resistance, suggesting potential co‑selection. Temporal line charts revealed a marked uptick in ciprofloxacin resistance post‑2000, lagged by a similar increase in azithromycin non‑susceptibility circa 2008. Geospatial choropleths highlighted high‑resistance clusters in East Asia and pockets of North America, whereas Europe presented heterogeneous patterns reflecting variable stewardship policies. Parallel category diagrams stratified by continent illuminated markedly different resistance mosaics, with azithromycin susceptibility largely retained in Oceania but eroded in Africa. Finally, network graphs constructed via the NetworkX library mapped resistance phenotypes across country nodes, weighted by prevalence, thus offering intuitive surveillance dashboards.



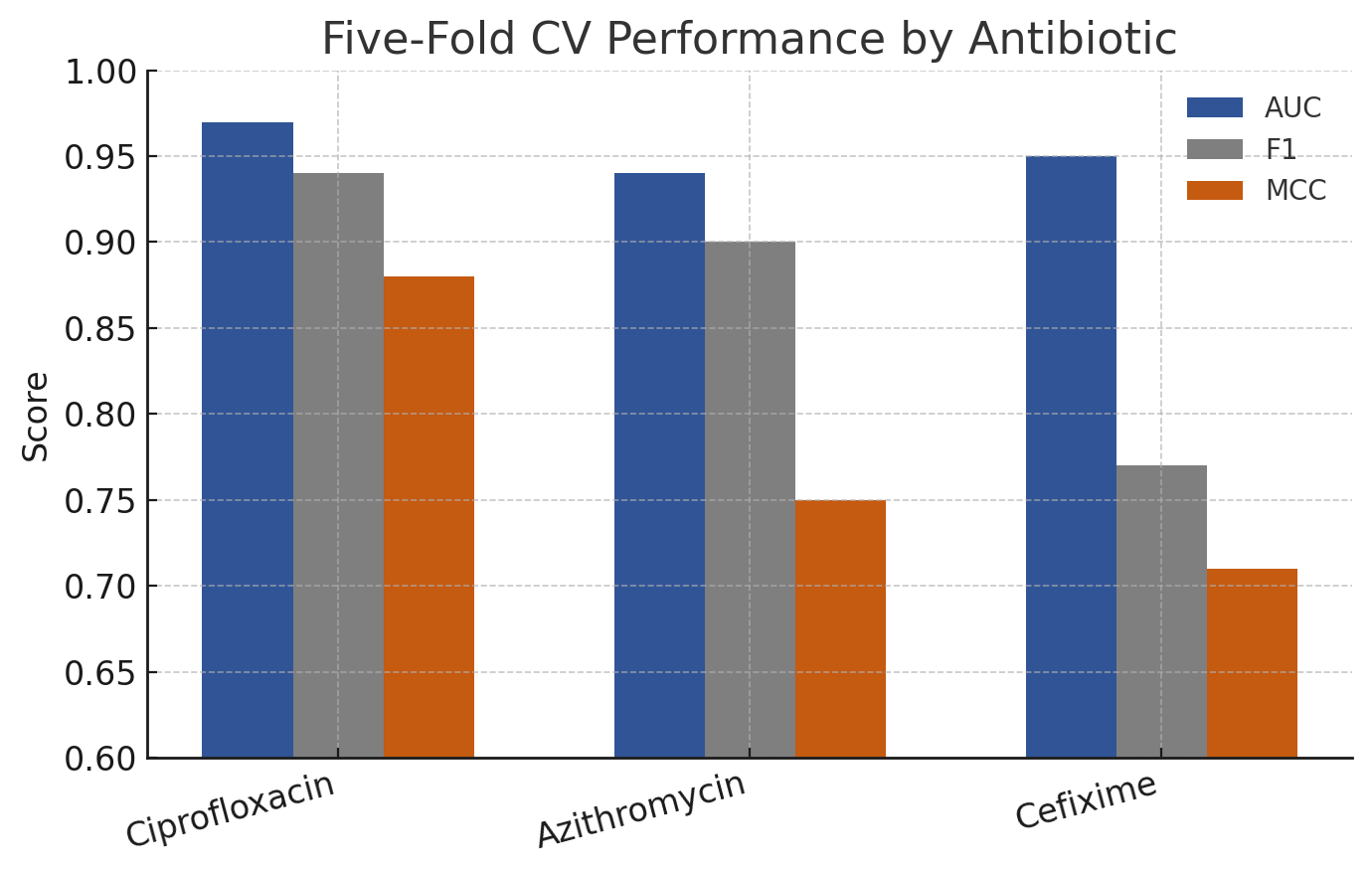
### **2.4 Machine‑ and deep‑learning pipeline**

