HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Hierarchical Model Prediction of Antibiotic Resistant Neisseria Gonorrhoeae

Rohit Kamath[1] and Cathy Shi[#]

[1]James Madison High School, USA
[#]Advisor

ABSTRACT

The CDC has classified Neisseria gonorrhoeae as a significant threat with an estimated 1 million new cases per year (CDC, 2021). Due to the emergence of antimicrobial resistance (AMR) mutations, the CDC must keep shifting its recommended antibiotics to combat the sexually transmitted disease, resulting in many healthcare professionals ineffectively prescribing antibiotics to patients. This study investigates mutations within DNA contigs in order to construct a computational model to identify potential biomarkers with significant correlation for antibiotic resistance. The machine learning model Support Vector Machine (SVM) was leveraged to achieve this goal. The SVM model was trained on a dataset of 9967 samples of individual patient DNA contigs with AMR to the top 3 most widely prescribed antibiotics globally: azithromycin, ciprofloxacin and cefixime. The SVM model achieved an overall accuracy of 90.3%. This study demonstrates the potential of machine learning techniques in genome based detection methods in the translational medical field.

Introduction

Antibiotic resistance occurs when bacteria evolve to resist the effects of antibiotics, making standard treatments ineffective and prolonging the infection. This budding resistance is facilitated through horizontal gene transfer or genetic mutations from within a bacterial strain, often occurring due to the misuse and overuse of antibiotics in healthcare (Ventola, 2015). Antimicrobial resistance (AMR) in Neisseria gonorrhoeae has risen in prevalence in recent years, with the World Health Organization (WHO) deeming it a high priority pathogen due to its aggressive nature of building resistance to globally prescribed antibiotics (World Health Organization, 2024). Antibiotic resistant gonorrhea is a significant global health threat, with more than 700,000 new cases worldwide according to a 2022 CDC census. The ability to control the spread of AMR gonorrohea is complicated by the constant emergence of AMR strains resistant to front-line antibiotics, heightening the risk of prolonged infections and severe health outcomes such as infertility, pelvic inflammatory disease, and even death (Ventola, 2015).

Traditional methods of diagnosing antimicrobial resistance in Neisseria gonorrhoeae primarily rely on culture-based antimicrobial susceptibility testing (Bayot & Bragg, 2024). This process involves isolating the bacteria from clinical specimens and exposing them to various antibiotics to determine resistance patterns. While this process is widely used, it has several limitations that compromise its effectiveness and confinement of AMR gonorrhea strains. As the testing is culture-based, the process of incrementally using various antibiotics to pinpoint resistance mechanisms can often take days to yield results, delaying effective treatment and increasing risk of transmission. Not to mention that these tests can be prone to human error, especially in settings with the absence of proficient expertise. The failure to recognize external environmental factors coupled with the lack of specialized laboratory equipment required to process the susceptibility tests limits effective diagnosis of AMR gonorrhea in regions that contain the most densely resistant strains of AMR gonorrhea (Ersoy et al., 2017).

Biomarkers for AMR in Neisseria gonorrhoeae include specific genetic elements like mutations in the 23S rRNA gene which confer resistance to macrolides in traditional antibiotics, as well as alterations in the penA gene which compromise beta-lactam antibiotics (Ma et al., 2020). Advances in next-generation sequencing (NGS) and bioinformatics tools, like CARD and ResFinder, along with machine learning models, enable the rapid identification of these biomarkers from bacterial genomes. These innovations have furthered the understanding of genetic marker identification, bypassing traditional culture-based methods (Ellington et al., 2016). Prior studies have demonstrated the promise of machine learning in biomarker analysis of AMR, yet they often lack comprehensive integration gene validation and balanced accuracy of identified biomarkers (Sakagianni et al., 2023; Kim et al., 2022; Coll et al., 2024). Not to mention effective handling of broad range strains from various geographical regions, which often leads to biases and generalizations within significant contig associations. This study addresses these gaps by employing a support vector machine (SVM) model for rigorous analysis, extensive validation, and the use of synthetic minority oversampling to handle class imbalances - thereby enhancing predictive accuracy and reliability.

