IET Systems Biology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Investigating the Impact of Antibiotics on Environmental Microbiota Through Machine Learning Models

Yiheng Du[1] | Khandaker Asif Ahmed[2] | Md Rakibul Hasan[3,4] | Md Zakir Hossain[1,4]

[1]Australian National University, Canberra, Australia | [2]CSIRO Australian Centre for Disease Preparedness, Geelong, Australia | [3]BRAC University, Dhaka, Bangladesh | [4]Curtin University, Bentley, Australia

Correspondence: Md Zakir Hossain (zakir.hossain1@curtin.edu.au)

Received: 19 November 2024 | **Revised:** 21 January 2025 | **Accepted:** 20 February 2025

Handling Editor: Hao Wu

Funding: The authors received no specific funding for this work.

Keywords: bioinformatics | biology computing | learning (artificial intelligence)

ABSTRACT

Antibiotic pollution in the environment can significantly impact soil microorganisms, such as altering the soil microbial community or emerging antibiotic-resistant bacteria. We propose three machine learning (ML) methods to investigate antibiotics' impact on microorganisms and predict microbial abundance. We examined the microbial abundances of various environmental soil samples treated with antibiotics. We developed 3 ML models: (Model 1) for predicting the most abundant bacterial classes in a specific treatment group; (Model 2) for predicting antibiotic treatment effects based on bacterial abundances; and (Model 3) for using data from short-term incubations to predict the data of community structure after stabilisation. In Model 1, the Random Forest model achieved the highest average accuracy, with a Coefficient of Variation mean of 0.05 and 0.14 in the training and test set. In Model 2, the accuracy of the random forest and SVM models have the highest accuracy (nearly 0.90). Model 3 demonstrates that the Random Forest can use data from short-term incubations to predict the abundance of bacterial communities after long-term stabilisation. This study highlights the potential of ML models as powerful tools for understanding microbial dynamics in response to antibiotic treatments. The code is publicly available at - https://github.com/DeweyYihengDu/ML_on_Microbiota.

1 | Introduction

Antibiotics have now been extensively utilised in various domains, including routine medical practices, research experiments, animal breeding, and crop production [1]. However, a significant portion of these antibiotics is not fully absorbed by the human or animal bodies, leading to the excess antibiotics eventually finding their way into the soil or water through various means. Among the antibiotics, the highest concentrations found in faeces are of tetracycline drugs, followed by fluoroquinolones and sulfonamides [2]. These antibiotics have a profound impact on the abundance of soil microbes as well as the overall microbial and enzyme activities, subsequently affecting the physical and chemical properties of the soil, such as pH, moisture, and organic matter content, which in turn leads to soil degradation [3–5]. In the process of extracting microbes, metagenomic extraction technology is predominantly employed; however, this technique faces substantial limitations when applied to soil samples. The uneven distribution of microbes within the soil and their adhesion to soil particles make microbial extraction particularly complex [6, 7]. Traditional methods for soil microbial analysis often demand extensive data

organisation and require numerous samples and repetitive experiments to ensure convincing results [8]. These challenges underscore the need for innovative analytical tools to better understand microbial dynamics in soil environments.

Antibiotics are known to exert significant effects on environmental microbiota. Among many antibiotics, the primary ones identified in environmental microbiota include amoxicillin, oxytetracycline dihydrate, sulfadiazine, trimethoprim, tylosin tartarate, and ciprofloxacin. These antibiotics are extensively used in medical and agricultural domains, such as pastoral fields, leading to their inevitable introduction into the environmental microbiota. Amoxicillin primarily functions by inhibiting bacterial cell wall synthesis, affecting the growth of a variety of Gram-positive bacteria and some Gram-negative bacteria [9]. Oxytetracycline dihydrate operates by binding to the 30S ribosomal subunit, interfering with the ability of bacteria to produce essential proteins, thus preventing the incorporation of amino acid residues into elongating peptide chains [10]. Sulfadiazine, on the other hand, inhibits bacterial folic acid synthesis through competitive antagonism of *para*-aminobenzoic acid (PABA), thereby impacting nucleic acid synthesis, and consequently affecting a diverse range of microorganisms, challenging the survival of most microbes [11]. Trimethoprim functions by reversibly inhibiting dihydrofolate reductase, obstructing the reduction of folic acid to tetrahydrofolate, thereby impeding bacterial nucleic acids and protein synthesis [12]. Tylosin tartarate inhibits bacterial protein synthesis by binding to the 50S ribosomal subunit [13]. Lastly, Ciprofloxacin exerts its antimicrobial activity by inhibiting bacterial DNA gyrase, thereby affecting bacterial DNA metabolism [14, 15]. Most studies focus on the effects of single antibiotics; however, real-world environments often exhibit complex antibiotic mixtures. This research addresses both individual and combined impacts of antibiotics on microorganisms.

Machine learning (ML), originally conceptualised by Arthur Samuel in 1959, serves as a methodological approach designed to uncover patterns within large datasets and is a specialised subset of artificial intelligence (AI) [16]. Fundamentally, ML employs algorithms to analyse data, recognise patterns, and subsequently make decisions or predictions based on these insights [17]. Unlike conventional software, which is programed to execute specific tasks, ML models learn from large-scale training data, thereby enabling them to accomplish tasks autonomously [18].

ML has found broad applications across various disciplines, including finance, image recognition [19], healthcare, and recently, in the domain of biology [20–25] and microbiology [26]. In this context, ML is commonly deployed for predictive modelling, feature extraction, time-series analysis, and image classification, among other applications [24, 25, 27]. ML currently has a wide range of applications in evaluating the impact of antibiotics on bacteria and in the screening of antibiotic resistance genes, such as in the screening of resistance genes within the environment of chicken farms [28]. Consequently, machine learning is becoming an increasingly popular data analysis method in the field of microbiology. In microbiology, key research areas include microbial taxonomy [29, 30], the intricate interplay between gut microbiota and their hosts [26, 31, 32], the study of pathogenic microorganisms, and drug discovery [33]. Further, environmental microbial evolution is a burgeoning field [34].

Currently, in the field of microbiology, the main applications of ML include microbial classification and identification [35], interactions between microbes [36], functional genomics [37], and the discovery of drugs and bioactive compounds [38], among other areas. This paper employs ML methods to investigate the impact of antibiotics on microbial community structures. With the advent of the big data era, and particularly the rise of metagenomics and high-throughput sequencing, enormous datasets are becoming increasingly common in microbiological research [39]. Due to the intrinsic complexity and often elusive nature of microbial entities, ML has become an indispensable tool for analysing large datasets and uncovering targeted insights. ML techniques are increasingly being applied to study the effects of antibiotics on bacteria. For example, previous studies have used decision tree methods to predict antibiotic resistance [40], explored *Pseudomonas aeruginosa* through whole-genome approaches [41], and employed XGBoost models to analyse antibiotic activity against multiple bacterial species [42]. These studies demonstrate the growing potential of ML in microbiological research. Building on this foundation, our study leverages ML to investigate the impact of antibiotics on bacterial abundance, offering new insights into these complex interactions.

This study is propelled by two primary objectives. The first objective is to engineer an ML model capable of predicting the level of antibiotic treatment based on bacterial abundance. As previously discussed, the extensive infiltration of antibiotics into various environments significantly influences microbial abundance. By ML algorithms, this study aims to intricately model the relationship between bacterial abundance and the levels of antibiotic treatment. This endeavour is anticipated to provide a pipeline for monitoring and evaluating antibiotic pollution across different environments, thereby contributing to mitigating the detrimental impacts associated with antibiotic pollution. The second objective of this study is to develop an ML model that utilises the abundance data of preceding bacterial generations to predict the abundance of subsequent generations. Understanding bacterial population dynamics across generations is instrumental in delineating the interactions between microbiota and their environment, especially in the context of antibiotic presence. By harnessing the predictive power of ML and historical abundance data, this model aims to furnish a robust analytical tool for forecasting bacterial population dynamics, which is crucial for both microbiological research and environmental management.

These two objectives primarily use ML methods to investigate the impact of antibiotics on microbes in soil environments. Using ML models, the study aims to find the complex relationships between antibiotic residues and microbial abundance, as well as predict microbial population dynamics across generations. The study underscores the potential of ML as a valuable tool in advancing the field of microbiology and environmental science, providing a foundation for future research (for instance, identifying beneficial microbes for crop protection).

