Machine Learning Prediction of Susceptible versus Non-susceptible Antimicrobial Susceptibility from Quality-Controlled Inpatient Microbiology Data

Jittiphat Sanom[1], Chanchanok Aramrat[2], Kittikorn Hantrakul[1], Part Pramokchon[1], Alongkot Gongmanee[1], and Somnuk Sinthupuan[1*]

[1]Department of Computer Science, Faculty of Science, Maejo University
[2]Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand

*Corresponding Author Email: somnuk@mju.ac.th

Abstract

Hospital microbiology data are frequently affected by missing values, duplicate records, inconsistent naming, and heterogeneous coding across information systems, all of which limit their suitability for machine-learning analysis. This study developed a machine-learning framework to predict antimicrobial susceptibility as a binary outcome, classified as susceptible (S) versus non-susceptible (non-S; combining intermediate and resistant), using quality-controlled inpatient hospital microbiology and patient-context data.

An end-to-end pipeline was designed to evaluate data quality and standardize organism names, specimen types, antibiotic names, and susceptibility-result fields using dictionary mapping, regular expressions, and fuzzy matching. Records that could not be interpreted reliably were flagged as REVIEW and excluded from supervised learning. The target variable was transformed from S/I/R into a binary outcome, where S was interpreted as susceptible and I or R was classified as non-susceptible.

A total of **1,097,459 inpatient culture records** were initially obtained from the IPD Culture database. After linking IPD culture records with IPD patient-context data, **528,134 records** were available. After excluding records with missing modeling features, **497,542 complete IPD modeling records** remained, comprising **337,302 susceptible records** and **160,240 non-susceptible records**. CTGAN-based augmentation generated **177,062 synthetic non-susceptible records**, resulting in a balanced analytical dataset of **674,604 records** with 337,302 records per class.

Four tree-based machine-learning models—Decision Tree, Random Forest, XGBoost, and CatBoost—were trained to predict susceptible versus non-susceptible outcomes under three experimental settings with sequential feature removal. Model performance was evaluated using accuracy, macro precision, macro recall, and macro F1-score. Random Forest achieved the best performance in all three settings. With the full feature set, Random Forest achieved **81.66% accuracy** and a **macro F1-score of 0.8166**. When organism information was removed, accuracy decreased to **72.33%**, and when both organism and ward information were removed, accuracy further decreased to **67.04%**. These findings indicate that systematically quality-controlled inpatient microbiology data can support machine-learning prediction of binary AST status and that organism information contributes substantially to predictive performance.

Keywords: data quality assessment; data preprocessing; machine learning; antimicrobial susceptibility testing; binary classification; antimicrobial susceptibility prediction; antimicrobial resistance; CTGAN; Random Forest; inpatient microbiology data.

1. Introduction

Culture and antimicrobial susceptibility testing (AST) data are fundamental for the treatment of infectious diseases, antimicrobial stewardship, and antimicrobial resistance (AMR) surveillance in hospitals. Cumulative antimicrobial susceptibility reports, or antibiograms, are widely used to summarize institutional susceptibility patterns and support empiric antimicrobial selection. The CLSI M39 guideline provides an important framework for the analysis and presentation of cumulative antimicrobial susceptibility test data so that such reports are reliable and clinically useful (Clinical and Laboratory Standards Institute [CLSI], 2022a).

In practice, hospital microbiology data are often affected by multiple data quality problems, including missing values, duplicate records, non-standard organism and antibiotic names, inconsistent specimen labels, and heterogeneous coding across information systems. These problems reduce not only the reliability of cumulative susceptibility reporting but also the readiness of the data for machine-learning analysis. Systems such as WHONET and BacLink were developed to facilitate microbiology data management and standardization, highlighting the importance of data quality assessment and preprocessing before advanced analysis (Stelling et al., 2007; WHONET, 2026).