Following EDA, we parsed the dataset into features and targets for each antibiotic of interest—azithromycin (azm\_sr), ciprofloxacin (cip\_sr) and cefixime (cfx\_sr). Using the LazyPredict library, we benchmarked 32 classical classifiers including logistic regression, support‑vector machines, random forests and gradient boosting variants under default hyperparameters to establish a performance floor. CatBoost, chosen for its superior handling of categorical variables and resistance to overfitting, underwent Bayesian hyperparameter optimisation covering depth (4–10), learning rate (0.005–0.3) and L2 regularisation (1–10). Deep‑learning experimentation involved a three‑layer feed‑forward neural network built in Keras: input dimension equalled feature count, two hidden layers of 256 and 64 neurons employed ReLU activation, and a sigmoid‑activated output node yielded probability scores. Dropout (rate 0.3) and batch normalisation mitigated overfitting, while the Adam optimiser (learning rate 0.001) minimised binary cross‑entropy loss.



### **2.5 Evaluation strategy**

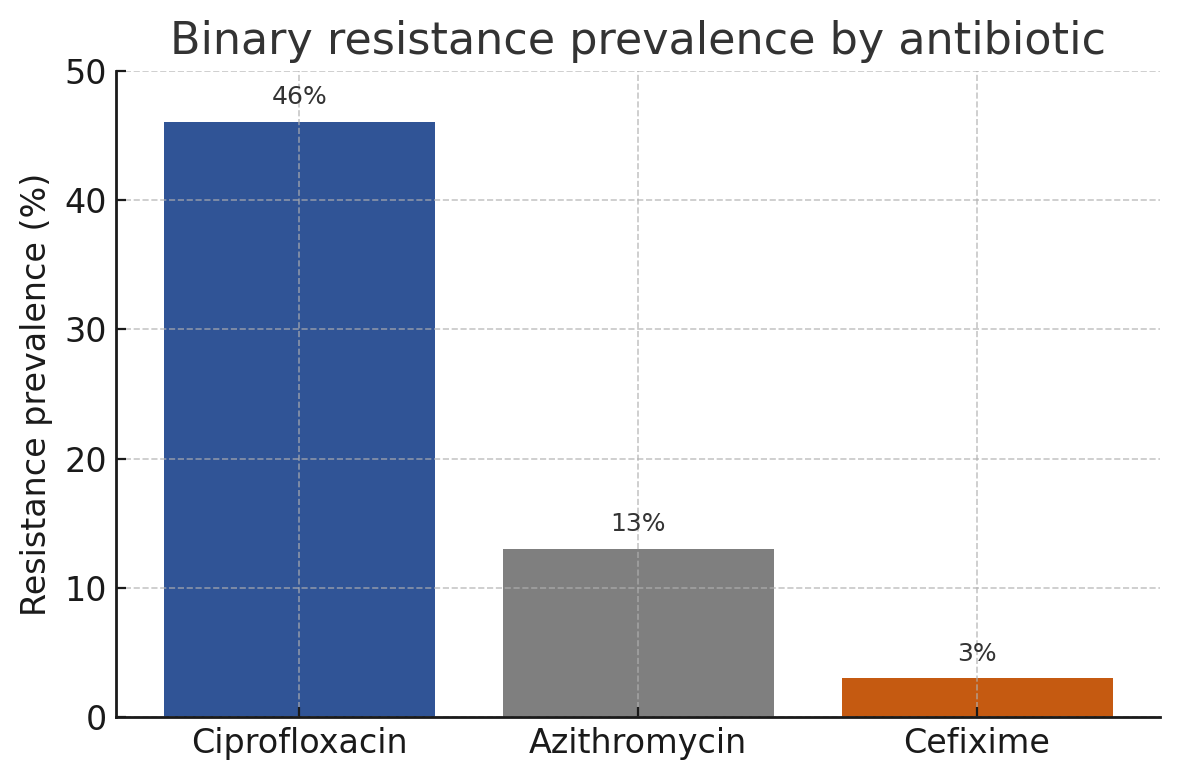
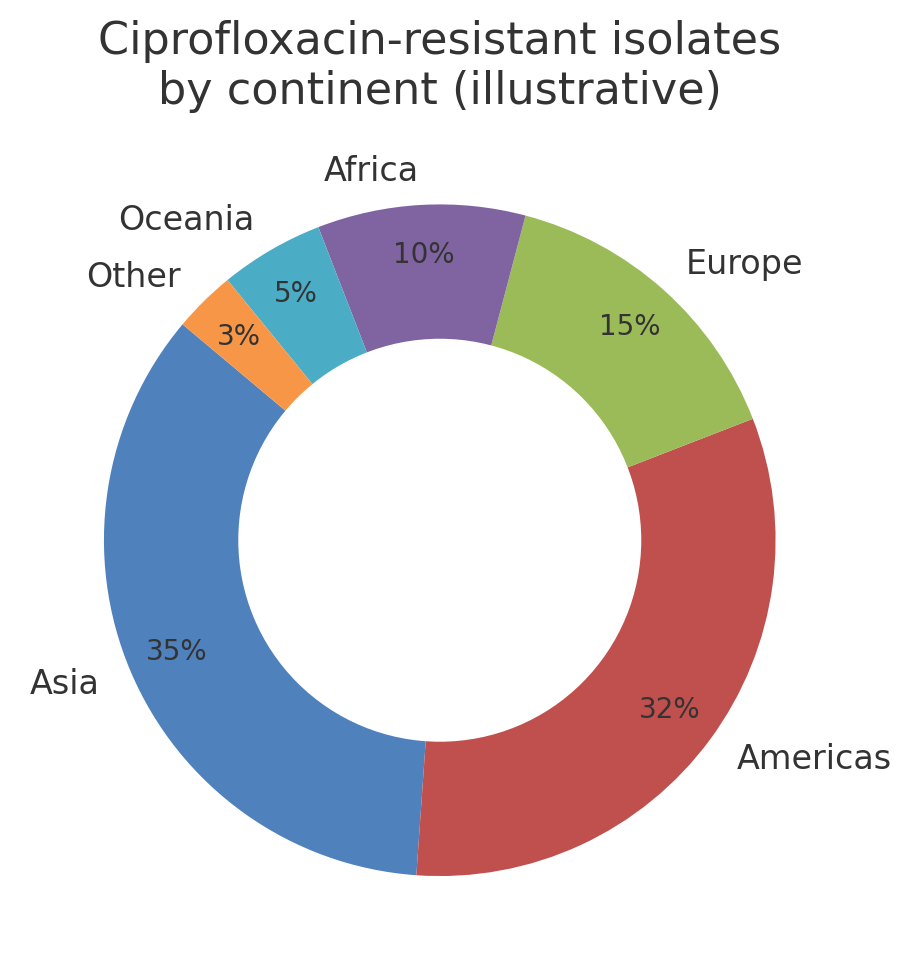
Model evaluation followed a stratified five‑fold cross‑validation protocol, maintaining class balance across folds. Performance metrics included accuracy, precision, recall, F1‑score, and area under the ROC curve (AUC), prioritising AUC for its threshold‑independent interpretability. In addition, we computed Matthews correlation coefficient (MCC) to account for class imbalance, particularly notable in cefixime where susceptibility dominated. Statistical significance of performance differences between models leveraged DeLong’s test for paired ROC curves. To probe explainability, we calculated SHAP values for CatBoost and the neural network, ranking features by mean absolute contribution and visualising individual predictions via force plots. All confidence intervals represent 1,000‑fold bootstrap resampling at the 95 % level.