2 | Methodology

2.1 | Data Description

In the project, the FASTA data originated from the public project PRJNA576637 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA576637).[1] This dataset includes 627 samples collected as part of The Biodiversity Exploratories, a research initiative funded by the German Science Foundation (DFG Priority Programme 1374). It provides sequencing data on bacterial communities under different conditions, including exposure to six antibiotics, and was designed to investigate the impact of antibiotics on bacterial abundance in soil. Primarily, two types of soil samples (grass soil and forest soil) were collected and microbes were cultured with a combination of six different antibiotics (namely amoxicillin, oxytetracycline dihydrate, sulfadiazine, trimethoprim, tylosin tartrate, and ciprofloxacin). Samples were harvested at 0 days, 3 days, 8 days, and 20 days. Therefore, we categorised the samples based on their cultured time into Incubation-days 0, Incubation-days 3, Incubation-days 8, and Incubation-days 20, used in ML models.

Different combinations of antibiotics have been shown in the supplementary Table S1. Samples from forest soil and grass soil were collected from separate plots, each designated by a unique plot ID. Due to the high level of randomness in bacterial abundance in individual samples, the average value from biological samples across different plot IDs was used as the input data to ensure the accuracy of the model. For the accuracy of the study, the predicted abundance of the bacteria was analysed at the order level to avoid the impact of too many unknown new species on the model.

2.2 | Data Preprocessing

To initially explore the impact of antibiotics on bacterial abundance, we first plotted an overall graph of bacterial abundance in different antibiotic treatment categories (Figure 1). Nonmetric multidimensional scaling (NMDS) analysis is a research method that simplifies objects in a multidimensional space into a lower-dimensional space while preserving the original relationships between the objects [43]. It is frequently applied in ecological analysis. To investigate the correlation between different categories (such as Antibiotics Group, Isolation Source, and Incubation-days) and the overall bacterial abundance, we first performed NMDS analysis on the samples, followed by a correlation analysis using the envfit function from the vegan package in R [44]. The envfit function evaluates the statistical significance of the calculated correlations through a Permutation Test. The Permutation Test is a non-parametric test that generates a null distribution of correlations by repeatedly randomising the association between environmental variable values and the ordination results. This allows for the calculation of the probability (p-value) of observing the current or more extreme correlation under the null hypothesis that the environmental variables are unrelated to the ordination results.

2.3 | Model Development

In this study, we primarily used Python for the overall analysis, especially using scikit-learn for ML [45]. Additionally, during this process, we employed Python libraries such as NumPy and Pandas for data organisation, and the Matplotlib library to visualise the data analysed.

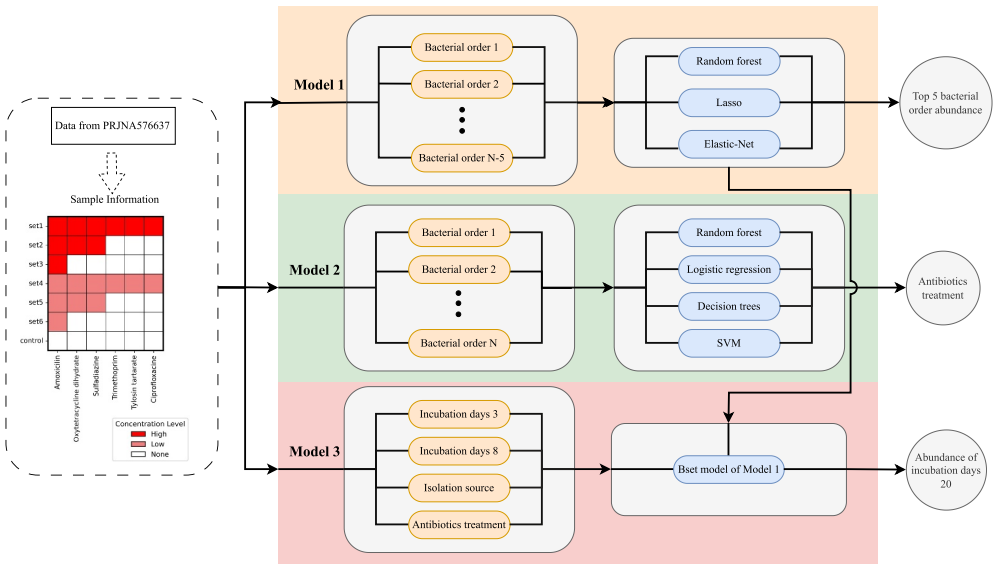

FIGURE 1 | The ML analysis framework used. OTU (operational taxonomic unit) information extracted from the PRJNA457637 FASTA dataset, illustrated through a flowchart depicting the classification and prediction of bacterial order using three distinct models. Within Model 1 and Model 2, the abundance of different bacterial orders is used, with N representing the number of bacterial orders obtained in the sample. Model 3 employs an ML model derived from the best-performing model in Model 1.

2.3.1 | Model 1—Predicting Top Bacteria in Single Sample

In Model 1, we explored the abundances of different bacterial orders to predict the abundances of top bacterial orders. Since this model used the abundance of other bacterial orders as predictive variables, relative abundance cannot be used as it would simplify the model to directly subtract all other relative abundances from 100%. Therefore, when using the abundance of other bacteria, we organised the data by subtracting the abundance of the bacterial orders we need to predict from the original data, and recalculating the abundance of predictive bacterial orders to use as our predictive variables.

During model building, we attempted 3 ML models for modelling, which are the Random Forest model [46], Elastic-Net model [47], and Lasso model [48]. A total of 627 samples were used, split into 80% (501 samples) for the training set and 20% (126 samples) for the testing set. RMSE (Root Mean Square Error) was the square root of the average of the squared errors. It is used to measure the performance of a prediction model, especially in regression problems [49]. RMSE provided a sense of the magnitude of the model's prediction error, with a smaller value indicating higher predictive accuracy of the model [50]. To evaluate the performance of the model, the Coefficient of Variation (CV) was chosen as the test score. The reason was that CV was a standardised form of RMSE, standardised by the mean of actual values, allowing for greater comparison between different bacterial orders, and obtaining the standard error [51].

Through Model 1, we could determine which ML model can perform well in the data. Then, we used the resulted best model, for later models. In later more complex models, we could directly use this ML model to simplify the analysis steps. In this analysis, we selected features with an importance score greater than 0.01 to enhance the interpretability of the model and provide a deeper understanding of the correlations between different bacterial orders.

2.3.2 | Model 2—Predicting Antibiotic Treatment With Bacterial Abundance

In Model 2, we used bacterial order abundance as features, applying ML to predict the antibiotic treatment conditions. Because the objective differed from Model 1, we used different ML models to build Model 2. We attempted to use ML models such as Logistic Regression model, Random Forest model, Decision Tree model, and Support Vector Machine model (SVM) to predict the antibiotic treatment conditions. The sample size was the same as Model 1, with 627 samples divided into 80% (501 samples) for training and 20% (126 samples) for testing. The model initially employed the chi-square test [43], targeting antibiotic treatment groups as the dependent variable, with the source of isolation and varying culture times as feature variables. The diversity of different bacteria was arranged according to their differences. Bacterial abundances were ranked from high to low, starting with the abundance of 10 bacteria as feature variables and incrementally increasing in steps of 10, up to the abundance of 100 bacteria as feature variables, to conduct preliminary machine learning training and test its accuracy. The number of feature variables yielding the highest accuracy was then selected for more detailed training subsequently. During the training process, methods of oversampling and undersampling were utilised to balance the number of datasets differently, and hyperparameter tuning was applied to refine the model.

In the methods section concerning hyperparameter tuning across various ML models, we optimised model performance by adjusting key parameters. For the Random Forest model, we set the number of trees in the range of 100–200 and determined the maximum tree depth to be between 10 and 20. In the SVM model, the penalty strength parameter varied from 1 to 10, and we chose between radial basis function and linear kernels. For the Decision Tree model, we controlled complexity by setting the tree's maximum depth between 5 and 10. Lastly, in the Logistic Regression model, we adjusted the regularisation strength, also in the range of 1–10. These modifications were targeted to strike a balance between the complexity of each model and its ability to generalise, aiming to improve the predictive accuracy of each model on the dataset.