In recent years, machine learning has increasingly been applied to antimicrobial resistance and susceptibility prediction. Reviews of the literature have shown that tree-based models, particularly Random Forest and gradient-boosted trees, often perform well with real-world healthcare data because they can capture nonlinear relationships and accommodate complex interactions among variables (Anahtar et al., 2021; Ardila et al., 2025; Sakagianni et al., 2023). Nevertheless, many AMR prediction studies still face common limitations, including poor source-data quality, class imbalance, inconsistent preprocessing pipelines, repeated-measurement bias, and limited discussion of how categorical AST labels were derived and standardized (Ardila et al., 2025; Sakagianni et al., 2023).

Accordingly, this study aimed to develop a machine-learning framework for predicting antimicrobial susceptibility under a binary classification setting, susceptible versus non-susceptible, from quality-controlled inpatient hospital microbiology data. The specific objectives were: (1) to assess and improve the quality of inpatient microbiology data before model development; (2) to standardize organism names, antibiotic names, specimen types, and susceptibility results into a machine-learning-ready format; (3) to develop and compare the predictive performance of multiple tree-based machine-learning models for binary antimicrobial susceptibility classification; and (4) to examine the contribution of microbiology variables and patient-context variables to predictive performance.

2. Related Work

Relevant studies can be grouped into four major areas: (1) surveillance systems and cumulative antimicrobial susceptibility reporting, (2) standardization and integration of microbiology data, (3) machine-learning applications for antimicrobial susceptibility prediction, and (4) methodological gaps related to data quality, repeated observations, class imbalance, and real-world implementation.

In the area of surveillance and reporting, CLSI M39 provides an important framework for the analysis and presentation of cumulative antimicrobial susceptibility test data to support clinical decision-making and antimicrobial stewardship (CLSI, 2022a). At the global level, the World Health Organization established the Global Antimicrobial Resistance and Use Surveillance System (GLASS) to standardize the collection, analysis, and sharing of AMR data (World Health Organization, 2022, 2023). In Thailand, the National Antimicrobial Resistance Surveillance Center of Thailand publishes annual antibiogram reports and surveillance summaries, underscoring the direct role of AST data in AMR monitoring and antibiotic stewardship (National Antimicrobial Resistance Surveillance Center of Thailand, 2024, 2025).

In the area of standardization and microbiology data integration, several studies have shown that the major challenge in hospital microbiology datasets is not simply the absence of data but also inconsistency of formats, nonuniform coding, and fragmentation across information systems. WHONET and BacLink were developed to support the capture, conversion, processing, and analysis of microbiology data from legacy systems, particularly for surveillance use (Stelling et al., 2007; WHONET, 2026). In the Thai context, previous work has demonstrated ETL-based processing and visualization pipelines for bacterial profiles and antibiotic susceptibility profiles from hospital data, emphasizing data cleaning, quality assessment, and bias reduction before further analysis (Aramrat & Boonma, 2023).

For machine-learning applications, previous studies have demonstrated that ML can support antimicrobial stewardship, resistance phenotype prediction, and clinical decision support. Personalized antibiogram approaches based on electronic health records have shown the ability to predict antibiotic susceptibility and support narrower-spectrum therapy while maintaining adequate coverage (Corbin et al., 2022). Multitask learning frameworks have also been proposed to predict individualized resistance probabilities across multiple antimicrobial classes (Goto et al., 2026). Collectively, these studies suggest that real-world clinical and microbiology data can be used for clinically meaningful susceptibility prediction (Anahtar et al., 2021; Corbin et al., 2022; Goto et al., 2026).

Despite these advances, major challenges remain. Recent reviews have emphasized that model performance in AMR prediction depends not only on algorithm choice but also on the quality and consistency of the underlying data, the handling of class imbalance, and the rigor of evaluation procedures (Ardila et al., 2025; Sakagianni et al., 2023). These limitations motivate the present study, which directly focuses on the integration of data quality assessment, systematic preprocessing, CTGAN-based class balancing, and binary antimicrobial susceptibility prediction using inpatient microbiology data.