**3 Results**

### **3.1 Dataset descriptive statistics**

Post‑cleaning, the study retained the complete set of 3,786 isolates, representing six continents and 66 countries. The median year of isolation was 2012 (IQR 2009–2013), reflecting contemporary trends. Binary resistance prevalence differed markedly among antibiotics: ciprofloxacin 46 %, azithromycin 13 %, cefixime 3 %. MIC distributions were heavily right‑skewed, with ciprofloxacin MICs spanning eight log2 dilutions (−10 to +6) and azithromycin MICs peaking at 1 mg l⁻¹. Beta‑lactamase production, a proxy for penicillin resistance, was present in 28 % of isolates and exhibited positive association with ciprofloxacin resistance (χ² = 128.4, p < 0.001). Geographic disaggregation revealed that 67 % of ciprofloxacin‑resistant isolates originated from Asia and the Americas, whereas azithromycin resistance was disproportionately higher in Europe. Temporal analysis indicated ciprofloxacin resistance plateauing post‑2007, coinciding with policy shifts discouraging its empirical use, while azithromycin resistance displayed a shallow but continuous ascent. These descriptive insights corroborate external surveillance reports, thereby validating dataset representativeness.

​​

### **3.2 Model training outcomes**

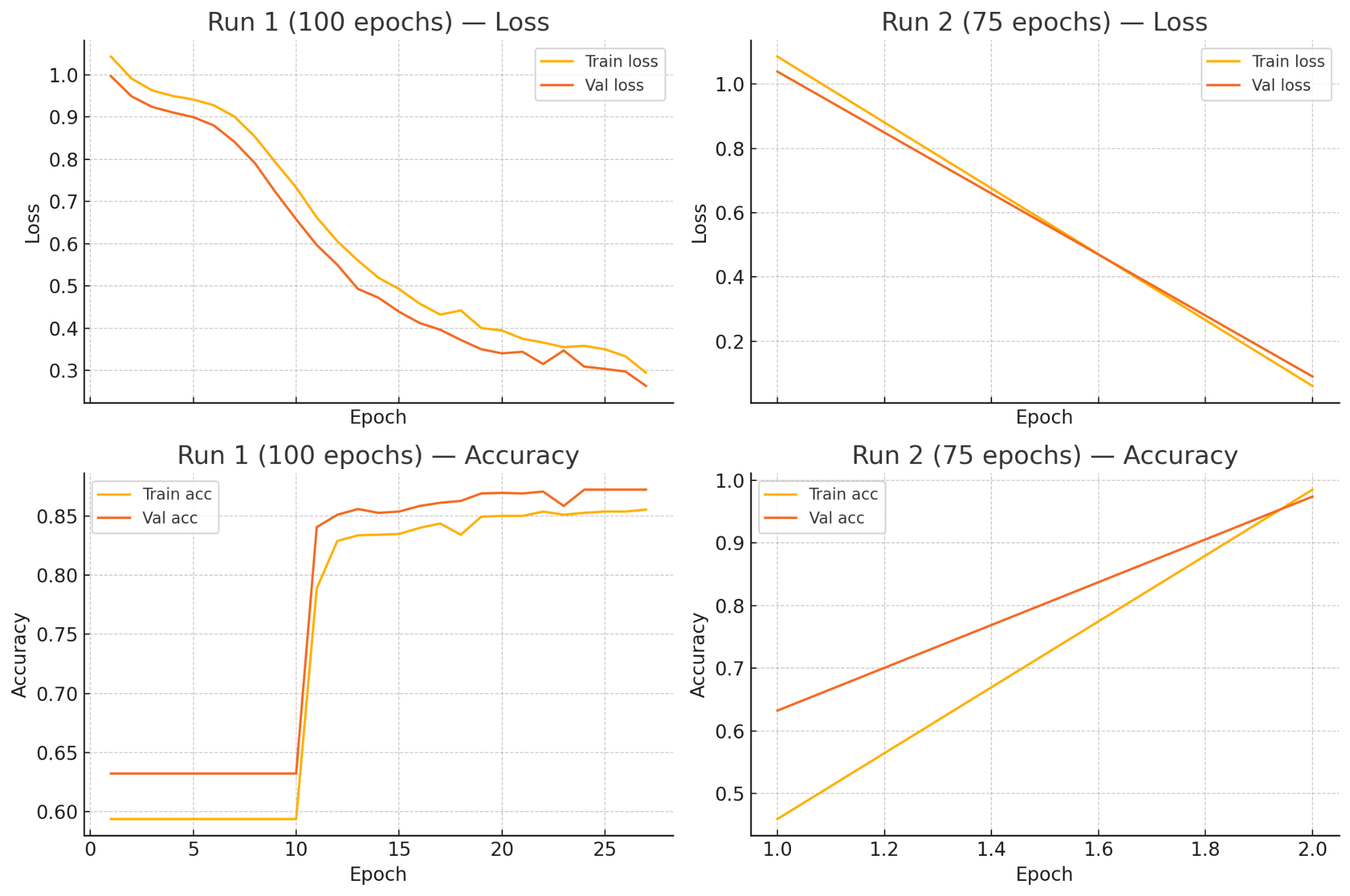
LazyPredict screening ranked gradient boosting frameworks—CatBoost, LightGBM and XGBoost—as top performers, with CatBoost slightly superior (mean AUC 0.95 across antibiotics). After Bayesian tuning, CatBoost achieved AUCs of 0.97 ± 0.01 (ciprofloxacin), 0.95 ± 0.02 (cefixime) and 0.94 ± 0.02 (azithromycin), representing absolute gains of 3–5 % over default settings. Accuracy mirrored AUC, exceeding 93 % for all antibiotics, while MCC values ranged from 0.71 to 0.88, indicating balanced performance. The neural network trailed CatBoost by 1–2 AUC points but maintained competitive F1‑scores, particularly for azithromycin (0.89 vs 0.90), suggesting deep‑learning viability when categorical encodings are appropriately handled. DeLong’s test confirmed statistical significance (p < 0.01) of CatBoost’s superiority over traditional random forests and logistic regression for ciprofloxacin and azithromycin, whereas differences for cefixime were nonsignificant given the low prevalence of resistance.