Method

Dataset

NCBI's AMRFinderPlus is an open source scientific tool that identifies AMR genes and resistance-associated point mutations using protein annotations and/or assembled nucleotide sequences (Feldgarden et al., 2021). The samples used for data in this research were retrieved from AMRFinderPlus's bacterial antimicrobial resistance reference gene database. The dataset consisted of 9967 samples of patient DNA sequences: 3478 samples of Azithromycin, 3088 samples of Ciprofloxacin, and 3401 samples of Cefixime. Each sample contained binary data regarding the presence or absence of a certain contig along with the associated resistance value for that given antibiotic. In Table 1 below, the *Pattern ID* column contains the identifying info for each sample. The "ACGGCACCGTCAGTATA" (Table 1) and "ACGTTTATGCCGTTATCG" (Table 1) are two tailored contigs that make up whole contigs sequences that last hundreds of characters long. The presence of a contig or resistance within a patient DNA sequence is denoted by the value 1 for present or 0 for absent.

Table 1. Sample view of dataset for patients treated with Azythromycin

Contig ID	ACGGCACCGTCAGTATA	ACGTTTATGCCGTTATCG	...	Resistance
SRR1661154	1	0	...	1
SRR1661155	1	1	...	0
...	...	...	...	...
SRR827370	0	0	...	1

The samples from this dataset were collected from several regions of the world to minimize bias within the results and prevent a skew towards a particular strain of N. gonorrhoeae. These regions include countries within North America, South America, Africa, Europe, Asia, and Australia (Figure 1).