To comprehensively evaluate the performance of our multi-class model, we first construct the confusion matrix for the entire model. Considering our research focus on distinguishing samples that have undergone antibiotic treatment from those that have not, we pay special attention to ROC (Receiver Operating Characteristic) analysis during the model evaluation process. The ROC curve is a graphical tool used for evaluating the performance of classification models. It displayed the model's performance at different classification thresholds by plotting the True Positive Rate (TPR) on the *y*-axis and the False Positive Rate (FPR) on the *x*-axis [52]. AUC (Area Under the Curve) is the area beneath the ROC curve, providing a single value to assess the model's overall classification performance [53]. Specifically, we perform ROC curve analysis on the test set for six different groups of antibiotic treatment conditions in comparison with the control group data. These datasets represent different classification scenarios in comparison to the control group samples that have not undergone antibiotic treatment. The higher the AUC value (maximum is 1), the higher the accuracy of the model. At the same time, we got the confusion matrix to visually observe the accuracy of the models.

2.3.3 | Model 3—Prediction in Different Generations

In Model 3, we used the ML model selected in Model 1, as we were predicting the abundance of bacterial orders based on the abundance of other bacterial orders, and the data structure was similar. Therefore, we can directly employ the ML model that performed the best in Model 1 to build Model 3. This differed from Model 2, as the purpose of Model 2 was to categorise the data, which differed from the aims and data structures of Model 1 and Model 3. The data we used was from the same antibiotic treatment experimental group, using the average bacteria abundance from biological samples with different plot IDs as the data for the model. We did not use the data from Incubation-days 0 because the bacteria in Incubation-days 0 had not yet

experienced antibiotic treatment conditions. As a result, we were left with only 202 sample combinations that met the requirements for model training and validation. In order to avoid the occurrence of extreme values in bacterial abundance in individual samples, we did not use the abundance from a single sample, but instead used the average bacterial abundance of a plot ID. Since we used the average bacterial abundance of plot IDs, the amount of data used in the model was relatively small, so we employed leave-one-out cross-validation to make full use of the data. The method involved using data from Incubation-days 3 and Incubation-days 8 to predict the abundance of the top 5 bacteria on Incubation-days 20. The combinations of features included using Incubation-days 3 data alone to predict the top 5 bacteria's abundance on Incubation-days 20, Incubation-days 8 data alone for the same prediction, and a combination of data from both Incubation-days 3 and 8 to predict the abundance on Incubation-days 20. The best-performing ML model was then selected based on Coefficient of Variation (CV) and Root Mean Square Error (RMSE) metrics.

3 | Results

3.1 | Data Variance

To support the development of ML models, the bacterial abundance data was analysed to investigate the impact of antibiotic treatment and to infer potential changes in bacterial abundance after incubation. This aimed to identify any possible correlations existing between the variations in bacterial abundance. Through the application of Nonmetric Multidimensional Scaling (NMDS) analysis on our samples, we observed significant findings concerning the relationship between bacterial abundance and various classification categories. Specifically, the analysis was conducted under the differentiation of antibiotics group, isolation source, and incubation days. Our results illustrate a high degree of correlation across these categories with respect to the bacterial communities present in the samples.

Firstly, in the bacterial order abundance graph (Figure 2a), we could see that in most of the antibiotic treatment sets, *Chthoniobacterales* had the highest abundance, followed by *iii1_15*. However, in different antibiotic treatment scenarios, a significant variation in bacterial abundance was observed. Even within a single antibiotic treatment set, the bacterial abundance showed significant differences, possibly due to different isolation sources or incubation days. This implied that we cannot simply judge the relationship between bacterial abundances.

In the NMDS analysis, if there is a significant difference in either NMDS1 or NMDS2 within a category, it indicates that the category has a substantial correlation with bacterial abundance. In the box plots of Figure 2b–d, significant differences are observed, indicating that the Antibiotics Group, Isolation Source, and Incubation days all have considerable variability in bacterial abundance. Moreover, when testing for NMDS1 and NMDS2, the *p*-values were significantly less than 0.05, indicating a clear correlation between the categories of Antibiotics Group, Isolation Source, and Incubation-days with the composition of bacterial species (Table 1). The residual variance (Sum Sq) is substantial, indicating that other unmeasured factors may influence bacterial abundance. These categories can be used as features in constructing models in subsequent machine learning analysis.

In conclusion, upon the addition of antibiotic treatment, different concentrations of antibiotic treatment would have very complex impacts on different generations of biological samples. Therefore, in subsequent modelling, it was necessary to use the abundance from multiple incubation-days, isolation source and the conditions of antibiotic treatment as features to predict the information regarding the abundance of the final generation.

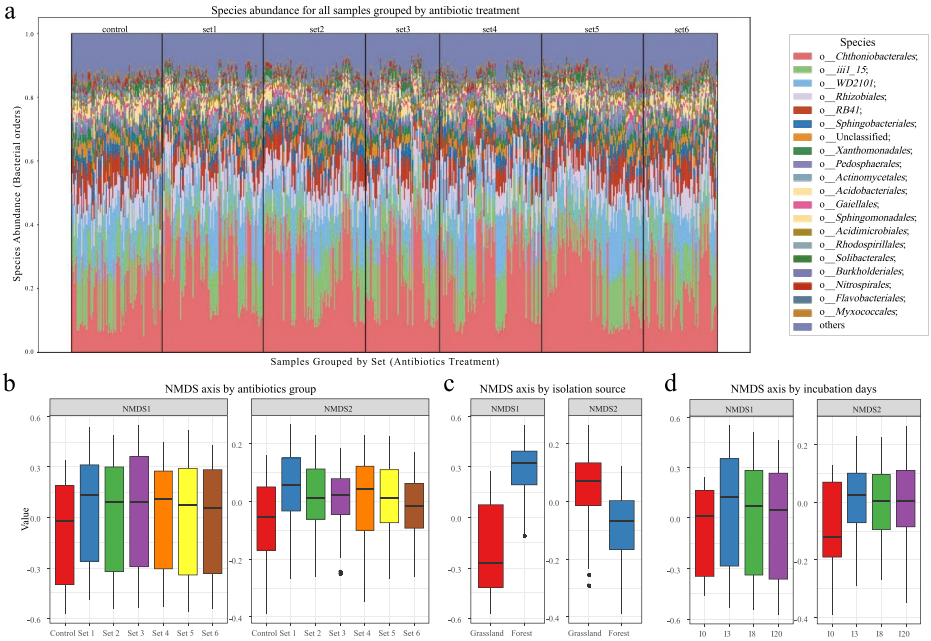

FIGURE 2 | Variance in different antibiotics treatment sets. (a) Species abundance for all samples grouped by antibiotics treatment; (b) Boxplot of NMDS axis value by antibiotics group; (c) Boxplot of NMDS axis value by isolation source; (d) Boxplot of NMDS axie value by incubation-days.

TABLE 1 | Combined analysis of variance tables. NMDS1 and NMDS2 represent the information of two coordinate axes after mapping the data onto a two-dimensional plane through NMDS analysis. The method used for correlation determination is the envfit function for testing correlation. The abbreviations Df, Sum Sq, Mean Sq, and Pr (> F) stand for Degrees of Freedom, Sum of Squares, Mean Square, and *p*-value. A *p*-value less than 0.05 indicates a significant correlation between the classification method and the community structure. Statistical significance indicators: "***" denotes *p-value* < 0.001, "**" denotes *p-value* < 0.01, "*" denotes *p-value* < 0.05, and "ns" indicates a non-significant difference (*p-value* ≥ 0.05), (i.e more * means higher statistical significance).

Response	Df	Sum Sq	Mean Sq	F value	Pr (> F)
NMDS1					
Incubation-days	3	0.742	0.247	5.3020	0.0012960**
Antibiotics	6	1.086	0.181	3.8789	0.0008244***
Isolation source	1	33.837	33.837	725.3946	$< 2.2e-16$***
Residuals	616	28.734	0.047		
NMDS2					
Incubation-days	3	0.2263	0.07543	7.2400	8.849e-05***
Antibiotics	6	0.4352	0.07253	6.9615	3.560e-07***
Isolation source	1	2.8075	2.80750	269.4763	$< 2.2e-16$***
Residuals	616	6.4177	0.01042		

TABLE 2 | The CV MSE and standard error of the top 5 bacteria in Generation 4's abundance under the Random Forest, Lasso, and Elastic-Net ML models (top performances are highligted as bold).

