Predicting Antibiotic Resistance Through Whole Genome Sequences and Machine Learning

Joel C. Hoefs | 20757218

Supervised by Md Redowan Mahmud

NPSC2001: Research, Leadership and Entrepreneurship in Science 1

October 12, 2023

Authors Note

As this report loosely resembles APA 7th Ed. formatting, I'll note that this isn't in any form a professional research paper and I've done this simply for good practice.

- Relevant code can be found over at this GitHub
- Primary model iteration was done on google collab here

Table of Contents

Abstract 3
Background 4
What is Antibiotic Resistance and Why Does it Matter? 4
Machine Learning in Bioinformatics 4
Patient Metadata and Electronic Health Records 6
Data acquisition 6
Performance and Suitability 7
Whole Genome Sequences 7
Proof of Concept 8
Variant Calling 13
Making the pipeline 15
Tracing Feature Importance's as SNP determinants 17
Multilabel Data 17
Oversampling 18
The Curse of dimensionality and feature selection 19
Random Forests, Classifier Chains and Stacking 19
Parametric and Performance centric approaches 23
Frequency-based Chaos Game Representation (FCGR) 23
Approaches 24
Insights & Future work 27
What's lacking? 28
Other approaches 28
Reflection of this work 28
References 29

Abstract

Antibiotic Resistance (ABR) is a global burden to the future of modern medicine and has garnered multidisciplinary efforts to regulate, understand and surveil the phenomenon. Federal stewardship plans have encouraged bioinformaticians to research statistical alternatives to ABR testing and optimize workflows in genomic ABR studies. Following this endeavor, we evaluate machine learning (ML) methods against the most common unary track infection, Escherichia Coli (E. Coli) and 4 commonly proscribed antibiotics; Ciprofloxacin, Cefotaxime, Ceftazidime and Gentamicin. Multiple ML techniques compete to predict and reveal ABR determinants from a Single Nucleotide Polymorphism (SNP) matrix of 809 E. Coli genome samples as an alternative to genome wide association studies and general-use empirical ABR testing. A stacking ensemble consisting of a logistic regression, random forest and support vector machine trained on SNP lists with additional random forest feature reduction outperformed all other parametric and non-parametric models with an average f1-score of 0.867. A single random forest classifier chain with 200 trees each ranked extremely closely suggesting it alone should suffice for SNP data and was further used to extract feature importance's via the average Gini impurity decrease method to rank significant SNP locations. Many resulting SNP loci lined up with ABR encoding gene regions proving it a valid empirical alterative to genome wide association studies (GWAS). However Alternative approaches have been shown to have greater predictive potential and due to the diminishing returns of these complex procedures may not be effective enough to have significant impact in clinical research. We also find areas for improvement for ML backed assistance software at the point of care.

Background

What is Antibiotic Resistance and Why Does it Matter?

Antimicrobial resistance (AMR) is a fast-spreading health crisis that threatens the effectiveness of modern antibiotic treatment, associated with millions of deaths annually ("Antibiotic resistance," 2020; "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," 2022). While AMR is concerned with all pathogenic organisms, here we focused on resistance in bacterial infections alone, referred to as antibiotic resistance.

We see in nature that some bacterial cells have the capacity to defend themselves from antibodies. Antibiotic degrading enzymes, efflux pumps and the ability to modify its outer binding target all rendering drug(s) useless. Rarely, resistance is initiated via spontaneous mutation, but more frequently this behavior is either inherited during replication or horizontal gene transfer, a process accelerated through the misuse of antibiotics (Sakagianni et al., 2023). When an antibiotic prescription kills most bacterial cells, those with resistance to the antibiotic remain and freely multiply without competition (Thänert et al., 2019). Thus, using the correct antibiotic and dosage is pinacol for controlling and eradicating resistant infections, and over time many strategies have been adopted in the form of antimicrobial stewardship programs to assist with global ramification. A recent innovation being clinical and digital decision support systems (DDSS) used to assist a professionals' prescription at the point of care via a web-based software solution (G. Feretzakis et al., 2021).

We see that levels of resistance vary between countries by a significant margin where antibiotics are either scarce or multidrug prescriptions are overused (Ricciardi et al., 2016). This and other factors suggest that a patient's contextual information, such as hereditary and medical history, may serve as a weak but un-invasive empirical prediction for DDSS solutions, an approach that later explored in depth.

Machine Learning in Bioinformatics

Over the last decade, a great surplus of unused data from a variety of disciplines has gathered the attention of the previously shunned backpropagation algorithm for regression and classification proposed in 1986 (Rumelhart et al., 1986), now university credited as the apex of modern machine learning. A generation's worth of research in optimizing accuracy and training speed has resulted in a full catalogue of architectures.

One such surplus in question being health and disease related data, which if taken advantage of promises life-saving insights into medical research and the future of biotechnology. Where medical records and experimental metadata have long since been prospects in the eyes of machine learning researchers, Next Generation Sequencing (NGS) and modern genomics poses a larger opportunity with more backing to maximize workflows in pathology and biomedical research. Today's machine learning models are mostly parametric, they adopt the backpropagation algorithm and possess up to billions of weights and biases, all slowly updated iteratively during training to minimize the model's degree of error. This has Allowed people to confront huge dataset surpluses and reveal highly complex dependencies human beings lack the eye or time to discover.

Traditional non-parametric models omit the murky ocean of parameters in deep neural networks in favor of a simpler, rigorous mathematics driven algorithms that exposes its inner workings and decision making for trust, comprehension, and easy traceable feature importance (Zhou et al., 2022). This traceability being an incredibly powerful tool in a research environment, allowing data scientists to query their models for markers and clusters that attribute the phenotypes it was initially trained to predict.

Patient Metadata and Electronic Health Records

Usually, antibiotic susceptibility testing involves the disk diffusion method, a plate covered by a growth medium for bacterial colonization is dotted with antibiotic wafers and incubated for a day. The penetration of the antibiotic against the infection is visible as a diffusion circle, the minimum inhibitory concentration (MIC) is determined by the width of each diffusion circle which then infers whether each antibiotic is resistant / intermediate / susceptible to the bacteria. However due to the rarity of dangerous resistance among non-life-threatening infections (Rivers, 2016), ABR testing is not routinely done; certain antibiotics diffuse faster than others incurring measurement error, facilities and resources for ABR testing are not always available (Ren et al., 2021) and some bacteria take significant time to fully colonize (Medicine, 2013). Broth microdilution can be used as an alternative where suspended bacteria is added to an array of nutrient-rich antibiotic solutions with descending dilution where the least dilution with visible diffusion represents the MIC (mcg/ml). Broth Macro-dilution speeds up this labor-intensive process by using a pre-inserted matrix of antibiotic wells, but still suffers from the same burdens as the disk diffusion method. As the greatest source of antibiotic misuse is diagnostic uncertainty and lack of knowledge, antimicrobial stewardship programs target alternatives to ABR testing at the point of care.

Empirical DDSSs aim to provide fast predictions based of electronic health record (EHR) data to alert physicians about possibly inappropriate choices whilst honoring their time and final decision (Georgios Feretzakis et al., 2021; Lewin-Epstein et al., 2020; Rezel-Potts & Gulliford, 2022). Whilst relying on medical history, prescription and metadata is fast and un-invasive, it will never come close to the accuracy of proper ABR testing, thus interoperability, speed and availability of such models is prioritized.

Data acquisition

In an attempt to produce a variant DDSS solution, we found that EHR datasets are incredibly sparse and locked behind professional research grants and board clearances unsuitable for

undergraduates. In fact, any remnants of patient information are rigorously protected by research hospitals, national repositories, large scale databases and respective researchers for ethical reasons. We contacted universities and researchers that seemed willing to comply or participate but had to eventually move on due to time constraints.

Performance and Suitability

Despite the many research endeavors into DDSS solutions for ABR, implementations are rarely used in clinical practice. This being primarily the result of the lack of urgency, impact and interoperable conclusions (Georgios Feretzakis et al., 2021). The models themselves also continue to be weak in predictive potential.