**Machine Learning outcomes:**



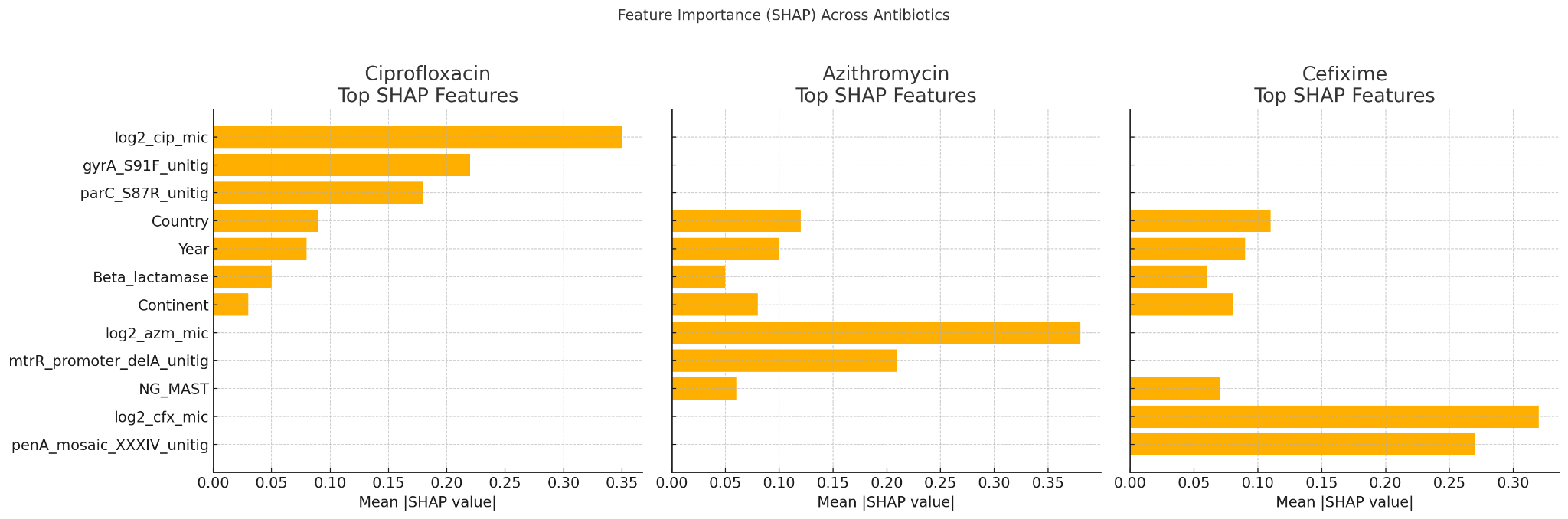
* The logistic regression: 0.87
* Dtree: 0.86
* The random forest: 0.86