Target column	Model	CV train	CV test	SE train	SE test
o__Chthoniobacterales	**Random Forest**	**0.06**	**0.17**	**7.51E–05**	**1.24E–03**
	Elastic-Net	0.51	0.50	2.47E–03	4.48E–03
	Lasso	0.50	0.53	2.45E–03	4.61E–03
o__iii1_15	**Random Forest**	**0.04**	**0.12**	**1.26E–05**	**3.43E–04**
	Elastic-Net	0.29	0.28	4.77E–04	9.10E–04
	Lasso	0.29	0.29	4.76E–04	9.87E–04
o__WD2101	**Random Forest**	**0.05**	**0.15**	**2.49E–05**	**3.81E–04**
	Elastic-Net	0.55	0.61	1.91E–03	4.03E–03
	Lasso	0.56	0.55	1.96E–03	4.03E–03
o__Rhizobiales	**Random Forest**	**0.05**	**0.14**	**3.63E–05**	**2.26E–04**
	Elastic-Net	0.30	0.31	5.79E–04	8.27E–04
	Lasso	0.30	0.32	5.84E–04	7.56E–04
o__RB41	**Random Forest**	**0.06**	**0.14**	**1.89E–05**	**1.81E–04**
	Elastic-Net	0.31	0.30	4.34E–04	8.21E–04
	Lasso	0.31	0.32	4.14E–04	9.20E–04
Mean	**Random Forest**	**0.05**	**0.14**	**3.35E–05**	**4.74E–04**
	Elastic-Net	0.39	0.40	1.17E–03	2.21E–03
	Lasso	0.39	0.40	1.18E–03	2.26E–03

3.2 | Performance of Predicting Top Bacteria Orders

The aim of this model (Model-1) was to use the abundance of other bacteria in the sample to predict the abundance of the top 5 bacteria in the sample. The five bacterial orders we aim to predict were *Chthoniobacterales*, *iii1_15*, *WD2102*, *Rhizobiales*, and *RB42*. The Coefficient of Variance (CV) was used to evaluate the model, and the Random Forest model was selected as it exhibited the highest accuracy (0.05 in Train Set and 0.14 in Test Set) in predicting bacterial abundance (Table 2). This model then served in the construction of the more complex Model 3 later on. Therefore, we used the random forest model for analysis and construction of Model-3 later. In the random forest model of model 1, we analysed the bacteria, and those with an importance of features greater than 0.01 were obtained in relation to these 5 bacteria, in order to understand the interactions between bacteria in the sample.

In Model 1, we observed that the accuracy of the Lasso model and Elastic-Net model were similar (0.39 in Train Set and 0.40 in Test Set), while the accuracy of the Random Forest model was significantly higher than these two models, as known by comparing the CV (Table 2). While there was a slight decline in accuracy in the test set, the degree of decline was within an acceptable range. In the comparison between the predicted values and actual values in the Random Forest model, we could clearly see that the predicted values and actual values were roughly equal (Figure 3a,b). The R^2 value, also known as the coefficient of determination, measures how well the model's predictions match the observed data. It indicates the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1 signifies a stronger fit between the predicted and actual values. In our analysis, the R^2 values were 0.992 for the training set and 0.901 for the test set, both of which exceed 0.90. These high R^2 values suggest that the model accurately captures the patterns in the data. The better performance of Random Forest compared to Elastic-Net and Lasso is likely due to the wide range of bacterial abundance values in the samples and the potential nonlinear relationships between the target variable and the features. Elastic-Net and Lasso are linear models, which are less effective than Random Forest in capturing such nonlinear dependencies.

In the analysis of feature importance, we used a heatmap (Figure 3c) to display the abundance of the top 5 bacteria and the features with importance of greater than 0.1. We could see that there was a relatively complex relationship among different bacteria orders. Especially when predicting *Chthoniobacterales* abundance, *B07_WMSP1* and *TCC2188* had a larger impact on its features, as shown in Figure 3c. It was worth noting that since we subtracted the abundance of the target bacteria and recalculated the abundance of other bacteria when making predictions, a situation of predicting itself would not occur. Meanwhile, *Sphingobacteriales* had a higher importance for *RB41*, suggesting that there may be a stronger influence relationship between them.

The Random Forest method was currently the most accurate, exhibiting high accuracy in both the training set and test set (0.05 in Train Set and 0.14 in Test Set). The accuracy of the Elastic-Net and Lasso models were quite similar (0.39 in Train Set and 0.40 in Test Set). Therefore, the ML model we used for predicting Incubation-days 20 in Model 3 was the Random Forest model.

3.3 | Performance of Predicting Antibiotic Treatment Groups

In the overall analysis of Model 2, the confusion matrix (Figure 4a) showed that Model-2 possessed the capability to accurately discern the presence of antibiotic contamination within samples. However, the model encounters difficulties in accurately distinguishing the specific degree of antibiotic

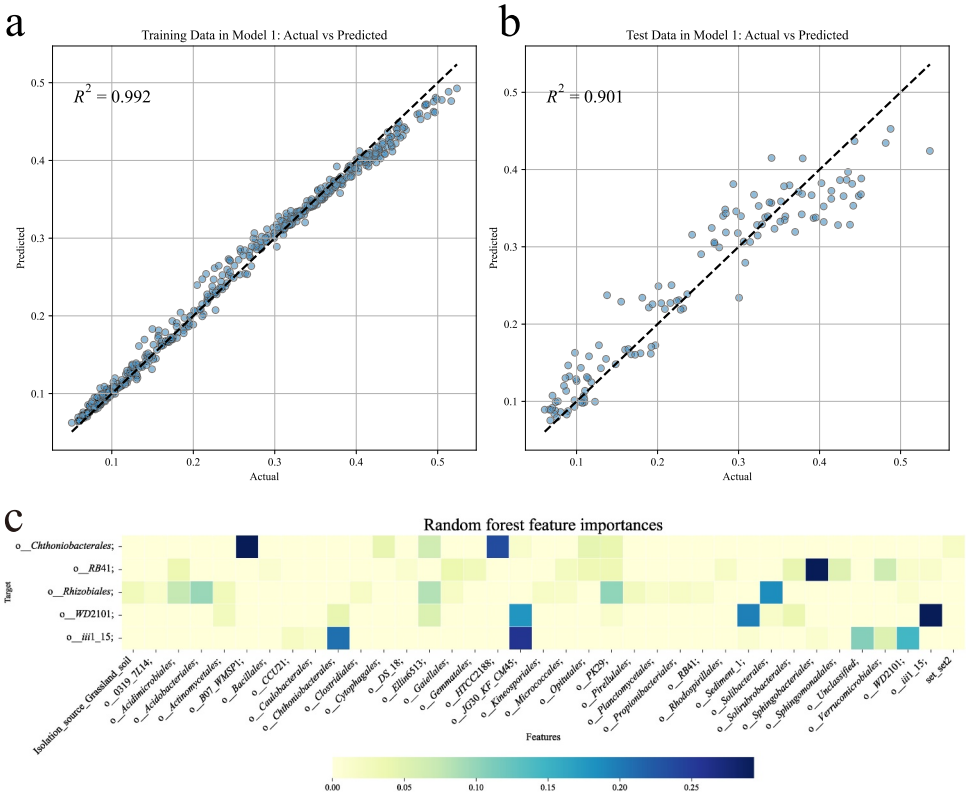

FIGURE 3 | Performance of model 1 and model 2. The R^2 values represent the coefficient of determination, which indicates the proportion of the variance in the observed data explained by the model. An R^2 value closer to 1 signifies a better fit between the predicted and actual values. (a) Performance of Random Forest model in training data: Actual versus Predicted; (b) Performance of Random Forest model in test data: Actual versus Predicted; (c) Heatmap of feature importance for specified bacteria.