EHR data is known to be inconsistent and broken, making generalizable machine learning interfaces near impossible. However, new endeavors in one-shot learning and natural language processing strive to interpret inconsistent, unstructured EHR data but are still in their infancy (Li et al., 2022; Liu et al., 2022).

Whole Genome Sequences

Genomics and NGS have given rise to many computational approaches in pathogen research. One such method is known as a genome-wide association study (GWAS), which involves the study of genome variation at the nucleotide level, single nucleotide polymorphisms (SNPs) and indels, as diagnostic markers for disease phenotype, virulence, regulation, and their related genes. As this method is purely statistical, it fails to capture the strength of phenotype-SNP relatedness and is overly sensitive (Shi et al., 2019). We loosely adapt this practice to the world of Machine learning, then evaluate its appropriateness and propose techniques for empirical SNP marker detection (Benkwitz-Bedford et al., 2021; Her & Wu, 2018; Ren et al., 2021; Shi et al., 2019). E. coli is favored for our pathogen of choice due to its genome availability, single chromosomal nature and the heavy amount of research and reference material about its genotype in academia. E. coli is a gram-negative bacteria, primarily found in the unary

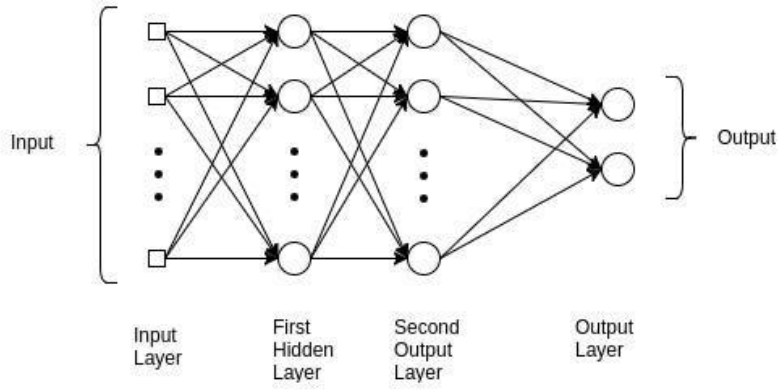
tract known for strong symptom recurrence (Thänert et al., 2019). Strains are frequently multi-drug resistant with a high volume of β-lactamase enzymes, and through rapid horizontal gene transfer have gain the capacity to resist virtually every antimicrobial (Poirel et al., 2018). The relatedness of disease virulence and E. coli serotype variation is emphasized due to a relatively low SNP count (Rahman et al., 2022). As a result, E. coli has become the gold standard for evaluating research in ABR.

Proof of Concept

Prior to beginning the genomic solution, a proof of concept was made to classify ABR gene sequences for demonstration purposes. The prototype was simple and had little resemblance to the GWAS but proves how easily genotype-phenotype relationships are revealed through supervised machine learning. The dataset contains 100,000 genes of undisclosed origin, and a binary resistance value to an undisclosed drug with equal class balance. In this approach, 3 models of interest compete for accuracy each being supported solutions for sequence data (Srinivasu et al., 2022). The first was a 3-layer feed forward neural network, also referred to as a Multilayer Perception (MLP).

Figure 1

Depiction of MLP with 2 hidden layers



Note. From Analytices Vidhya By F. Perixoto, 2020, “A Simple overview of Multilayer Perceptron (MLP)” https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/

A feed forward neural network contains layers of parameters, weights and biases that repeatedly transform an input vector into a single output representing the confidence that the gene is resistant. This is done by multiplying neurons against weights done as matrix-like combinations shown in *figure 1* before adding a bias to the result at each layer. As this alone represents a naive linear transformation, an activation function is required both to prevent linear regression and consistently normalize output. When a loss function captures its degree of error after an attempted prediction, the backpropagation algorithm is used to update theses parameters to lower the highly parameterized loss function by taking a step backwards from their respective gradient vector.

With the given notation the *four horsemen of backpropagation* (Nielsen, 2015) are as follows:

$$\delta^l = \nabla_a C \odot \sigma'(z^l) \quad (1)$$

$$\delta^l = \left((w^{l+1})^T \delta^{l+1}\right) \odot \sigma'(z^l) \quad (2)$$

$$\frac{\boldsymbol{\delta C}}{\boldsymbol{\delta b_j^l}} = \delta_j^l \quad (3)$$

$$\frac{\boldsymbol{\delta C}}{\boldsymbol{\delta w_{jk}^l}} = a_k^{l-1}\, \delta_j^l \quad (4)$$

- $C :=$ cost / loss function
- $\sigma :=$ activation function
- $a_k^l :=$ input value of the k^{th} neuron in the l^{th} layer,
- $z^l :=$ represents the l^{th} layer's output of a neuron prior to activation: $z^l \equiv w^l a^l + b^l$
- $w^l := l^{th}$ layer's weight matrix
- $w_{jk}^l :=$ weight connecting the j^{th} neuron in the $(l+1)^{th}$ layer to the k^{th} neuron in the l^{th} layer.

These equations done in sequence returns the partial derivative of a weight or bias with respect to the cost function, giving way to the final gradient step applied to every parameter in the network after a training batch:

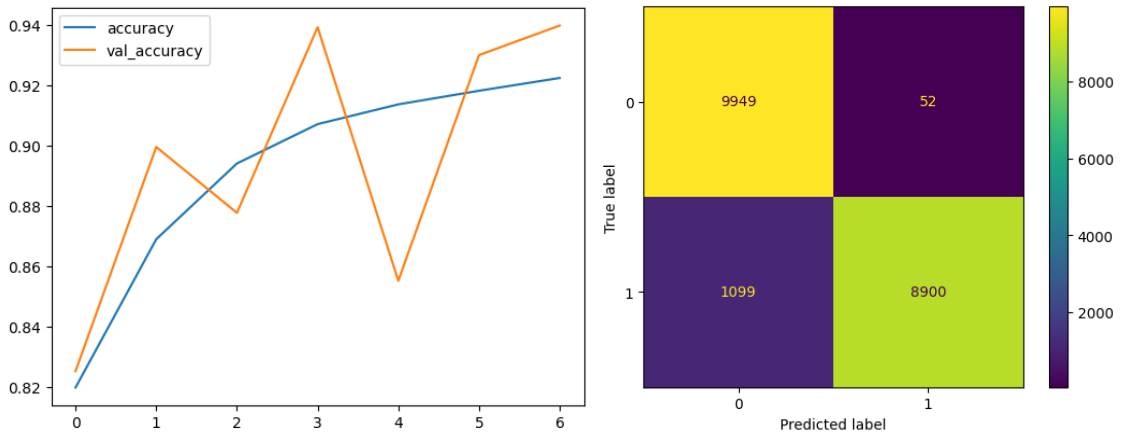
$$w_{jk}^{l} \rightarrow w_{jk}^{l} - \alpha \frac{\boldsymbol{\delta C}}{\boldsymbol{\delta w_{jk}^{l}}} \qquad b_{j}^{l} \rightarrow b_{j}^{l} - \alpha \frac{\boldsymbol{\delta C}}{\boldsymbol{\delta b_{j}^{l}}}$$

Where the constant α is wisely chosen by the programmer to perfect this process; a variable known as a *hyperparameter* of which many others exist. For conciseness, understanding of this algorithm is not required for this report, nor will I dive deeper into the other niece behaviors and models mathematically and instead refer only by name.

The MLP in question possessed 2 hidden layers of 20 & 10 neurons each, ReLU activation, L2 regularization, binary cross entropy loss and ADAM optimization. 20% of the dataset was used for testing, and another 20% of the remaining training data was used for per epoch validation data. Neural networks of this mass suffer from a preventable problem called *overfitting,* where the model fits too closely to the training data and in losing its ability to generalize, suffers when its tested against the new unseen validation data. Overfitting is detectable, as the per-epoch training accuracy will begin to exceed the validation accuracy significantly during training. Thusly TensorFlow provides *early stopping* capability to stop the training loop to restore weights at the optimal time. Overfitting also being the motivation for L2 (Lasso) regularization, adding a portion of the weights to the loss function to punishing high weights for making extreme assumptions about the training set. The MLP achieved a final accuracy of 94%.