3. Materials and Methods

3.1. Study Design and Data Source

This retrospective study used inpatient microbiology and anonymized patient-context data from Chiang Mai University Hospital, a tertiary hospital in Chiang Mai province, Thailand, collected from 8 September 2011 to 2 January 2020. A total of **1,097,459 inpatient culture records** were initially obtained from the IPD Culture database. Patient-context variables were obtained from the IPD patient database and included sex, age, and ward of admission.

The records underwent sequential preprocessing to clean, standardize, and transform the data. The eligibility criteria for retaining records were defined as follows: (1) records must originate from inpatient cultures; (2) records must successfully merge with the IPD patient-context database using admission identifiers; (3) AST results must be interpretable as S, I, or R; (4) standardized organism, antibiotic, specimen, and susceptibility-result fields must pass the data quality assessment; and (5) records with missing modeling features must be excluded before model development.

3.2. Research Ethics

This study was reviewed and approved by the Human Research Ethics Committee, Faculty of Medicine, Chiang Mai University, Committee 5 (Research ID: 8889; Study Code: FAM-2565-08889). The present study is a secondary analysis of data derived from a previously approved research project. It involved retrospective hospital microbiology and clinical record data, with no direct patient contact or intervention. All data were de-identified prior to analysis. The requirement for individual informed consent was waived by the ethics committee because the study posed minimal risk and did not adversely affect the rights or welfare of the patients.

3.3. Data Quality Assessment

This study assessed data quality in two primary dimensions, completeness and consistency, because both dimensions directly affect data readiness for machine learning. Let the full dataset be denoted by $D=\{x_1,x_2,\ldots,x_N\}$, where Nis the total number of records and $x_{iv}$is the value of variable vin record i.

The completeness of variable v was defined as:

$$\text{Completeness}(v)=\frac{1}{N}\sum_{i=1}^{N} I(x_{iv} \text{ is not missing}), \quad (1)$$

where $I(\cdot)$ is the indicator function.

The consistency of variable v relative to a predefined standard-value set S_v was defined as:

$$\text{Consistency}(v)=\frac{1}{N}\sum_{i=1}^{N} I(x_{iv}\in S_v). \quad (2)$$

These measures were used to identify missing values, non-standard names, unresolved mappings, ambiguous codes, and invalid AST result formats before the modeling stage.

3.4. Data Preprocessing

The preprocessing workflow was designed for inpatient microbiology data originating from routine hospital information systems and containing inconsistent naming conventions. It consisted of five main steps.

3.4.1. Data Standardization

Let $r$denote a raw value, such as an organism name, antibiotic name, specimen type, or susceptibility-result string, and let $T=\{t_1,t_2,\ldots,t_m\}$ denote the corresponding set of standardized terms. The standardization function f(r)was defined as:

$$f(r)=\begin{cases} \text{dictionary match}, & \text{if } r \text{ matches a predefined mapping},\\ \text{regex-normalized match}, & \text{if } r \text{ matches a standard pattern},\\ \arg\max_{t\in T}\text{sim}(r,t), & \text{if } \max_{t\in T}\text{sim}(r,t)\geq\tau,\\ \text{REVIEW}, & \text{otherwise.}\end{cases} \quad (3)$$

where sim(r,t) is the string similarity score between the raw value and a standardized term, and τ is the fuzzy-matching threshold. This procedure was applied to organism names, antibiotic names, specimen types, and susceptibility-result fields. Terms that could not be reliably standardized were assigned REVIEW status and excluded from supervised model development.

3.4.2. Intrinsic-Resistance Checking

Let $o_i$denote the standardized organism for record i, $a_i$the standardized antibiotic, $y_i$the recorded susceptibility category, and $R_{int}$the rule base of organism-antibiotic pairs with known intrinsic resistance. The intrinsic-resistance inconsistency flag was defined as:

$$\text{IRFlag}_i=\begin{cases}1, & \text{if } (o_i,a_i)\in R_{int} \text{ and } y_i=S,\\ 0, & \text{otherwise.}\end{cases} \quad (4)$$

If $\text{IRFlag}_i=1$, the record was flagged for review.