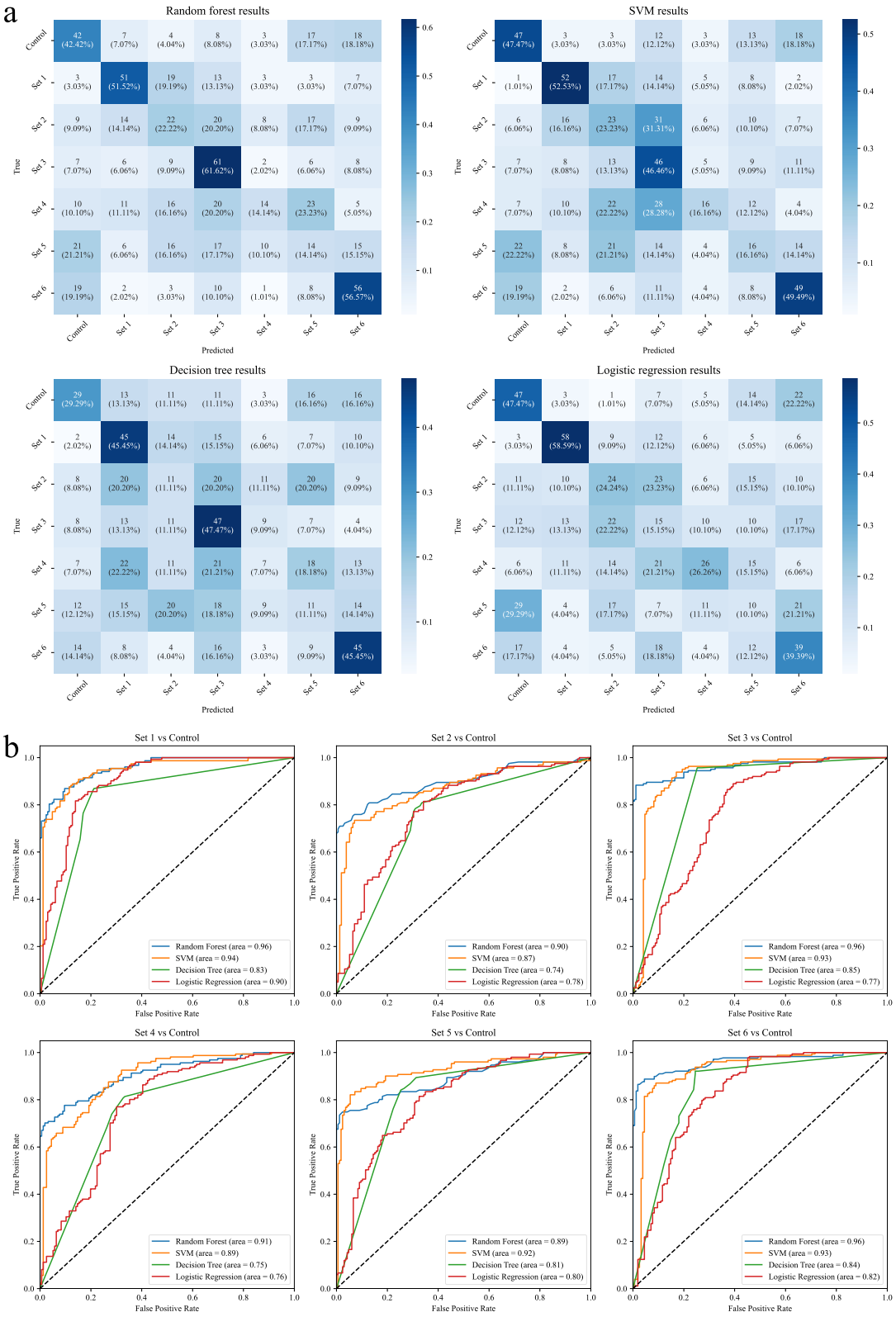

FIGURE 4 | Performance of Model 2. (a) The confusion matrix of Model two includes confusion matrices for machine learning methods such as Random Forest, SVM (Support Vector Machine), Decision Tree, and Logistic Regression; (b) The ROC curves of Model 2, used to distinguish between the control group and various antibiotic treatment groups, evaluate the model's ability to accurately determine whether an environment is contaminated by antibiotics.

contamination and identifying the exact antibiotic agents responsible for the contamination. When distinguishing between the control group and different antibiotic contamination groups (Figure 4b), the random forest and SVM models have the highest accuracy, with AUC values around 0.90. In contrast, decision tree and logistic regression models have lower

accuracy, with AUC values both around 0.80. Model-2 could effectively identify the presence of antibiotic contamination in samples, but it performed poorly in predicting the specific types and levels of antibiotic contamination, leading to potential misidentification of antibiotic contamination combinations.

Furthermore, during this analysis, we incorporated the isolation source as an input variable, namely (forest soil and grass soil). The inclusion of the isolation source significantly enhanced the accuracy of the model. It suggested that if Model 2 aimed to accurately identify antibiotic treatment conditions, it needed re-training in different environmental contexts to ensure the model's accuracy. This showed the importance of tailoring the model according to the unique characteristics and conditions of the environment in which it was applied.

3.4 | Performance of Generations Prediction ML Model

The objective of this model (Model-3) was to use the mean abundance of bacterial Incubation-days 3 and 8 within the same plot to predict the abundance of Incubation-days 20 under this antibiotic treatment. Guided by the results of Model 1, we employed the Random Forest model for Model 3. Subsequently, hyperparameter tuning was utilised to select the appropriate parameters for the model. Since we employed different plot IDs, Leave-One-Out Cross-Validation (LOOCV) was used. The Mean Squared Error (MSE) was 0.0003, and the R^2 was 0.9539 (Figure 5a). The predictions were essentially in line with the actual values, signifying good accuracy of our model. So, the use of mean abundance across different plots was a good choice aimed at enhancing the model's accuracy. However, in scenarios where there were not enough diverse plots or adequate sample quantities, employing individual sample counts as input factors could be an alternative approach.

In order to evaluate the impact of using different input data on the model, we compared the usage of data from Incubation-days 3 and Incubation-days 8 to predict the abundance of bacterial orders in Incubation-days 20. As shown in Figure 5b, only using the data from Incubation-days 3 to predict Incubation-days 20 got a smaller CV than using data from Incubation-days 8, but the model using a combination of data from both Incubation-days 3 and Incubation-days 8 exhibited the smallest CV (0.14), indicating that the amalgamation of data from these generations can improvement the model's accuracy. When enough data was available, using the data from the generation closest to the target prediction generation enhanced our model's accuracy further. And using the combined data could also improve the model's accuracy.

In conclusion, we could use Model 3 to predict data across different incubation days. To enhance the model's accuracy, the mean value of similar samples could be used to reduce data variance. Additionally, the use of combined data from different incubation days also served to improve the model's precision.

4 | Discussion

In this study, we primarily focused on three machine learning models. Model 1 predicts the abundance of the top 5 bacteria using the abundance data of bacteria other than the top f5 in abundance. Model 2 uses the abundance of various bacteria to predict whether soil samples are contaminated with antibiotics. Model 3 employs bacterial abundance information from incubation days 3 and 8 to predict the abundance of bacteria in samples on incubation days 20.

4.1 | Variance Between Different Generations With Antibiotics Treatments

A significant aspect of the research is identifying classification levels with clear correlations to start the construction of ML models. During the NMDS analysis, it became apparent that antibiotics, isolation, and incubation days significantly influence

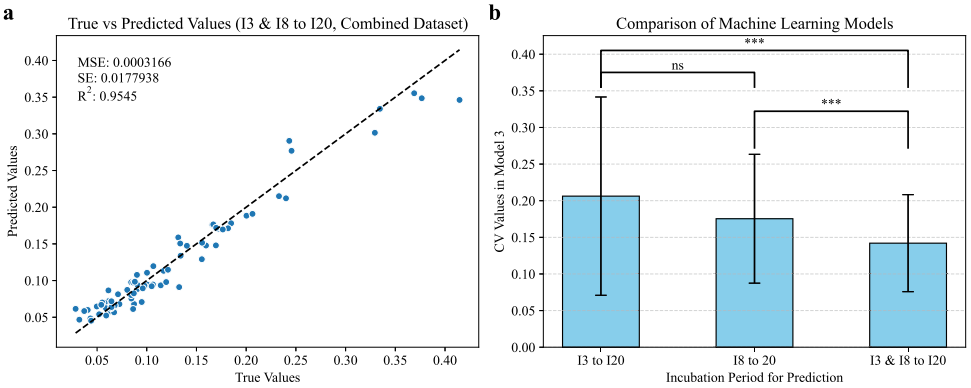

FIGURE 5 | Performance of Model 3 and its evaluation. (a) True versus Predicted values (Incubation-days 3 & Incubation-days 8 to Incubation-days 20 in combined data set). The R^2 values represent the coefficient of determination, which indicates the proportion of the variance in the observed data explained by the model. An R^2 value closer to 1 signifies a better fit between the predicted and actual values; (b) Coefficient of Variance (CV) between actual and predicted relative abundances for different ML of model 3 with error bars representing the standard error of the mean. Here, we compared 3 methods with different information about the Incubation-days 3 and Incubation-days 8. Statistical significance indicators: "***" denotes *p-value* < 0.001, "**" denotes *p-value* < 0.01, "*" denotes *p-value* < 0.05, and "ns" indicates a non-significant difference (*p-value* ≥ 0.05).