Figure 2

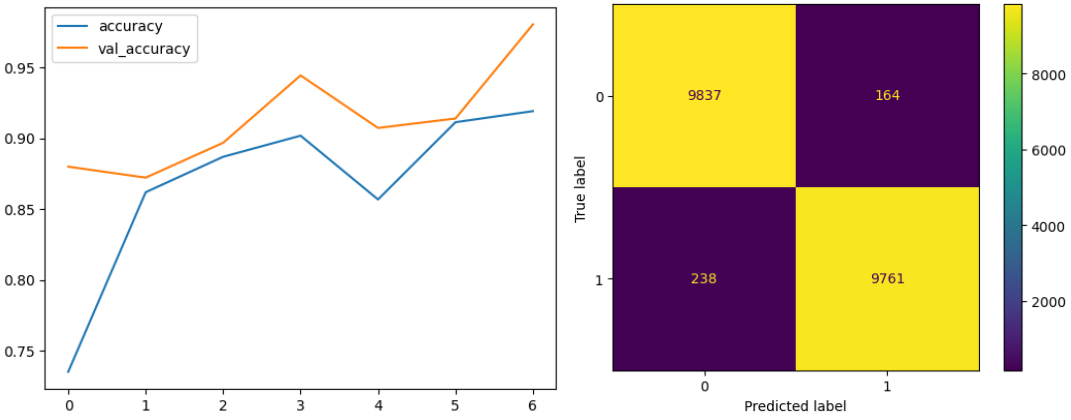
Training curve and confusion matrix (0 representing susceptibility and 1 resistance) matrix of the MLP, showing a clear generosity to false positives.



Next was a bidirectional gated recurrent unit neural network (GRU), an optimization of the recurrent neural network (RNN) model proposed in 2014. RNN's function by taking variable length sequence data and retaining a portion of its output for its own subsequent input. Long sequence dependencies are conserved within this hidden state and is used to make the final prediction. A GRU enhances the RNN cell allowing it to control (gate) how much of the hidden state to be covered or passed for a feature, a solution to the vanishing gradient problem that burdens traditional RNN's. A bidirectional GRU makes one pass across the sequence, and then another back to the start to further solidify each nucleotides equal importance in the sequence regardless of order. The Bi-GRU achieved an accuracy of 98%.

Figure 3

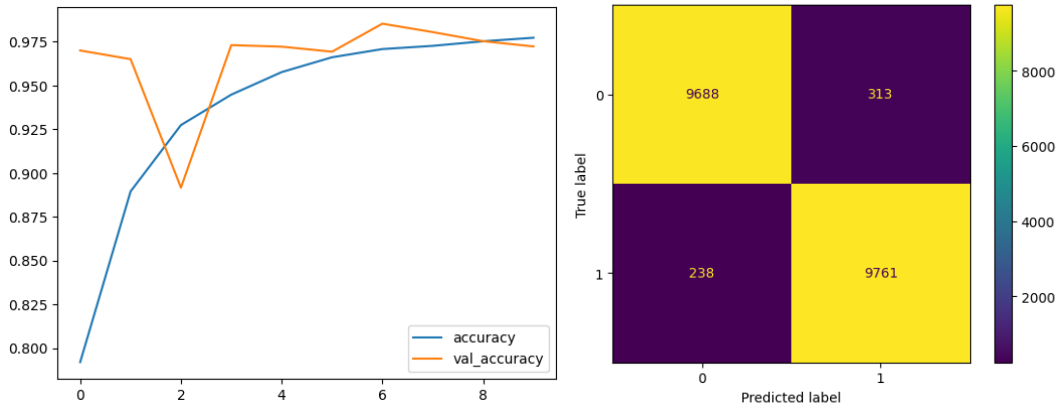
Training curve and confusion matrix of the Bi-GRU



Finally, these were matched against a standard 1-dimensional Convolutional Neural Network (CNN). An architecture usually found in 2 dimensions for image-based tasks with support for genetic sequences and even whole genome datasets for supervised predictions (Ren et al., 2021). Despite little relevancy or focus on bioinformatics, CNNs are still routinely used in genomics due to its ability to tackle and simplify data with an extremely high number of related features. A 1-dimensional CNN holds its parameters in vector-like filter maps, which are used via mathematical convolution to produce a set of convolutional layers each extracting a separate pattern inscribed by the filter. This is usually followed by a pooling layer reducing the dimension of the resultant maps until the remaining maps can be flattened into an output vector sent to a single sigmoid binary classification neuron. In place of regularization, *dropout layers* were used to combat overfitting. These layers exclude a percentage (40%) of the preceding neurons from participating in a training batch. This prevents the network from relying on only a subset of its filters – meaning it's learnt to be sensitive to very specific features of the data resulting in overfitting.

Figure 3

Training curve and confusion matrix of the CNN



The dataset was originally gathered from Kaggle.com but was later removed by the author before this report was written. This dataset was not at all representative of real genomic data but may indicate the superiority of these models for nucleotide data moving forward.

Variant Calling

Adopting the SNP-phenotype strategy above, we first ask ourselves why this method is meaningful. Traditional Antibiotic Resistance (ABR) predictions rely wholly on known ABR genes. Studying variation at the nucleotide level allows for a generic approach for locating related ABR encoding genes and intergenic regulatory elements in the core genome (Segerman, 2012; Shi et al., 2019), and assisting the analysis of understudied Whole Genome Sequence (WGS) data when new resistances emerge.

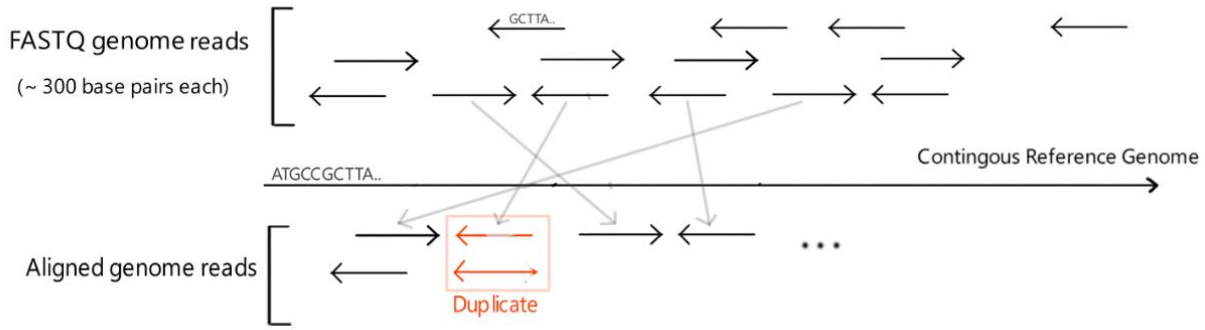
The Dataset was openly supplied from The European nucleotide Archive, where the original genome dumps where supplied for open use (Moradigaravand et al., 2018). The dataset contained accession numbers for 1500 pair-end reads of E. coli K-12 and 4 of its most proscribed antibiotics with a simple susceptibility list per sample.

To extract SNP's a large preprocessing pipeline must first be curated. Modern Pair-end Genome reads come in the FASTQ file format, containing unsorted random pair-end reads of roughly 300 base-

pairs each and associated base quality scores. For our purpose these reads are useless in their current form and must firstly be arranged into order by lining it up with a serotype reference genome using an alignment similarity algorithm. This process is called alignment, a hefty computation where the algorithm must work with FASTQ files often a gigabyte each to reveal a contiguous genome within the reads.