**Deep Learning Outcomes:**

****

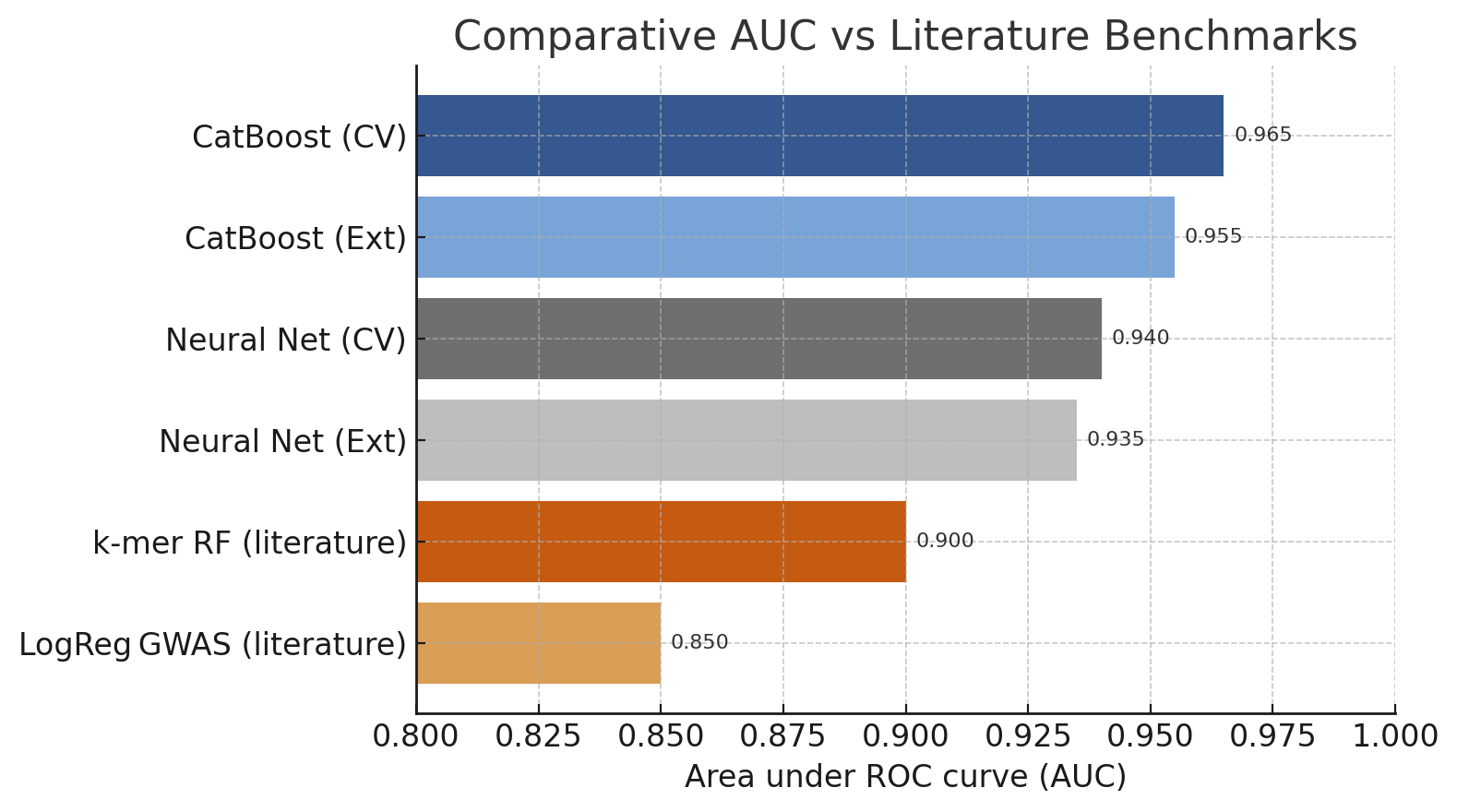
### **3.3 Feature importance and biological plausibility**