the bacterial community structure. Within the antibiotics treatment group, a substantial difference was observed between the control group and the samples treated with antibiotics, whereas the differences among samples treated with various antibiotics were less significant. This aligns with the findings from Model 2, which showed lower accuracy in identifying the specific concentrations of antibiotic treatments. The source of isolation has a significant impact on the bacterial community structure, with considerable differences observed in NMDS, indicating that the bacterial community structures differ markedly across different isolation sources. In the analysis of incubation days, varying incubation periods also introduced significant differences in the bacterial community structure, suggesting that the community structure of bacteria changes after antibiotic treatment is applied. However, NMDS analysis only indicates correlation and does not imply causation; therefore, at this point, we cannot conclude that a succession in the bacterial community has occurred.

In the analysis of functions, antibiotics are formulated to either exterminate or inhibit bacterial growth by targeting essential biological processes within bacterial cells. However, the emergence of antibiotic resistance challenges the efficacy of antibiotics. Over time, the appearance of antibiotic resistant or degrading bacteria underscores the adaptability and resilience of bacterial communities [4]. They evolve mechanisms to modify antibiotic targets, decrease antibiotic uptake, increase antibiotic input, or produce enzymes that neutralise the antibiotics, showing a complex level of evolutionary adaptation.

In summation, a profound understanding of the complex interactions between antibiotic treatments, bacterial abundance, and antibiotic resistance is quintessential for accurately predicting microbial community dynamics and advancing antibiotic resistance research.

4.2 | Random Forest Algorithm Showed Best Accuracy for Model 1

In Model 1, the Random Forest model demonstrated a strong accuracy in predicting the abundance of the top five bacteria in the samples by recalculating their abundance using the abundance data of other bacteria. While the standalone results from Model 1 may not hold substantial importance in practical research, the model serves as a valuable tool for delving into the potential collaborative relationships existing among different bacteria within the samples rather than deriving conclusions solely through simple statistical tests.

Moreover, during the analytical phase, many response variables or features were identified as uncultured bacteria. Despite their uncultured status, advancements in genomics assembly techniques have facilitated the recognition of these bacteria as new orders, distinct from other known bacterial orders. Hence, these uncultured bacteria were incorporated as response variables or features for the construction of ML models. Expanding on the concept of bacterial abundance, it embodies a crucial facet of microbial ecology. The relative abundance of various bacterial species in a given environment can significantly influence the overall microbial community structure, functionality, and its response to external pressure, such as antibiotic treatments. The dynamics of bacterial abundance can explain potential symbiotic or antagonistic relationships among different bacterial taxa, shaping the community's collective behaviour and its impact on the environment.

The interactions among different bacterial orders revealed through ML models like Random Forest show the complex network of relationships underpinning microbial community dynamics. These interactions between different orders may include mutualistic, antagonistic, and symbiotic relationships. These complex interactions can lead to the formation of relatively stable communities in environmental samples. Furthermore, the emergence of uncultured bacteria as significant features in ML models accentuates the vast unexplored diversity within microbial communities.

4.3 | Bacteria Abundance to Predict the Antibiotics Treatment

In the process of using bacterial abundance to predict the outcome of antibiotic treatment in Model 2, we observed that ML models like Random Forest model or Support Vector Machine (SVM) could get satisfactory results. The significance of this model lies in our ability to directly use the abundance of soil microorganisms to explore the extent of antibiotic contamination in the soil. We noted that our ML model performed well in samples treated with high concentrations of antibiotics, while there was a decline in accuracy in samples treated with lower antibiotic concentrations. From this, we inferred that perhaps in the face of light antibiotic contamination, soil microorganisms exhibit a self-restorative phenomenon. For example, certain soil microorganisms possess the capability to degrade or neutralise antibiotics, thus reducing their concentration and toxicity over time [54]. Bacterial genera such as *Pseudomonas* and *Bacillus* have been recognised for their ability to degrade antibiotics [55]. Moreover, some soil microorganisms can exhibit antibiotic resistance, allowing them to thrive in contaminated environments and potentially outcompete sensitive species over time. This could lead to a gradual restoration of microbial diversity, albeit with a different community composition reflecting a new equilibrium adapted to the presence of antibiotics.

In situations where a soil sample initially identified as heavily contaminated is later identified as normal in subsequent analyses using Model 2, it could suggest a potential microbial-mediated remediation of antibiotic contaminants. This hypothesis aligns with known microbial capabilities to biodegrade or neutralise antibiotic substances, underscoring the dynamic and adaptive nature of soil microbiomes in response to anthropogenic stressors like antibiotic pollution. It is noteworthy that in the modelling process, when we did not include the isolation source as a feature, the model's accuracy was quite low. However, upon including it, the model's accuracy significantly improved. This suggests that different environmental samples might necessitate the reconstruction of models. Therefore, we utilised hyperparameter tuning in our parameter settings to allow simple model training across different environmental samples, albeit at the cost of increased computational time.

In our study, we employed machine learning models as classifiers, and the majority of our models achieved ROC-AUC values greater than 0.90. This performance surpasses that of many existing models reported in the literature. For example, Yasir et al. [40] utilised decision tree methods to predict antibiotic resistance, achieving an ROC-AUC value of 0.82, which is notably lower than 0.90. Similarly, Stanton et al. [41] conducted studies on *Pseudomonas aeruginosa* and reported ROC-AUC values below 0.70 for their machine learning models. Moran et al. [42] used XGBoost models to predict antibiotic activity against *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*, but their ROC-AUC values were approximately 0.70. In comparison, our machine learning models demonstrate superior performance relative to these studies.

These intricate interactions between soil microorganisms, antibiotics, and the surrounding environment underscore the multifaceted impact of antibiotic pollution on microbial ecosystems. In a broader perspective, the findings from this modelling endeavour contribute to the burgeoning field of environmental microbiology.

4.4 | Successful Prediction in Different Generations

In Model 3, we successfully predicted the abundance of bacteria in incubation 20 days using the abundance of bacteria in incubation 3 and 8 days. This implies that we can use the abundance data from short-term culturing of environmental samples to obtain the bacterial abundance data that might require long-term culturing. However, there is a downside. Although we can predict the data for incubation-20-day, we still need a small amount of data from incubation-20-day to train our model during the model-building process. Therefore, our model is more suited for scenarios within long-term projects or long-term industrial research to predict the outcomes of microbial contamination. Model 3 provides a valuable approach to environmental microbiology research, allowing for the inference of microbial evolutionary processes across different generations. Expanding further on this discussion, the ability to predict bacterial abundance in later generations based on earlier generations can be a significant asset in long-term microbial ecology studies and industrial applications. It allows for a more efficient allocation of resources by minimising the need for prolonged culturing and monitoring. Especially in industrial settings, where microbial contamination can lead to substantial financial and operational challenges, having a predictive model like Model 3 can enable early intervention strategies to manage microbial populations effectively.

4.5 | Limitations and Future Works

Using our Model 2, it is possible to determine whether an environment is contaminated with antibiotics solely based on bacterial abundance. By employing Model 3, we can predict information about bacterial abundance after the community structure has stabilised, thereby avoiding extensive cultivation time and significantly enhancing the work efficiency of environmental scientists or biologists.