3.4.3. REVIEW Flag Assignment

Let status_idenote the final preprocessing status of record i. A record was assigned REVIEW status if it met at least one of the following conditions: unresolved organism mapping, unresolved antibiotic mapping, unresolved specimen mapping, invalid or ambiguous susceptibility-result coding, or intrinsic-resistance inconsistency.

$$\text{status}_i=\begin{cases}\text{REVIEW}, & \text{if the record cannot be reliably interpreted,}\\ \text{KEEP}, & \text{otherwise.}\end{cases} \quad (5)$$

The supervised-learning dataset before modeling was therefore defined as:

$$D_{sup}=\{i:\text{status}_i=\text{KEEP and } y_i\in\{S,I,R\}\}. \quad (6)$$

3.4.4. Binary Target Transformation

The original AST interpretation was recorded as S, I, or R. For model development, the target variable was transformed into a binary outcome:

$$y\in\{S, \text{non-}S\},$$

where S was interpreted as susceptible, and I or R was classified as non-susceptible. The study used the categorical AST interpretations recorded by the hospital laboratory at the time of testing. Because the study period spanned 2011 to 2020, breakpoint criteria may have varied over time. The study did not retrospectively recalculate all AST categories using a single breakpoint version.

3.4.5. Final Dataset for Modeling

After IPD culture and IPD patient-context linkage, **528,134 records** were available. The merged dataset had no missing sex values, **10,774 missing age values**, and no missing ward values. After excluding records with missing modeling features, **497,542 complete records** were included in the modeling dataset.

The explanatory feature vector for the full model was:

$$X=[\text{organism, antibiotic, specimen, sex, age, ward}]. \quad (7)$$

The complete modeling dataset contained **337,302 susceptible records** and **160,240 non-susceptible records**.

3.5. Handling Class Imbalance

The complete modeling dataset was imbalanced, with susceptible records more frequent than non-susceptible records. CTGAN was used to generate synthetic records for the minority non-susceptible class. Before augmentation, the dataset contained **337,302 susceptible records** and **160,240 non-susceptible records**. CTGAN generated **177,062 synthetic non-susceptible records**, resulting in a balanced analytical dataset of **674,604 records**, with **337,302 records per class**.

3.6. Experimental Design

The CTGAN-balanced analytical dataset was used for model comparison. It was split into training and test sets using an 80:20 split, resulting in **539,683 training records** and **134,921 test records**. The test set contained **67,461 susceptible records** and **67,460 non-susceptible records**.

Three experimental settings were defined to evaluate the contribution of feature groups:

- **Setting 1:** organism + antibiotic + specimen + sex + age + ward
- **Setting 2:** antibiotic + specimen + sex + age + ward
- **Setting 3:** antibiotic + specimen + sex + age

This design assessed the contribution of organism identity and ward context while retaining antibiotic information across all settings.

3.7. Machine-Learning Models

This study compared four tree-based machine-learning models: Decision Tree, Random Forest, XGBoost, and CatBoost. Let x_i denote the feature vector of sample i. The final class prediction of each model can be expressed as:

$$\hat{y}_i=\text{argmax}_{c\in S,\text{non-}S}P(y_i=c|x_i) \quad (8)$$

where ŷ_i is the predicted class for record i, x_i is the feature vector of record i, c is a candidate class, S denotes the susceptible class, non-S denotes the non-susceptible class, and P(y_i = c | x_i) is the estimated probability that record i belongs to class c given its feature vector.

3.7.1. Decision Tree

Decision Tree selects splits by maximizing impurity reduction. In this study, Gini impurity was used:

$$G(t)=1-\sum_{c\in S,non\text{-}S} p(c|t)^2 \quad (9)$$

where G(t) is the Gini impurity at node t, c is a class label, and p(c | t) is the proportion of samples belonging to class c at node t. A smaller Gini impurity indicates a purer node.