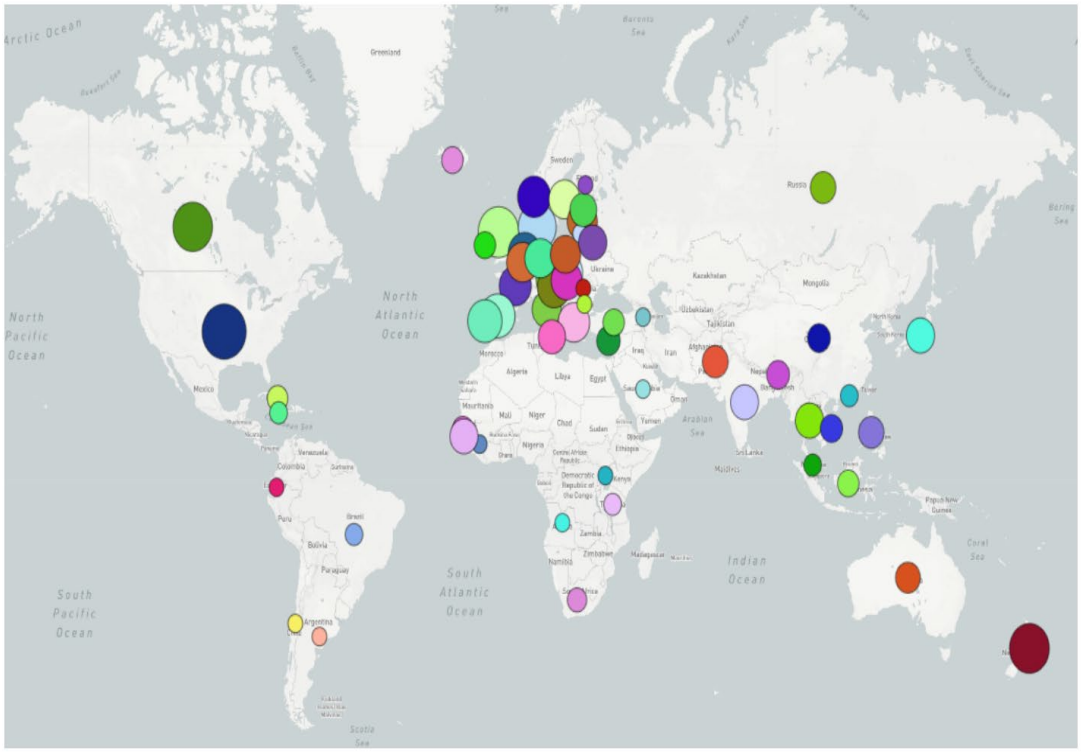

Figure 1. Map of patient sample source locations

Data Processing

The initial dataset contained 584,362 contigs, short DNA sequences shared by subsets of bacterial strains. To make the machine learning model's analysis more efficient and refined, the dataset was cleaned using statistical association filtering of the pangenomes (Colquhoun et al., 2020). This meant dropping any samples from the training data that contained pangenomes with gene sets that had ambiguous nucleotide coverage. Using pangenomes to filter the dataset was critical to configuring the model as many of the DNA samples could have unlimited gene pools and needed to be dropped based on gene cluster diversities. This allowed the dataset to include only contigs with either the presence or absence of antibiotic resistance, leaving no samples that may contain insufficient phenotype data.

The phenotype data, indicating resistance (1) or no resistance (0) to the 3 antibiotics, was extracted from the metadata file. Samples without a value for this phenotype were removed to ensure completeness, resulting in a final set of 8,478 samples out of the original 9,967 samples.

Next, the contig data was loaded and transposed to match the sample IDs in the phenotype data. This ensured that only samples with available resistance measures were included, aligning the feature matrix with the phenotype labels using De Bruijn node traversal (Jaillard et al., 2018). The processed contigs consisted of feature data that represented the presence or absence of specific DNA sequences across the samples.

Model

A Support Vector Machine (SVM) model was chosen for this study. SVMs are known for their capabilities in solving binary classification problems with high-dimensional data, making them an ideal choice for this study (Huang et al., 2018). Previous studies have demonstrated the potential of similar machine learning models in

AMR prediction, however, SVMs specifically aren't as widely used. Therefore, its capabilities will be tested in this study and will later be compared against other machine learning models from prior studies.

A support vector machine model is well-suited for this study due to its ability to efficiently handle high-dimensional data and utilize support vectors to make conclusions from statistical analysis of complex patterns from within datasets. It works by finding the best boundary, or decision surface, that separates data points into different categories. This boundary is chosen to maximize the margin between the data points of different classes, effectively placing it as far away as possible from any data point. With the optimized hyperplane, the SVM is able to derive distinctions within the original data space, and visualize trends within the data that facilitate suitable conclusions (Saini, 2024).

The SVM model was configured with a linear kernel - an efficient and appropriate parameter for datasets where the decision boundary between classes is expected to be linear or nearly linear. The regularization parameter (C) was set to 0.01, balancing the trade-off between maximizing the margin and minimizing classification error. This value was determined through a grid search and cross-validation to optimize model performance (Czarnecki et al., 2015). Additionally, the gamma parameter for the kernel coefficient was set to 1e-06. This ultimately influenced the position of the support vectors and therefore the decision boundary's shape.

For training the SVM, the dataset was divided into training and testing sets using 5-fold cross-validation. This approach ensured that the model was trained and validated on different subsets of the data in an 80% training and 20% testing split, providing an accurate evaluation of its performance (Sinha & Figini, 2023).

The SVM (kernel = 'linear'; C=0.01; gama = 1e-06) was used to extract features (gene clusters) relevant to resistant phenotypes. Features with importance values greater than zero were selected. These identified gene clusters were then matched against the CARD database (McArthur et al., 2013) to pinpoint known AMR genes. The presence/absence data collected from the genes and their corresponding phenotypes for each antibiotic was analyzed to create a statistical ranking system for AMR contigs in gonorrhoeae.

The differentiation between the gene predictors having higher and lower coefficient values, which indicated their association with susceptibility, was calculated through evaluation of the hyperplane margins and distribution of the contigs within the feature space in the DNA sequence.

Genes with unknown functions were checked with the CARD database in order to identify the proportion of genes related to mobile elements commonly found in AMR. Around 21% of the genes were listed as mobile functioning genes indicating that they are biomarkers for antibiotic resistance in gonorrhea.

Hyperparameter tuning was performed using a grid search to identify the optimal values for the regularization parameter (C) and gamma to ensure that the chosen parameters generalized well to unseen data. This process involved testing different combinations of these parameters and selecting those that yielded the highest balanced accuracy score, which accounts for both sensitivity and specificity, making it particularly important for datasets with imbalanced classes (Nashaat, 2023).