Figure 4

Abstract representation of alignment process with a pair-end FASTQ genome



Once the binary alignment files are sorted and indexed, the variant calling process can take place. Here, the alignment file is referred again to the reference genome where SNPs and Indels (variations of more than 1 nucleotide) are extracted based on read support and quality. The resulting variant call files are filtered removing low quality calls before SNP and indel data is written to the dataset. Thankfully as E. coli only has a single chromosome the complication of alleles and genotype can safely be ignored.

Making the pipeline

Automating genome downloading, alignment, variant calling, and dataset curation was done using python and bash in a Linux environment. The popular Burrows Wheel alignment algorithm (Li & Durbin, 2009) was used for alignment, the samtools package was used for sorting, indexing and diagnostics, and the bcftools package for variant calling and cleaning (Multiple genome alignment was not an option due the multiple terabyte size of all the genomes together with the resultant files). Preprocessing was done on a per-genome basis deleting all data besides the variants before another was downloaded.

A cloud-based solution was adopted distributing partitions of the data over 3 *c5.large* AWS compute instances allowing an initially 10-day task to be complete autonomously in just 2. AWS's virtual CPU token system being particularly suitable for the burstable IO / alignment behavior of the pipeline. However, it would have been more appropriate to take advantage of AWS's genomics tailored solutions for faster transfer and better storage potential.

Finally, a SNP-matrix (Benkwitz-Bedford et al., 2021; Ren et al., 2022) was formed for ease of use taking samples as rows and SNP loci as columns. The pipeline was maintained on GitHub and can be found here, however, after strong examination of the resultant dataset issues were eventually found rendering the data ineffective. After trying multiple pipelines and diagnostics I was never able to find the source of the issue and solemnly resorted to a sub-optimal dataset openly supplied by Ren et al. (2022) with only 600 genomes of the same species and antibiotics late in the year. This dataset adopted the same SNP-phenotype strategy above and the same SNP-matrix form. FASTQC was used to evaluate the quality of the genomes prior to alignment, showing no immediate issues.

Figure 4

Per-base sequence quality of a random E. coli genome from the dataset

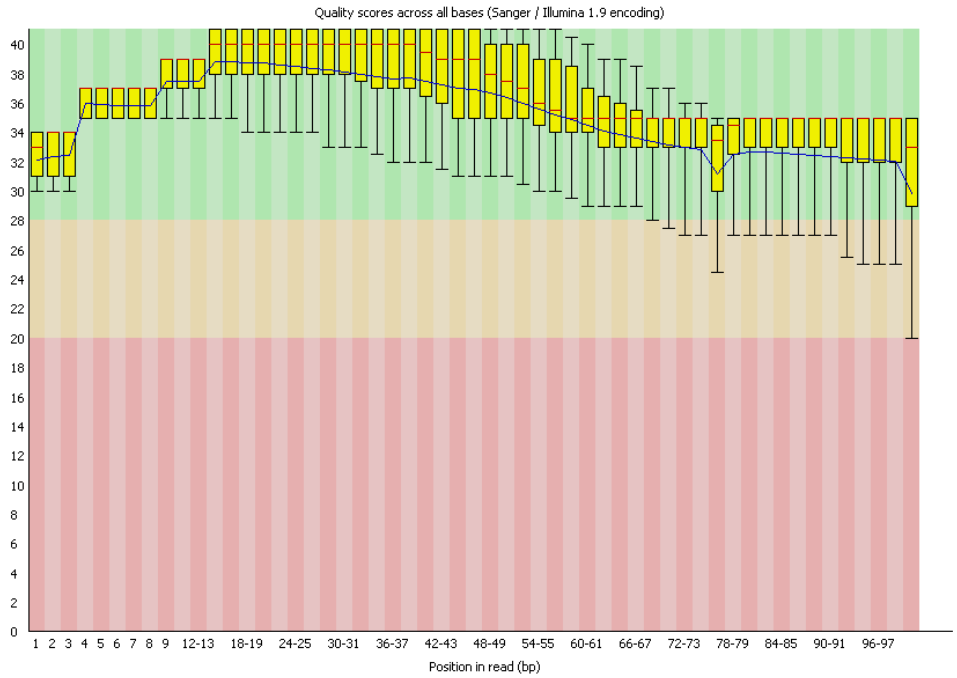
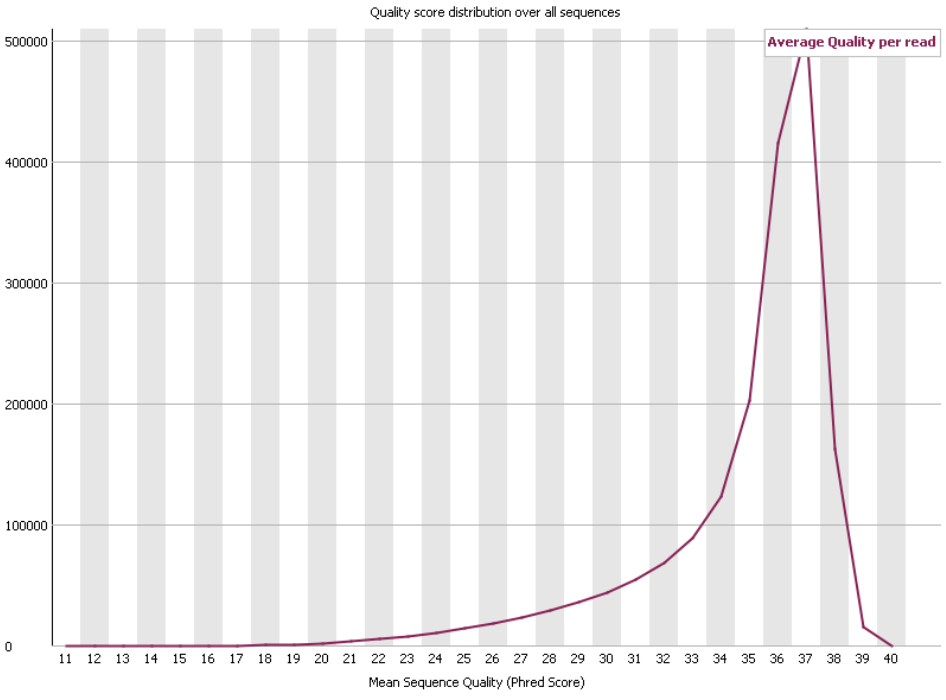


Figure 5

Per-sequence Pthred quality scores of the same genome



Tracing Feature Importance's as SNP determinants

By depriving a model of all priori ABR knowledge, features that the model deems most significant in its decisions become valuable resistance determinants like in GWAS. This with further genome annotation and amplification of SNP sites can help draw attention to casual gene regions and help surveil bacteria that acquire new resistance methods over time, a vital process in ABR research.

Multilabel Data

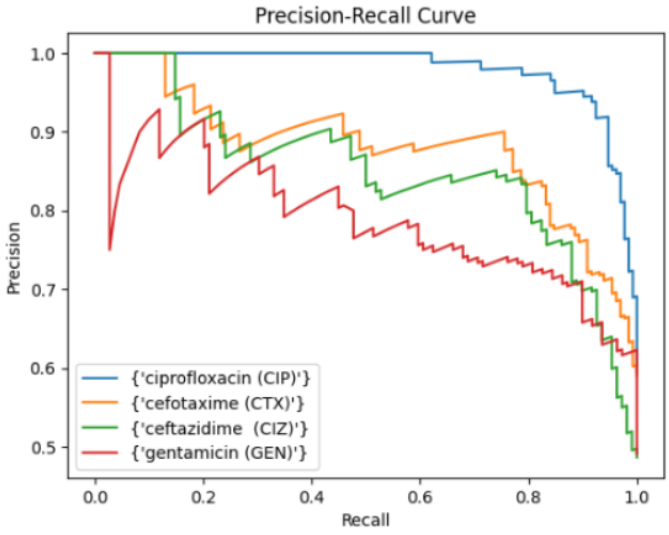
As almost all the samples are multidrug resistant, the dataset and model become *multilabel*, complicated almost every component of machine learning especially in evaluation metrics and package support. The Imbalanced, multi class and multilabel dataset can no longer be evaluated naively on accuracy. Instead, the accuracy of false and true positives, called *precision*, and that of negatives, called *recall* is considered. This is most visually represented through the *confusion matrix* diagram as shown in *figure 2*. Due to the significant class imbalance, 4 separate confusion matrices must be used for each antibiotic. The *f1* score is used to capture an accuracy-like metric per label where:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

We may further evaluate a model's full set of precision and recall metrics when adjusting the prediction assumption threshold, initially assumed to be 0.50 for each class, demonstrating all the possible model configurations if a particularly strict or lenient model is required, called a precision-recall curve. This, again, needing 4 separate curves:

Figure 5

Examples of a Precision recall curve, noticing that the closer it hugs the top right corner, the better the model is performing over all thresholds.