3.7.2. Random Forest

Random Forest is an ensemble of multiple decision trees. The final prediction is obtained by majority vote:

$$\hat{y}_i=\text{mode}\{h_1(x_i),h_2(x_i),\ldots,h_B(x_i)\} \quad (10)$$

where ŷ_i is the final Random Forest prediction for record i, h_b(x_i) is the class predicted by tree b for record i, B is the total number of trees in the ensemble, and mode{.} denotes the majority-vote operator.

The class probability can be written as:

$$P(y_i=c|x_i)=(1/B)\sum_{b=1}^{B} I(h_b(x_i)=c) \quad (11)$$

where P(y_i = c | x_i) is the estimated probability that record i belongs to class c, I(.) is an indicator function, I(h_b(x_i) = c) equals 1 if tree b predicts class c and 0 otherwise, and B is the total number of trees.

3.7.3. XGBoost

XGBoost is based on gradient boosting, in which trees are added sequentially to reduce prediction error:

$$\hat{y}_i=\sum_{k=1}^{K} f_k(x_i), f_k\in F \quad (12)$$

where ŷ_i is the prediction score for record i, K is the total number of boosting iterations or trees, f_k is the tree function added at iteration k, x_i is the feature vector of record i, and F is the function space of regression trees.

3.7.4. CatBoost

CatBoost is a gradient-boosting algorithm designed specifically for categorical features and based on ordered boosting:

$$\hat{y}_i^{(m)}=\hat{y}_i^{(m-1)}+\eta f_m(x_i) \quad (13)$$

where ŷ_i^(m) is the prediction score for record i after boosting iteration m, ŷ_i^(m−1) is the prediction score from the previous iteration, η is the learning rate, f_m(x_i) is the base learner added at iteration m, and x_i is the feature vector of record i.

3.7.5. Model Configuration and Experimental Environment

Categorical variables were encoded using Label Encoding for Decision Tree, Random Forest, and XGBoost. CatBoost handled categorical features directly. Hyperparameters were

not optimized using Grid Search or Random Search; instead, manual tuning was applied. Decision Tree used max_depth = 10. Random Forest used n_estimators = 100. XGBoost used n_estimators = 100, learning_rate = 0.1, and max_depth = 6. CatBoost used iterations = 100, learning_rate = 0.1, and depth = 6. A random seed of 42 was used for reproducibility.

3.8. Evaluation Metrics

Model performance was evaluated using accuracy, macro precision, macro recall, and macro F1-score. Particular emphasis was placed on macro F1-score because the original binary class distribution was imbalanced.

Accuracy was defined as:

$$\text{Accuracy}=\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i=y_i). \quad (14)$$

For each class c, precision and recall were defined as:

$$\text{Precision}_c=\frac{TP_c}{TP_c+FP_c},\text{Recall}_c=\frac{TP_c}{TP_c+FN_c}. \quad (15)$$

The F1-score for class cwas defined as:

$$F1_c=\frac{2\times\text{Precision}_c\times\text{Recall}_c}{\text{Precision}_c+\text{Recall}_c}. \quad (16)$$

The macro F1-score was computed as:

$$\text{Macro F1}=\frac{1}{2}\sum_{c\in\{S,\text{non-S}\}} F1_c. \quad (17)$$

4. Results

4.1. Dataset Flow and Class Distribution

A total of **1,097,459 inpatient culture records** were initially obtained from the IPD Culture database. After linking IPD culture records with IPD patient-context data, **528,134 records** were available. The linked dataset had no missing sex values, **10,774 missing age values**, and no missing ward values. After excluding records with missing modeling features, **497,542 complete records** remained for binary classification.

The complete modeling dataset contained **337,302 susceptible records** and **160,240 non-susceptible records**, corresponding to **67.79% susceptible** and **32.21% non-susceptible**.

Table 1. Dataset flow and class distribution.