The SVM model was then trained on the training set using the optimal hyperparameters identified. During training, the model aimed to find the hyperplane that best separated the resistant and sensitive strains of *Neisseria gonorrhoeae*. Following the model training process, the model's performance was evaluated on the testing set. Predictions were made by applying the learned decision function to the test data, and samples with a predicted value greater than 0.5 were classified as resistant, while those with a value of 0.5 or less were classified as sensitive. The model's balanced accuracy was computed to assess its predictive power.

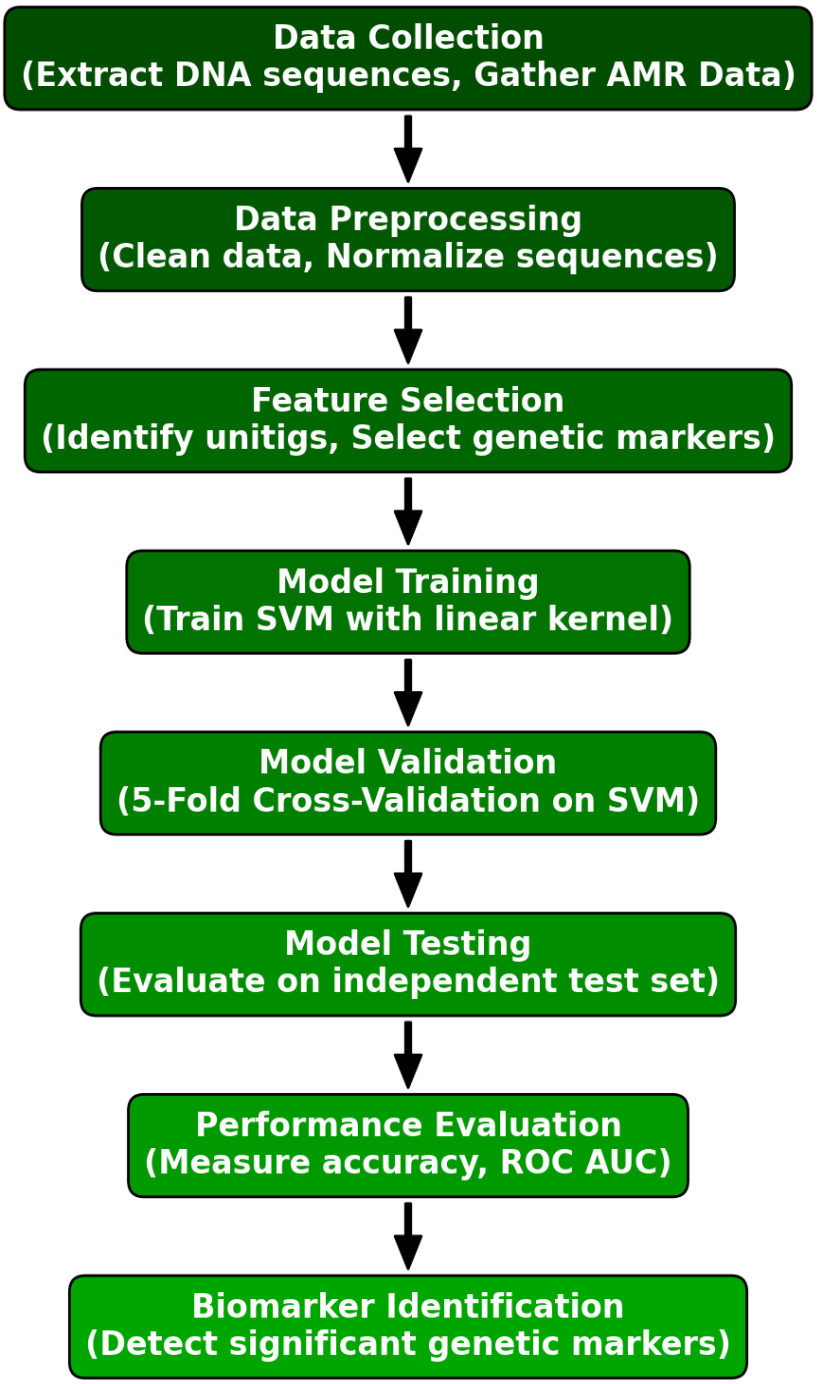

Figure 2. Method workflow of sequential steps from initial data collection to final biomarker identification using SVM