Currently, in the fields of microbiology and antibiotics, the main research focus using machine learning is on exploring the synthesis methods of antibiotics. There is scant use of machine learning to predict the changes in bacterial abundance across different generations under various antibiotic treatments. Therefore, this study did not compare its model with other published models.

In studying the impact of antibiotics on soil microorganisms, through Model 2, we can use the abundance of bacterial order in biological samples to predict and distinguish the antibiotic content in soil environments. However, the accuracy of using Model 2 to predict the degree of antibiotic contamination still needs to be improved. In future research, we should select more features to enhance the model's accuracy. When we use Model 3 to predict the abundance across different generations, we still need a small amount of predicted bacterial abundance data to train our model. Moreover, if we do not include the isolation source as feature during model training, the model's accuracy will be significantly low. This implies that in different experiments, we cannot use whole trained data for prediction; we need to use the project-specific data to retrain for tailored data suitable for that particular project. Therefore, in future work, we should make the entire analysis process more straightforward, making it easier for individuals without a strong computing background to use it easily.

5 | Conclusion

The aim of our study was to annotate the microbial 16s metagenomic sequence data obtained from soil environments, and utilise ML techniques to explore the impact of varying levels of antibiotic treatment on bacterial abundance. Additionally, we aimed to predict the bacterial abundance of subsequent generations using data from previous generations. Our findings showed that upon antibiotic treatment, species diversity in environmental samples sharply declined initially but gradually recovered with the passage of culturing time and increase in generations. Through Model 3, we could accurately predict the abundance of the next generation, thus reducing the experimental period. Meanwhile, Model 2 enabled us to use bacterial abundance for predicting antibiotic contamination, presenting a potential approach for antibiotic detection. Our research, by melding ML with metagenomic data, provides valuable tools for the complex interactions within microbial communities and their responses to antibiotic treatments, laying a groundwork for further explorations in microbial ecology and antibiotic pollution monitoring.

Author Contributions

Yiheng Du: data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing – original draft. **Khandaker Asif Ahmed:** conceptualisation, data curation, methodology, project administration, supervision, validation, writing – review and editing.

Md Rakibul Hasan: methodology, software, supervision, validation, visualisation, writing – review and editing. **Md Zakir Hossain:** conceptualisation, funding acquisition, methodology, project administration, resources, software, supervision, writing – review and editing.

Acknowledgements

Open access publishing facilitated by Curtin University, as part of the Wiley - Curtin University agreement via the Council of Australian University Librarians.

Ethics Statement

Since the data used in this article is from a public dataset, ethical statements are not applicable.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors have nothing to report.

Data Availability Statement

The data that support the findings of this study are openly available under NCBI BioProject ID PRJNA576637 and are available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA576637.

Endnotes

[1] The data comes from the NCBI public database: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA576637.

References

1. M. Jampani, J. Mateo-Sagasta, A. Chandrasekar, et al., "Fate and Transport Modelling for Evaluating Antibiotic Resistance in Aquatic Environments: Current Knowledge and Research Priorities," *Journal of Hazardous Materials* 461 (2024): 132527, https://doi.org/10.1016/j.jhazmat.2023.132527.

2. D. Massé, N. Saady, and Y. Gilbert, "Potential of Biological Processes to Eliminate Antibiotics in Livestock Manure: An Overview," *Animals* 4, no. 2 (2014): 146–163, https://doi.org/10.3390/ani4020146.

3. M. Apreja, A. Sharma, S. Balda, K. Kataria, N. Capalash, and P. Sharma, "Antibiotic Residues in Environment: Antimicrobial Resistance Development, Ecological Risks, and Bioremediation," *Environmental Science & Pollution Research* 29, no. 3 (2021): 3355–3371, https://doi.org/10.1007/s11356-021-17374-w.

4. T. Ma, X. Pan, L. X. Chen, et al., "Effects of Different Concentrations and Application Frequencies of Oxytetracycline on Soil Enzyme Activities and Microbial Community Diversity," *European Journal of Soil Biology* 76 (2016): 53–60, https://doi.org/10.1016/j.ejsobi.2016.07.004.

5. Y. Xu, W. Yu, Q. Ma, J. Wang, H. Zhou, and C. Jiang, "The Combined Effect of Sulfadiazine and Copper on Soil Microbial Activity and Community Structure," *Ecotoxicology and Environmental Safety* 134 (2016): 43–52, https://doi.org/10.1016/j.ecoenv.2016.06.041.

6. L. Ranjard and A. Richaume, "Quantitative and Qualitative Microscale Distribution of Bacteria in Soil," *Research in Microbiology* 152, no. 8 (2001): 707–716, https://doi.org/10.1016/S0923-2508(01)01251-7.

7. L. Grundmann, "Spatial Scales of Soil Bacterial Diversity – the Size of a Clone," *FEMS Microbiology Ecology* 48, no. 2 (2004): 119–127, https://doi.org/10.1016/j.femsec.2004.01.010.

8. N. Lombard, E. Prestat, J. D. Elsas, and P. Simonet, "Soil-specific Limitations for Access and Analysis of Soil Microbial Communities by Metagenomics," *FEMS Microbiology Ecology* 78 (2011): 31–49, https://doi.org/10.1111/j.1574-6941.2011.01140.x.

9. B. J. Akhavan and P. Vijhani, *Amoxicillin* (StatPearls Publishing, 2023), https://www.ncbi.nlm.nih.gov/books/NBK482250/.

10. W. Zhang, B. D. Ames, S.-C. Tsai, and Y. Tang, "Engineered Biosynthesis of a Novel Amidated Polyketide, Using the Malonamyl-specific Initiation Module From the Oxytetracycline Polyketide Synthase," *Applied and Environmental Microbiology* 72, no. 4 (2006): 2573–2580, https://doi.org/10.1128/aem.72.4.2573-2580.2006.

11. J. E. Maddison, A. D. J. Watson, and J. Elliott, "Chapter 8 - Antibacterial Drugs," in *Small Animal Clinical Pharmacology*, eds. J. E. Maddison, S. W. Page, and D. B. Church. 2nd ed. (W.B. Saunders, 2008), 148–185, https://doi.org/10.1016/B978-070202858-8.50010-5.

12. R. N. Brogden, A. A. Carmine, R. C. Heel, T. M. Speight, and G. S. Avery, "Trimethoprim," *Drugs* 23, no. 6 (1982): 405–430, https://doi.org/10.2165/00003495-198223060-00001.

13. J. K. Aronson, *Meyler's Side Effects of Drugs*. 16th ed.) (Elsevier, 2016), 233, https://doi.org/10.1016/B978-0-444-53717-1.01609-7.

14. K. Drlica and X. Zhao, "DNA Gyrase, Topoisomerase IV, and the 4-quinolones," *Microbiology and Molecular Biology Reviews* 61, no. 3 (1997): 377–392, https://doi.org/10.1128/mmbr.61.3.377-392.1997.

15. Y. Pommier, E. Leo, H. Zhang, and C. Marchand, "DNA Topoisomerases and Their Poisoning by Anticancer and Antibacterial Drugs," *Chemistry & Biology* 17, no. 5 (2010): 421–433, https://doi.org/10.1016/j.chembiol.2010.04.012.

16. A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development* 3 (1959): 210–229, https://doi.org/10.1147/rd.33.0210.

17. M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science* 349, no. 6245 (2015): 255–260, https://doi.org/10.1126/science.aaa8415.

18. P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM* 55, no. 10 (2012): 78–87, https://doi.org/10.1145/2347736.2347755.

19. G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging," *Journal of Medical Imaging and Radiation Sciences* 50, no. 4 (2019): 477–487, https://doi.org/10.1016/j.jmir.2019.09.005.

20. Y. Sun, W. Dai, and W. He, "Identification of Key Immune-Related Genes and Immune Infiltration in Diabetic Nephropathy Based on Machine Learning Algorithms," *IET Systems Biology* 17, no. 3 (2023): 95–106, https://doi.org/10.1049/syb2.12061.

21. L. Wang, W. Zhang, Q. Gao, and C. Xiong, "Prediction of Hot Spots in Protein Interfaces Using Extreme Learning Machines With the Information of Spatial Neighbour Residues," *IET Systems Biology* 8, no. 4 (2014): 184–190, https://doi.org/10.1049/iet-syb.2013.0049.