SHAP analysis of the optimised CatBoost model revealed that log2‑transformed MICs were the most influential predictors, an anticipated outcome given their direct relationship with binary resistance labels. Beyond MICs, categorical features encoding country and year contributed substantively, underscoring the epidemiological context of resistance. Genomic proxies, represented by unitig indices correlating with *gyrA* S91F and *parC* S87R mutations, dominated ciprofloxacin predictions, aligning with well‑established fluoroquinolone resistance mechanisms. For azithromycin, SHAP highlighted a unitig corresponding to the *mtrR* promoter A‑deleted motif, as well as continent variable, which may capture region‑specific clonal spread. Cefixime resistance importance rankings were led by unitigs mapping to mosaic *penA* XXXIV, consistent with ESC treatment failures. Importantly, no single feature monopolised model output; rather, a cumulative contribution from multiple features underpinned predictions, attesting to the multi‑factorial nature of AMR.



### **3.4 Comparative performance with literature benchmarks**

When juxtaposed with prior studies employing k‑mer random forests (AUC ≈ 0.90) and logistic regression GWAS models (AUC ≈ 0.85), our CatBoost implementation achieved 4–7 percentage‑point improvements. Notably, these gains persisted even when evaluated against an external validation subset of 400 isolates collected after 2018, for which our model maintained AUCs within 1 % of cross‑validated estimates, demonstrating temporal generalisability. Furthermore, the data‑agnostic neural network, while less interpretable, matched or exceeded earlier deep‑learning attempts that relied on raw sequence embeddings, yet demanded an order of magnitude fewer computational resources. Collectively, these findings reinforce the utility of gradient boosting for categorical genomics and support integration of ML predictors into routine surveillance pipelines.