Step	Records
IPD Culture raw	1,097,459
After linking with IPD patient-context data	528,134
Complete modeling dataset after excluding missing modeling features	497,542
Susceptible (S) records	337,302
Non-susceptible (non-S) records	160,240
Synthetic non-S records generated by CTGAN	177,062

Balanced analytical dataset after CTGAN	674,604
Training set after 80:20 split	539,683
Test set after 80:20 split	134,921

4.2. Effect of CTGAN-Based Class Balancing

Before CTGAN-based augmentation, the complete modeling dataset was imbalanced, with 337,302 susceptible records and 160,240 non-susceptible records. CTGAN generated 177,062 synthetic non-susceptible records so that both classes contained 337,302 records. After augmentation, the analytical dataset contained 674,604 records and was balanced at 50.00% per class.

Table 2. Distribution of binary susceptibility classes before and after CTGAN-based augmentation.

Class	Before CTGAN (n)	Before CTGAN (%)	Synthetic records added	After CTGAN (n)	After CTGAN (%)
Susceptible (S)	337,302	67.79	0	337,302	50.00
Non-susceptible (non-S)	160,240	32.21	177,062	337,302	50.00
Total	497,542	100.00	177,062	674,604	100.00

4.3. Model Comparison Results

The comparison results of the four models across the three experimental settings are presented in Table 3. Random Forest achieved the best performance in all settings.

Table 3. Prediction results for binary antimicrobial susceptibility classification across three experimental settings.

Setting	Features	Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Setting 1	Org, Anti, Specimen, Sex, Age, Ward	Decision Tree	74.41%	0.7448	0.7441	0.7439
		Random Forest	81.66%	0.8167	0.8166	0.8166
		XGBoost	77.37%	0.7738	0.7737	0.7737
		CatBoost	75.83%	0.7583	0.7583	0.7583
Setting 2	Anti, Specimen, Sex, Age, Ward	Decision Tree	66.14%	0.6617	0.6614	0.6612
		Random Forest	72.33%	0.7233	0.7233	0.7233
		XGBoost	69.05%	0.6905	0.6905	0.6904
		CatBoost	67.74%	0.6775	0.6774	0.6773
Setting 3	Anti, Specimen, Sex, Age	Decision Tree	65.64%	0.6565	0.6564	0.6563
		Random Forest	67.04%	0.6704	0.6704	0.6704
		XGBoost	66.80%	0.6681	0.6680	0.6680
		CatBoost	65.93%	0.6595	0.6593	0.6593

In Setting 1, which included organism, antibiotic, specimen, sex, age, and ward, Random Forest achieved the highest accuracy of **81.66%** and macro F1-score of **0.8166**. In Setting 2, after removing organism information, Random Forest accuracy decreased to **72.33%**. In Setting 3, after removing ward information while retaining antibiotic, specimen, sex, and age, Random Forest accuracy further decreased to **67.04%**.

4.4. Effect of Feature Reduction

To evaluate the contribution of feature groups, Random Forest performance was compared across the three settings.

Table 4. Comparison of Random Forest performance under sequential feature reduction.

Setting	Features	Accuracy	Macro Precision	Macro Recall	Macro F1	Accuracy Difference from Setting 1
Setting 1	Org, Anti, Specimen, Sex, Age, Ward	81.66%	0.8167	0.8166	0.8166	—
Setting 2	Anti, Specimen, Sex, Age, Ward	72.33%	0.7233	0.7233	0.7233	-9.33 percentage points
Setting 3	Anti, Specimen, Sex, Age	67.04%	0.6704	0.6704	0.6704	-14.62 percentage points

The largest reduction occurred after organism information was removed, indicating that organism identity contributed substantially to model discrimination. Additional removal of ward information further reduced performance, suggesting that ward context also contributed to prediction.

4.5. Class-Level Performance of Random Forest

Class-wise Random Forest performance is presented in Table 5. The test set used in the notebook output contained 67,461 susceptible records and 67,460 non-susceptible records after CTGAN-balanced analytical splitting.

Table 5. Class-wise classification performance of Random Forest.