22. X. Lin, Z. Quan, Z. Wang, T. Ma, and X. Zeng, "Kgnn: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction," *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (2020): 2739–2745, https://doi.org/10.24963/ijcai.2020/380.

23. X. Lin, L. Dai, Y. Zhou, et al., "Comprehensive Evaluation of Deep and Graph Learning on Drug–Drug Interactions Prediction," *Briefings in Bioinformatics* 24, no. 4 (2023): bbad235, https://doi.org/10.1093/bib/bbad235.

24. T. Wang, J. Geng, X. Zeng, R. Han, Y. E. Huh, and J. Peng, "Exploring Causal Effects of Sarcopenia on Risk and Progression of Parkinson Disease by Mendelian Randomization," *npj Parkinson s Disease* 10, no. 1 (2024): 164, https://doi.org/10.1038/s41531-024-00782-3.

25. T. Wang, H. Shu, J. Hu, et al., "Accurately Deciphering Spatial Domains for Spatially Resolved Transcriptomics With Stcluster," *Briefings in Bioinformatics* 25, no. 4 (2024): bbae329, https://doi.org/10.1093/bib/bbae329.

26. L. Maier, C. V. Goemans, J. Wirbel, et al., "Unravelling the Collateral Damage of Antibiotics on Gut Bacteria," *Nature* 599, no. 7883 (2021): 1–5, https://doi.org/10.1038/s41586-021-03986-2.

27. C. Cao, C. Wang, Q. Dai, Q. Zou, and T. Wang, "Crbpsa: Circa-rbp Interaction Sites Identification Using Sequence Structural Attention Model," *BMC Biology* 22, no. 1 (2024): 260, https://doi.org/10.1186/s12915-024-02055-0.

28. M. Baker, X. Zhang, A. Maciel-Guerra, et al., "Machine Learning and Metagenomics Reveal Shared Antimicrobial Resistance Profiles Across Multiple Chicken Farms and Abattoirs in china," *Nature Food* 4, no. 8 (2023): 707–720, https://doi.org/10.1038/s43016-023-00814-w.

29. D. Bulgarelli, K. Schlaeppi, S. Spaepen, E. V. L. Themaat, and P. Schulze-Lefert, "Structure and Functions of the Bacterial Microbiota of Plants," *Annual Review of Plant Biology* 64, no. 1 (2013): 807–838, https://doi.org/10.1146/annurev-arplant-050312-120106.

30. Z. Yu, Y. Jia, Y.-H. Du, Z.-J. Du, and D. S. Mu, "Description and Genome Analysis of Actobacterium Tashihtau gen. nov., sp. nov., a Noval Denitrifying and Carbon-Fixing Bacterium Isolated from Marine Sediment," *Research Square* (2022), https://www.researchsquare.com/article/rs-1633326/v1.

31. M. L. Jones, J. G. Ganopolsky, C. J. Martoni, A. Labbé, and S. Prakash, "Emerging Science of the Human Microbiome," *Gut Microbes* 5, no. 4 (2014): 446–457, https://doi.org/10.4161/gmic.29810.

32. W. E. Ruff, T. M. Greiling, and M. A. Kriegel, "Host–microbiota Interactions in Immune-Mediated Diseases," *Nature Reviews Microbiology* 18, no. 9 (2020): 521–538, https://doi.org/10.1038/s41579-020-0367-2.

33. W. H. Moos, C. A. Pinkert, M. H. Irwin, et al., "Epigenetic Treatment of Persistent Viral Infections," *Drug Development Research* 78, no. 1 (2016): 24–36, https://doi.org/10.1002/ddr.21366.

34. J. Zhang, Y.-X. Liu, X. Guo, et al., "High-throughput Cultivation and Identification of Bacteria From the Plant Root Microbiota," *Nature Protocols* 16, no. 2 (2021): 988–1012, https://doi.org/10.1038/s41596-020-00444-7.

35. R. Knight, A. Vrbanac, B. C. Taylor, et al., "Best Practices for Analysing Microbiomes," *Nature Reviews Microbiology* 16, no. 7 (2018): 410–422, https://doi.org/10.1038/s41579-018-0029-9.

36. K. Faust and J. Raes, "Microbial Interactions: From Networks to Models," *Nature Reviews Microbiology* 10, no. 8 (2012): 538–550, https://doi.org/10.1038/nrmicro2832.

37. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun Metagenomics, From Sampling to Analysis," *Nature Biotechnology* 35, no. 9 (2017): 833–844, https://doi.org/10.1038/nbt.3935.

38. J. M. Stokes, K. Yang, K. Swanson, et al., "A Deep Learning Approach to Antibiotic Discovery," *Cell* 180, no. 4 (2020): 688–70213, https://doi.org/10.1016/j.cell.2020.01.021.

39. J. Galloway-Peña and B. Hanson, "Tools for Analysis of the Microbiome," *Digestive Diseases and Sciences* 65, no. 3 (2020): 674–685, https://doi.org/10.1007/s10620-020-06091-y.

40. M. Yasir, A. M. Karim, S. K. Malik, A. A. Bajaffer, and E. I. Azhar, "Application of Decision-Tree-Based Machine Learning Algorithms for Prediction of Antimicrobial Resistance," *Antibiotics* 11 (2022): 1593, https://doi.org/10.3390/antibiotics11111593.

41. R. A. Stanton, D. Campbell, G. McAllister, et al., "Whole-genome Sequencing Reveals Diversity of Carbapenem-Resistant *pseudomonas aeruginosa* Collected through Cdc's Emerging Infections Program, United States, 2016–2018," *Antimicrobial Agents and Chemotherapy* 66, no. 9 (2022): e0049622, https://doi.org/10.1128/aac.00496-22.

42. E. Moran, E. Robinson, C. Green, M. Keeling, and B. Collyer, "Towards Personalized Guidelines: Using Machine-Learning Algorithms to Guide Antimicrobial Selection," *Journal of Antimicrobial Chemotherapy* 75, no. 9 (2020): 2677–2680, https://doi.org/10.1093/jac/dkaa222.

43. R. Miller and D. Siegmund, "Maximally Selected Chi Square Statistics," *Biometrics* 38, no. 4 (1982): 1011, https://doi.org/10.2307/2529881.

44. P. Dixon, "VEGAN, a Package of R Functions for Community Ecology," *Journal of Vegetable Science* 14, no. 6 (2003): 927–930, https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.

45. G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn," *GetMobile* 19, no. 1 (2015): 29–33, https://doi.org/10.1145/2786984.2786995.

46. L. Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32, https://doi.org/10.1023/a:1010933404324.

47. H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society B* 67, no. 2 (2005): 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

48. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society B* 58, no. 1 (1996): 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

49. R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting* 22, no. 4 (2006): 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

50. C. J. Willmott and K. Matsuura, "On the Use of Dimensioned Measures of Error to Evaluate the Performance of Spatial Interpolators," *International Journal of Geographical Information Science* 20, no. 1 (2006): 89–102, https://doi.org/10.1080/13658810500286976.

51. G. F. Reed, F. Lynn, and B. D. Meade, "Use of Coefficient of Variation in Assessing Variability of Quantitative Assays," *Clinical and Vaccine Immunology* 9, no. 6 (2002): 1235–1239, https://doi.org/10.1128/cdli.9.6.1235-1239.2002.

52. K. H. Zou, W. J. Hall, and D. E. Shapiro, "Smooth Non-parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests," *Statistics in Medicine* 16, no. 19 (1997): 2143–2156, https://doi.org/10.1002/(sici)1097-0258(19971015)16:19<2143::aid-sim655¡3.0.co;2-3.

53. J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology* 143, no. 1 (1982): 29–36, https://doi.org/10.1148/radiology.143.1.7063747.

54. D. Borthakur, M. Rani, K. Das, M. P. Shah, B. K. Sharma, and A. Kumar, "Bioremediation: An Alternative Approach for Detoxification of Polymers From the Contaminated Environment," *Letters in Applied Microbiology* 75, no. 4 (2021): 744–758, https://doi.org/10.1111/lam.13616.

55. C. Tran, I. E. Cock, X. Chen, and Y. Feng, "Antimicrobial *Bacillus*: Metabolites and Their Mode of Action," *Antibiotics* 11, no. 1 (2022): 88, https://doi.org/10.3390/antibiotics11010088.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.