Sometimes receiver operating curves (ROC) are used instead, however when there is significant class imbalance a precision-recall curve is favored. *Area under the precision recall curve* (AUC-PR) is used to encapsulate the above diagram into a single value, between 0 and 1 similarly read to per class accuracy. It's worth noting as well that multilabel problems render many evaluation packages useless and interrupt the logic of many non-parametric algorithms, a problem later tackled in depth.

Oversampling

As biological data is often unpredictable, running into incredibly uneven data is expected in bioinformatics. The Synthetic Minority Oversampling Technique (SMOTE) is a popular choice for closing the sample number gap between imbalanced classes (Georgios Feretzakis et al., 2021; Pearcy et al., 2021; Ren et al., 2021). This method involves using a K-nearest neighbors' algorithm to interpolate new data points halfway along the line separating 2 existing data points. A powerful method but inappropriate for tokenized data like nucleotides, where adaptations of SMOTE exist, but none of which

facilitate multilabel output. Eventually conceding to minority duplication oversampling, same effect was achieved but running the risk of overfitting. Multiple correspondence analysis, hashing and ML-SMOTE all seem to lack proper support for datasets of this nature.

The Curse of dimensionality and feature selection

The number features (14,972 unique SNP's) far exceeded that number of samples (809) posing a classic challenge in machine learning called "*The curse of dimensionality*" (Geron, 2022). Making visualization, computational complexity, and overfitting a nightmare without strong regularization and dimensionality reduction. Following the latter, a throw-away Random Forest was fitted to the training set.

Random Forests are ensemble voting classifiers that train hundreds of decision trees, where the final classification is the most popular choice (vote) among them. Decision trees are iterative models that optimize a tree of splitting conditions as to "filter out" each class as purely as possible. By first using a per-tree Gini Impurity decrease algorithm, taking the Gini Impurity representing how useless a split is based on the number of samples it separates and compares how greatly it decreases when a feature is entirely removed. Features with minimal Gini decrease are considered important and once all feature importance are averaged among every tree in the ensemble, significant SNP loci can easily be gathered with precision, and the top 600 most informative SNPs were selected to form a reduced dataset.

Random Forests, Classifier Chains and Stacking

For marker detection, the same Gini impurity decrease method was used against the oversampled full dataset with optimized hyperparameters. To train other non-parametric models on multilabel data, a *classifier chain* is used.

Initially, 4 separate instances of a Random Forest were created for binary classification against each class simulating a multilabel classifier. However, entirely independent classifiers ignore inter-label relationships. Classifier chains retain a portion of the output of each binary classifier as input for each

other (Read et al., 2011) resonant of a RNN. Classifier chains of support vector machines and random forests were trained and evaluated. The predefined random order of classifiers for each feature gives chain classifiers significant variance, thus a voting classifier of 10 chains of random forests was made to curb this effect.

Finally, a stacking ensemble classifier was trained. In stacking, multiple models are trained against the dataset and an additional model is trained on the output of each candidate to make the final prediction, sometimes repeating this process and stacking classifier layers atop each other. The ensemble contained a random forest, support vector machine and logistic regression with another random forest as the decider model. Results were almost identical to the initial random forest tests suggesting it had little to contest within the ensemble.

Table

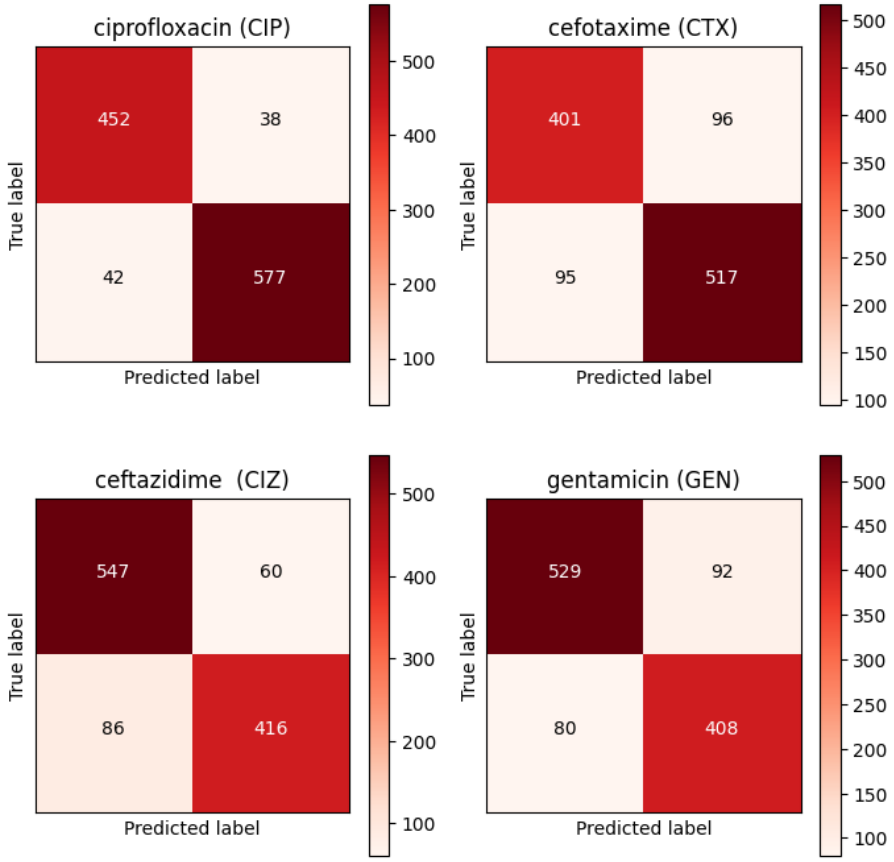
Per antibiotic F1 scores of each model

Antibiotic	**F1 scores**			
	Optimized Random Forest (RF)	**RF Classifier Chain**	**SVM Classifier Chain**	**LR Classifier Chain**
Ciprofloxacin (CIP)	*0.94*	*0.94*	*0.87*	*0.93*
Cefotaxime (CTX)	*0.84*	*0.83*	*0.77*	*0.82*
Ceftazidime (CIZ)	*0.85*	*0.84*	*0.73*	*0.81*
Gentamicin (GEN)	*0.84*	*0.83*	*0.42*	*0.82*

	RF Classifier Chain Ensemble	Stacking Ensemble
Ciprofloxacin (CIP)	*0.94*	*0.94*
Cefotaxime (CTX)	*0.83*	*0.84*
Ceftazidime (CIZ)	*0.84*	*0.85*
Gentamicin (GEN)	*0.83*	*0.83*