Setting	Class	Precision	Recall	F1-score	Support
Setting 1	Susceptible (S)	0.81	0.82	0.82	67,461
	Non-susceptible (non-S)	0.82	0.81	0.82	67,460
Setting 2	Susceptible (S)	0.72	0.73	0.72	67,461
	Non-susceptible (non-S)	0.72	0.72	0.72	67,460
Setting 3	Susceptible (S)	0.67	0.66	0.67	67,461
	Non-susceptible (non-S)	0.67	0.68	0.67	67,460

Random Forest demonstrated balanced performance across both classes in Setting 1, with F1-scores of 0.82 for both susceptible and non-susceptible classes. Class-level performance decreased progressively as organism and ward information were removed.

5. Discussion

This study developed and evaluated a machine-learning framework for predicting antimicrobial susceptibility under a binary classification setting, susceptible versus non-susceptible, using quality-controlled inpatient microbiology and patient-context data. The results show that standardized inpatient microbiology data can support machine-learning-based prediction of binary AST outcomes.

Random Forest achieved the best performance across all three experimental settings. In the full-feature setting, Random Forest achieved 81.66% accuracy and a macro F1-score

of 0.8166. This strong performance reflects the ability of ensemble tree-based models to capture nonlinear relationships between microbiological characteristics and patient-context variables.

Performance declined after feature removal. Removing organism information reduced Random Forest accuracy from 81.66% to 72.33%, representing a decline of 9.33 percentage points. Further removal of ward information reduced accuracy to 67.04%, representing a total decline of 14.62 percentage points from the full-feature setting. These findings are clinically plausible because susceptibility patterns are strongly influenced by organism identity and antimicrobial agent, while ward information may capture additional epidemiological or hospital service-level patterns.

The CTGAN-based balancing step addressed the imbalance between susceptible and non-susceptible records in the modeling dataset. Before augmentation, non-susceptible records represented 32.21% of the complete modeling dataset. CTGAN generated synthetic non-susceptible records to balance the analytical dataset, allowing the models to learn from both classes more evenly.

However, these results should be interpreted as performance on the CTGAN-balanced analytical evaluation derived from the available notebook output. For stricter clinical ML validation, the analysis should be repeated using a workflow in which the real-world dataset is split before CTGAN, CTGAN is fitted only on the training set, and final model evaluation is conducted on an untouched real-world test set. In addition, duplicate reduction should follow a rolling 30-day first-isolate rule per patient, organism, and antibiotic.

5.1. Limitations

This study has several limitations. First, the data originated from a single tertiary-care hospital, which may limit generalizability to other institutions. Second, the retrospective design relied on routine hospital data collected over multiple years, and laboratory practices or breakpoint criteria may have varied during the study period. The study used the categorical S/I/R results recorded in the hospital database and converted them into a binary target of susceptible versus non-susceptible; it did not retrospectively recalculate susceptibility categories using a single breakpoint version across the full period.

Third, CTGAN-based augmentation may not fully capture the true biological and epidemiological structure of minority-class non-susceptible observations. Fourth, the model-performance values reported here are based on the available notebook output from a CTGAN-balanced analytical evaluation. Additional rerunning is recommended using a leakage-safe workflow with CTGAN applied only to the training set and final evaluation on an untouched real-world test set. Fifth, external validation was not performed. Sixth, categorical variables were label-encoded for Decision Tree, Random Forest, and XGBoost for computational convenience; arbitrary ordinal effects cannot be ruled out completely. Finally, unless a temporal split is additionally performed, the evaluation does not fully represent prospective deployment.

Future work should include external validation across hospitals, temporal validation using later-year cohorts, patient-level or admission-level splitting, corrected first-isolate duplicate reduction, additional comparison of encoding strategies and class-imbalance methods, model calibration, uncertainty estimation, and clinically interpretable deployment frameworks for antimicrobial decision support.