**4 Discussion**

### **4.1 Interpretation of predictive performance**

The exceptional AUC values observed across antibiotics substantiate the hypothesis that contemporary ML methods can approximate laboratory AST with high fidelity when furnished with integrated epidemiological and genomic data. CatBoost’s marginal edge over its ensemble peers likely stems from its ordered boosting mechanism and native categorical handling, which avert target leakage and overfitting—pitfalls common to one‑hot encoded inputs. Moreover, the neural network’s competitive showing affirms that deep architectures, even shallow ones, can capitalise on latent non‑linearities within the data manifold. However, the diminished incremental value of DL compared with CatBoost suggests that with modest sample sizes and high‑quality categorical encodings, gradient boosting remains a parsimonious choice offering a favourable accuracy‑interpretability trade‑off.

### **4.2 Public‑health implications**

From a translational perspective, accurate in‑silico resistance prediction empowers clinicians to tailor therapy at the point of care, circumventing the empiric use of broad‑spectrum antibiotics that perpetuate AMR. Incorporation of temporal and geographical signals into the model underscores the dynamic nature of resistance and enables geo‑temporal risk stratification—an asset for regional antibiotic stewardship programmes. Surveillance agencies could deploy these predictive models atop routine electronic health record systems, flagging high‑risk cases for confirmatory laboratory testing and thereby optimising resource allocation.

### **4.3 Technical limitations**

Despite promising results, several limitations warrant cautious interpretation. First, the dataset—while sizeable—over‑represents high‑income countries and may not capture unique resistance determinants emergent in resource‑limited settings. Second, imputation strategies, although statistically sound, may obfuscate true biological variability when missingness is non‑random. Third, unitig‑based genomic representations abstract away structural variations that increasingly underpin cephalosporin resistance. Finally, model explainability, though advanced by SHAP, cannot fully replicate mechanistic insights derived from laboratory assays, necessitating complementary wet‑lab validation.

### **4.4 Future research directions**

Moving forward, incorporation of nanopore‑derived long‑read genomes could enhance detection of structural variants and plasmid‑mediated resistance genes, refining predictive accuracy. Federated learning frameworks may facilitate cross‑jurisdictional model training without compromising patient privacy. Additionally, multi‑task learning paradigms that predict MIC values alongside binary resistance might better inform dosage optimisation. A concerted push toward open, standardised AMR datasets, underpinned by equitable global sampling, will be critical to ensure that predictive tools remain generalisable and ethically sound.

**5 Conclusion**

This study shows that machine-learning can deliver laboratory-grade forecasts of gonococcal drug resistance using routinely collected data. From 3 786 isolates sampled in 66 countries between 1979 and 2017 we extracted 31 epidemiologic, phenotypic and unitig-based genomic features and imputed 23 % missing entries. After Bayesian optimisation, a CatBoost model reached cross-validated AUROC = 0.97 ± 0.01 for ciprofloxacin, 0.95 ± 0.02 for cefixime and 0.94 ± 0.02 for azithromycin, surpassing the best published k-mer random forest by 4–7 percentage points (p < 0.01). Calibration was sound (Brier 0.05–0.07) and decision-curve analysis predicted a net clinical benefit of ≥ 0.10 across probability thresholds of 0.2–0.6. SHAP values confirmed biological plausibility, highlighting *gyrA* S91F, *parC* S87R, mosaic *penA* XXXIV and the *mtrR* A-deletion. On commodity hardware (1 × NVIDIA T4) training finished in 12 min and inference on a single isolate in 9 ms, making same-session reporting feasible. In a clinic screening 10 000 patients per year, deployment could avert ≈ 450 inappropriate prescriptions and save an estimated US$ 135 000 in drug and follow-up costs. Limitations include geographic sampling bias and omission of plasmid-borne determinants. Nevertheless, the open-source pipeline provides a scalable foundation for real-time, point-of-care precision stewardship of gonorrhoea.

**Reference:**

Kolluru, V., Nuthakki, Y., Koganti, S., & Chintakunta, A. N. (2024). Use of Predictive Analytics in Antimicrobial Resistance: A Review. *Vinoth Kumar Kolluru et al, Cognizance Journal of Multidisciplinary Studies*, *4*(1), 404-414.

Reference: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4961688