Figure 5

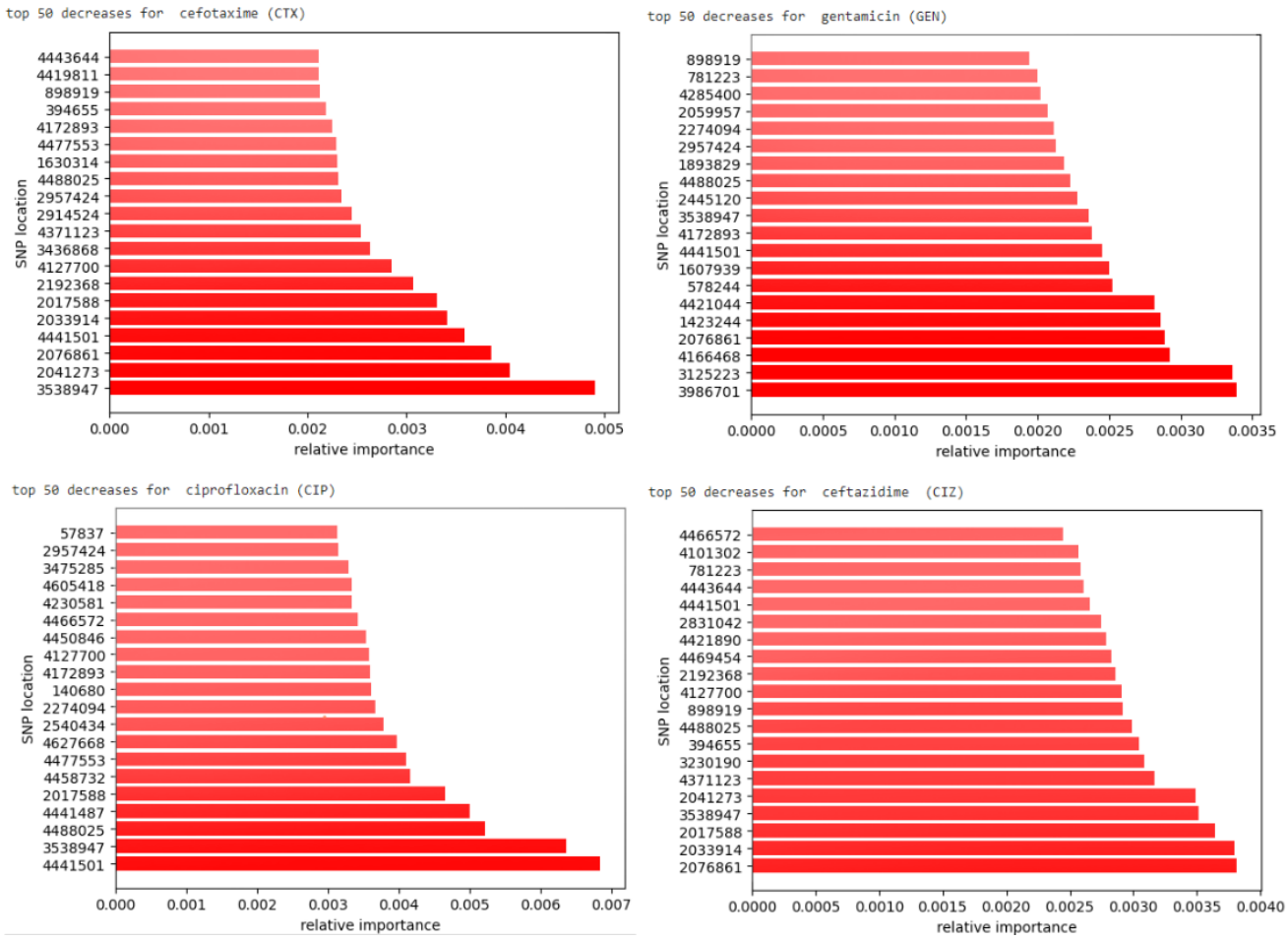
Per antibiotic confusion matrix of the best performing model – stacking ensemble



As all variations including the optimized random forest had minimal to no impact on the performance, the initial random forest was used for marker detection for simplicities sake.

Figure 6

Per antibiotic SNP feature importance's



Parametric and Performance centric approaches

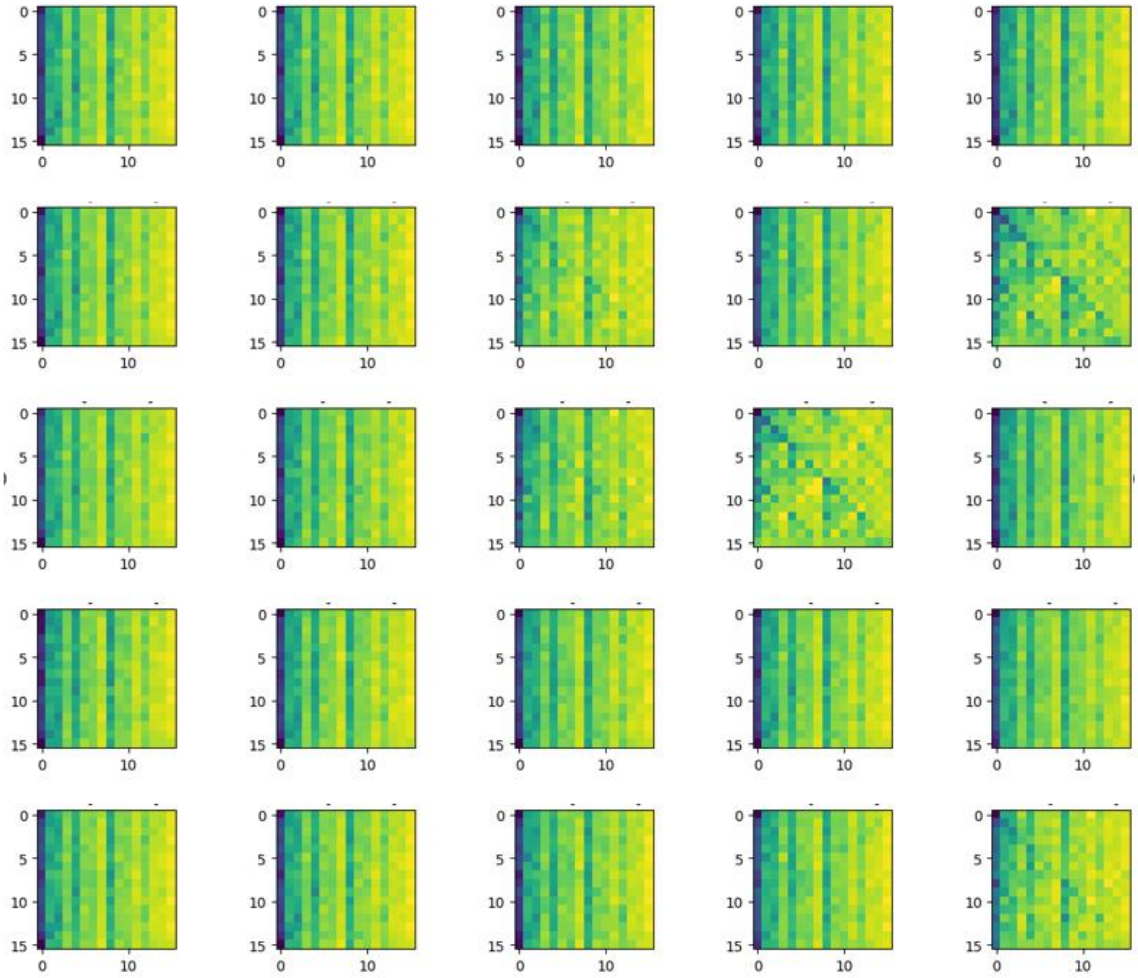
Alongside non-parametric solutions, studies also focus on highly parameterized neural networks for greater precision and sometimes in place of traditional models for feature selection (Shi et al., 2019). Catered preprocessing techniques were employed to maximize their potential.

Frequency-based Chaos Game Representation (FCGR)

Chaos Game representation (CGR) is a simple mathematical process famous for producing fractals and deducing the quality of random number generators through a random pixel drawing "game". The starting point for each pixel barely influences the outcome, thus CGR has garnered great attention in alignment-free genomics allowing whole genome sequences to be encoded identically regardless of read order, freeing a pipeline of the alignment burden (Löchel & Heider, 2021). A variant of CGR, Frequency CGR (FCGR), averages the pixel values of select cell sizes and effectively lowers its resolution, used solely for feature reduction (Lichtblau, 2019).