6. Conclusion

This study proposed a machine-learning framework for predicting antimicrobial susceptibility under a binary classification setting, susceptible versus non-susceptible, using quality-controlled inpatient microbiology data. The framework integrated data standardization, binary AST target transformation, CTGAN-based class balancing, and tree-based machine-learning models.

From 1,097,459 initial inpatient culture records, 528,134 records were linked with IPD patient-context data, and 497,542 complete records were available for binary modeling. The complete modeling dataset contained 337,302 susceptible records and 160,240 non-susceptible records. CTGAN generated 177,062 synthetic non-susceptible records, resulting in a balanced analytical dataset of 674,604 records.

Random Forest achieved the best performance across all experimental settings. With organism, antibiotic, specimen, sex, age, and ward features included, Random Forest achieved 81.66% accuracy and a macro F1-score of 0.8166. When organism information was removed, accuracy decreased to 72.33%, and when ward information was additionally removed, accuracy decreased further to 67.04%.

Overall, the findings indicate that quality-controlled inpatient microbiology data can support binary antimicrobial susceptibility prediction and that organism information is a major contributor to predictive performance. Further validation using a leakage-safe workflow, untouched real-world test data, temporal cohorts, and external datasets is required before clinical deployment.

References

Anahtar, M. N., Yang, J. H., & Kanjilal, S. (2021). Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research. *Journal of Clinical Microbiology, 59*(7), e01260-20.

Aramrat, C., & Boonma, P. (2023). Development of data processing and visualization for bacterial and antibiotic susceptibility profile. *Data Science and Engineering Record, 4*(1).

Ardila, C. M., González-Arroyave, D., & Tobón, S. (2025). Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review considering antimicrobial susceptibility tests in real-world healthcare settings. *PLOS ONE, 20*(2), e0319460.

Clinical and Laboratory Standards Institute. (2011). *Performance standards for antimicrobial susceptibility testing* (21st informational supplement, CLSI document M100-S21).

Clinical and Laboratory Standards Institute. (2020). *Performance standards for antimicrobial susceptibility testing* (30th ed.; CLSI supplement M100).

Clinical and Laboratory Standards Institute. (2022a). *M39: Analysis and presentation of cumulative antimicrobial susceptibility test data* (5th ed.).

Corbin, C. K., Sung, L., Chattopadhyay, A., Noshad, M., Chang, A., Deresinski, S., Baiocchi, M., & Chen, J. H. (2022). Personalized antibiograms for machine learning-driven antibiotic selection. *Communications Medicine, 2*, Article 38.

Goto, M., Bandyopadhyay, A., Shi, Q., Wang, Y., Perencevich, E. N., Hernandez, D., & Street, W. N. (2026). Personalized antibiogram: A novel multitask machine learning framework for simultaneous prediction of antimicrobial resistance profile with enhanced detection of carbapenem resistance in Enterobacteriaceae. *Clinical Infectious Diseases*, ciag027. https://doi.org/10.1093/cid/ciag027

National Antimicrobial Resistance Surveillance Center of Thailand. (2024). *Annual report of the National Antimicrobial Resistance Surveillance Center of Thailand*.

National Antimicrobial Resistance Surveillance Center of Thailand. (2025). *Antibiograms by health service region*.

Sakagianni, A., Koufopoulou, C., Feretzakis, G., Kalles, D., Verykios, V. S., Myrianthefs, P., &

Fildisis, G. (2023). Using machine learning to predict antimicrobial resistance: A literature review. *Antibiotics, 12*(3), 452.

Stelling, J. M., Kulldorff, M., & O'Brien, T. F. (2007). WHONET and BacLink: Software tools for laboratory-based surveillance of infectious diseases and antimicrobial resistance. *Advances in Disease Surveillance, 2*, 121.

WHONET. (2026). *WHONET microbiology laboratory database software* [Computer software].

World Health Organization. (2022). *Global antimicrobial resistance and use surveillance system (GLASS) report: 2022*.

World Health Organization. (2023). *Global antimicrobial resistance and use surveillance system (GLASS)*.