Figure 7

Colorized, class-wise FCGR of a few samples



Whilst FCGR is designed to shine with unaligned whole genomes, studies have still adopted it for forced feature reduction on SNP data (Ren et al., 2022) and is best practice for parametric models (Shi et al., 2019).

Approaches

A deep neural network (DNN) was used on the oversampled FCGR and feature-reduced SNP matrix. Fitted with 3 hidden layers (256, 128, 64), 2 dropout layers, L2 regularization and ADAM optimization. Learning rate scheduling and early stopping was used during training.

Additionally, a 2D Convolution neural network (CNN) was fitted to the oversampled FCGR dataset. The CNN had 4 convolutional layers (32, 64, 128) with batch normalization, pooling and drop layers between, L2 regularization and ADAM optimization.

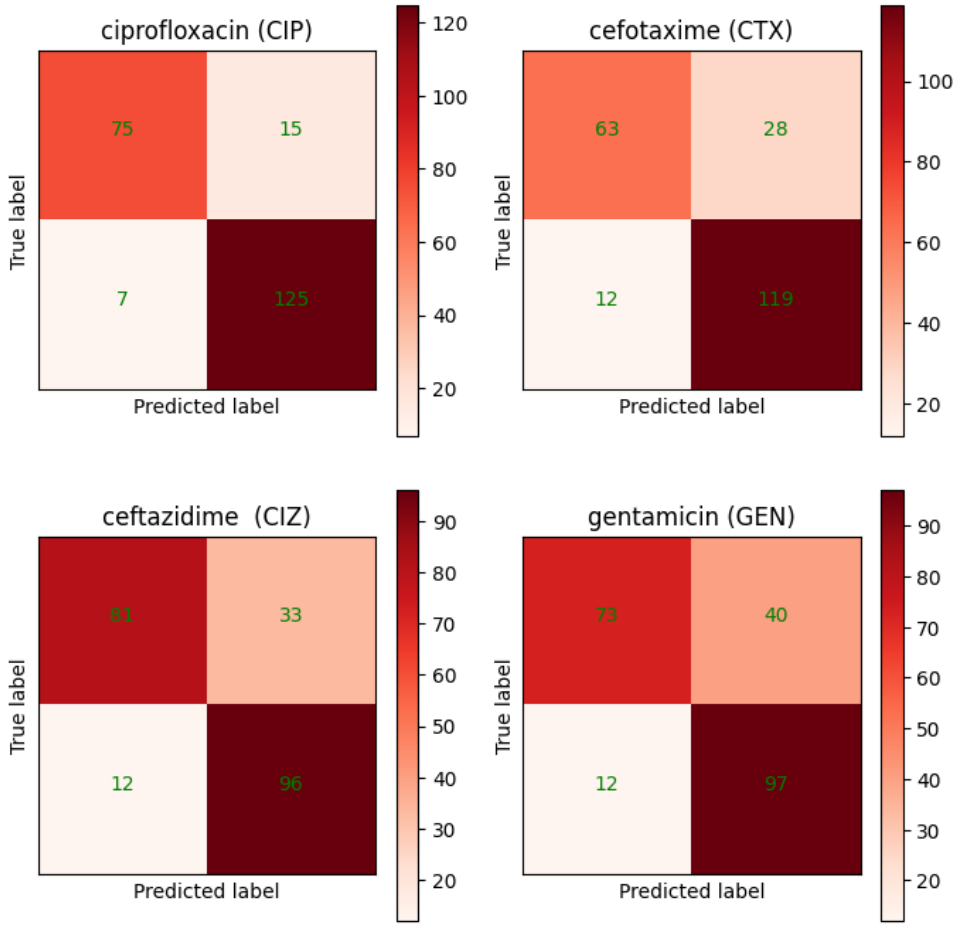
Table 2

Per-antibiotic f1 scores of all parametric approaches

Antibiotic	**F1 scores**		
	DNN & SNP matrix	**DNN & FCGR**	**CNN & FCGR**
Ciprofloxacin (CIP)	*0.83*	*0.93*	*0.92*
Cefotaxime (CTX)	*0.73*	*0.82*	*0.86*
Ceftazidime (CIZ)	*0.71*	*0.81*	*0.81*
Gentamicin (GEN)	*0.56*	*0.76*	*0.79*

Figure 8

Confusion matrices of highest performing model - CNN



Note; test set is significantly smaller as cross validation was not used.

The FCGR showed clear superiority to the feature selection method used on the SNP raw data. However, during FCGR, SNP determinants cannot be retraced easily.

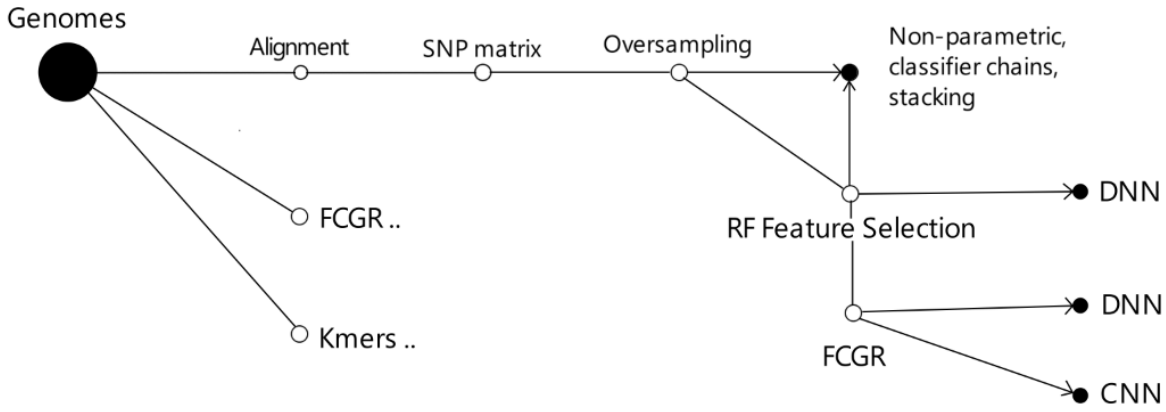
Insights & Future work

The parametric approaches were inferior to initial random forests, suggesting that ABR determinants aren't nested in complex patterns beyond their limited field of view. Tweaks and optimization had little to no effect on accuracy; the quality of this approach seems entirely dependent on the size and capacity of the dataset. F1 scores and AUC-PR's where on par with adjacent studies and expectedly higher that patient metadata approaches (Georgios Feretzakis et al., 2021), but it's important to note that predictive potential will vary significantly between different pathogens and drugs (Shi et al., 2019). Non-machine learning, statistical predictions seemed to have greater potential accuracy wise but lacks the bias-free, empirical intuition that this guides this approach.

Many SNP markers traced by feature importance had lined up with known resistance encoding genes regions as compared against the gene annotation of the same dataset (Ren et al., 2021). I was regrettably unable to commit to a proper gene annotation myself.

Figure 8

Depiction of approaches and possibilities explored in this report



Many improvements and reworks can be made here, curating a larger, diverse genomic dataset against proper resistance profiles is a requirement for developing reliable machine learning algorithms for professional use.

What's lacking?

The current state of machine learning in ABR research contains little evidence of progress, significant insights, or action. It seems to be in its very early stages and not nearly enough potential has been recognized to garner implementation. This approach lacks purposeful integration with existing workflows outside genome wide association studies of which it only contends with as an alternative. The lack of urgency and impact of this research should be considered moving forward.

Other approaches

A Supervised Pan-genome dataset of accessory genes provides faster, comprehendible profiles for miscellaneous use (Her & Wu, 2018).

Certain Neural networks are also capable of tracking SNP determinants and feature reduction on a SNP matrix, and if done wisely can give better results their non-parametric counterparts (Shi et al., 2019). Machine learning in genomics is certainly not confined to nucleotide variants, deep neural networks are more than capable of training on full unalignment genomes via the k-Mers and FCGR strategies for supervised, clustering and visualization purposes.

Reflection of this work

The overall execution of this project was a little messy and endured lots of wasted effort due to the many oversights. 70% of the work was eventually scrapped and the remaining project felt as if it lacked connection to purpose, action, and medicine. Far too much research weighed in medical sciences of which I am entirely untrained in, yet I committed extensively to as much surrounding knowledge as I could digest. Many lessons were hard learned this year and I'm eager to show my improvement moving forward.

References

Antibiotic resistance. (2020). [newsletter]. *2021*(2/28). Retrieved 31/July/2020, from https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance#:~:text=Antibiotic%20resistance%20occurs%20when%20bacteria,caused%20by%20non%2Dresistant%20bacteria.

Benkwitz-Bedford, S., Palm, M., Demirtas, T. Y., Mustonen, V., Farewell, A., Warringer, J., Moradigaravand, D., & Parts, L. (2021). Machine learning prediction of resistance to sub-inhibitory antimicrobial concentrations from Escherichia coli genomes. *bioRxiv*, 2021.2003.2026.437296. https://doi.org/10.1101/2021.03.26.437296

Feretzakis, G., Sakagianni, A., Loupelis, E., Kalles, D., Skarmoutsou, N., Martsoukou, M., Christopoulos, C., Lada, M., Petropoulou, S., Velentza, A., Michelidou, S., Chatzikyriakou, R., & Dimitrellos, E. (2021). Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy. *Healthc Inform Res*, *27*(3), 214-221. https://doi.org/10.4258/hir.2021.27.3.214

Feretzakis, G., Sakagianni, A., Loupelis, E., Kalles, D., Skarmoutsou, N., Martsoukou, M., Christopoulos, C., Lada, M., Petropoulou, S., Velentza, A., Michelidou, S., Chatzikyriakou, R., & Dimitrellos, E. (2021). Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy. *Healthcare Informatics Research*, *27*(3), 214-221. https://doi.org/10.4258/hir.2021.27.3.214

Geron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition* (3 ed., Vol. 1). O-Reilly Media.

Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. (2022). *Lancet*, *399*(10325), 629-655. https://doi.org/10.1016/s0140-6736(21)02724-0

Her, H.-L., & Wu, Y.-W. (2018). A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. *Bioinformatics*, *34*(13), i89-i95. https://doi.org/10.1093/bioinformatics/bty276

Lewin-Epstein, O., Baruch, S., Hadany, L., Stein, G. Y., & Obolski, U. (2020). Predicting Antibiotic Resistance in Hospitalized Patients by Applying Machine Learning to Electronic Medical Records. *Clinical Infectious Diseases*, *72*(11), e848-e855. https://doi.org/10.1093/cid/ciaa1576

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, *46*, 100511. https://doi.org/https://doi.org/10.1016/j.cosrev.2022.100511

Lichtblau, D. (2019). Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, *20*(1), 742. https://doi.org/10.1186/s12859-019-3330-3

Liu, X., Duan, R., Luo, C., Ogdie, A., Moore, J. H., Kranzler, H. R., Bian, J., & Chen, Y. (2022). Multisite learning of high-dimensional heterogeneous data with applications to opioid use disorder study of 15,000 patients across 5 clinical sites. *Sci Rep*, *12*(1), 11073. https://doi.org/10.1038/s41598-022-14029-9

Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Comput Struct Biotechnol J*, *19*, 6263-6271. https://doi.org/10.1016/j.csbj.2021.11.008

Medicine, S. (2013). Antibiotic Resistance In *Antibiotics - Lecture 9* (Vol. 2022). Youtube.

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., & Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLOS Computational Biology*, *14*(12), e1006258. https://doi.org/10.1371/journal.pcbi.1006258

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com/index.html

Pearcy, N., Hu, Y., Baker, M., Maciel-Guerra, A., Xue, N., Wang, W., Kaler, J., Peng, Z., Li, F., & Dottorini, T. (2021). Genome-Scale Metabolic Models and Machine Learning Reveal Genetic Determinants of Antibiotic Resistance in Escherichia coli and Unravel the Underlying Metabolic Adaptation Mechanisms. *mSystems*, *6*(4), 10.1128/msystems.00913-00920. https://doi.org/doi:10.1128/msystems.00913-20

Poirel, L., Madec, J. Y., Lupo, A., Schink, A. K., Kieffer, N., Nordmann, P., & Schwarz, S. (2018). Antimicrobial Resistance in Escherichia coli. *Microbiol Spectr*, *6*(4). https://doi.org/10.1128/microbiolspec.ARBA-0026-2017

Rahman, M. M., Lim, S. J., & Park, Y. C. (2022). Development of Single Nucleotide Polymorphism (SNP)-Based Triplex PCR Marker for Serotype-specific Escherichia coli Detection. *Pathogens*, *11*(2). https://doi.org/10.3390/pathogens11020115

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*(3), 333-359. https://doi.org/10.1007/s10994-011-5256-5

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., Hauschild, A.-C., Schwengers, O., & Heider, D. (2021). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, *38*(2), 325-334. https://doi.org/10.1093/bioinformatics/btab681

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., Hauschild, A. C., Schwengers, O., & Heider, D. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, *38*(2), 325-334. https://doi.org/10.1093/bioinformatics/btab681

Rezel-Potts, E., & Gulliford, M. (2022). Electronic Health Records and Antimicrobial Stewardship Research: a Narrative Review. *Current Epidemiology Reports*. https://doi.org/10.1007/s40471-021-00278-1

Ricciardi, W., Giubbini, G., & Laurenti, P. (2016). Surveillance and Control of Antibiotic Resistance in the Mediterranean Region. In *Mediterranean Journal of Hematology and Infectious Diseases* (Vol. 8, pp. e2016036).

Rivers, J. (2016). Antibiotic Resistance: An Old Solution but a New Problem. *addGene*. https://blog.addgene.org/antibiotic-resistance-an-old-solution-but-a-new-problem#:~:text=For%20routine%20infections%2C%20like%20strep,these%20infections%20is%20so%20rare.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

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 (Basel)*, *12*(3). https://doi.org/10.3390/antibiotics12030452

Segerman, B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories [Perspective]. *Frontiers in Cellular and Infection Microbiology*, *2*. https://doi.org/10.3389/fcimb.2012.00116

Shi, J., Yan, Y., Links, M. G., Li, L., Dillon, J. R., Horsch, M., & Kusalik, A. (2019). Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics*, *20*(Suppl 15), 535. https://doi.org/10.1186/s12859-019-3054-4

Srinivasu, P. N., Shafi, J., Krishna, T. B., Sujatha, C. N., Praveen, S. P., & Ijaz, M. F. (2022). Using Recurrent Neural Networks for Predicting Type-2 Diabetes from Genomic and Tabular Data. *Diagnostics (Basel)*, *12*(12). https://doi.org/10.3390/diagnostics12123067

Thänert, R., Reske, K. A., Hink, T., Wallace, M. A., Wang, B., Schwartz, D. J., Seiler, S., Cass, C., Burnham, C.-A. D., Dubberke, E. R., Kwon, J. H., & Dantas, G. (2019). Comparative Genomics of Antibiotic-Resistant Uropathogens Implicates Three Routes for Recurrence of Urinary Tract Infections. *mBio*, *10*(4), 10.1128/mbio.01977-01919. https://doi.org/doi:10.1128/mbio.01977-19

Zhou, R., Wang, W., Padoan, A., Wang, Z., Feng, X., Han, Z., Chen, C., Liang, Y., Wang, T., Cui, W., Plebani, M., & Wang, Q. (2022). Traceable machine learning real-time quality control based on patient data. *Clin Chem Lab Med*, *60*(12), 1998-2004. https://doi.org/10.1515/cclm-2022-0548