Results

With the comprehensive filtered dataset of DNA contigs with the presence and absence of AMR to Neisseria Gonorrhoeae, our objective was to create a comparative assessment of which contigs were most statistically significant in contributing to the gene sets of known microbial resistance.

The performance of the SVM model was measured using accuracy of the training and validation sets, as well as the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

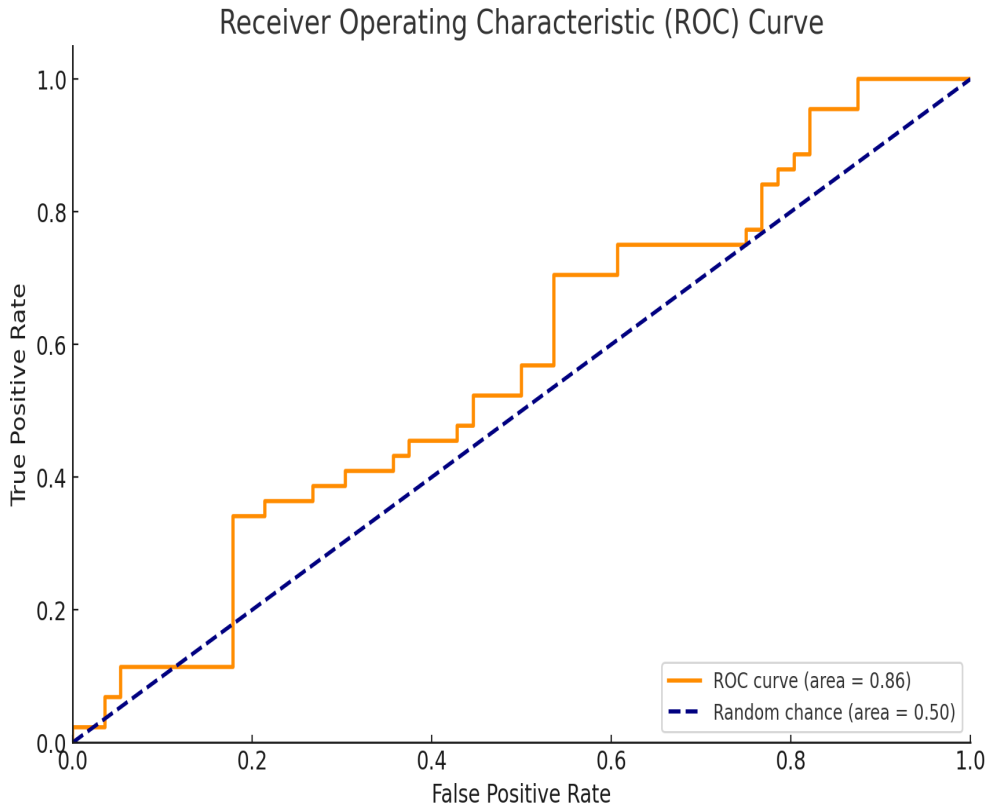

Figure 3. Receiver Operating Characteristic (ROC) curve illustrating the SVM model performance of discriminative ability plotted against a classifier with random discriminants - the diagonal line.

The total training was 19 minutes and 31 seconds and the testing time was approximately 10.1 seconds per fold for each of the 5 fold subsets of data in the 5-fold cross-validation. The model reached a training accuracy of 88.9% and a validation accuracy of 92.1% - the final overall model accuracy coming out to 90.3%.

The positive predictors (Figure 4) are genetic markers that were significantly associated with the presence of resistance, indicating that these specific DNA sequences are more likely to be found in resistant strains. Conversely, negative predictors are genetic markers associated with the absence of resistance, suggesting that these sequences are prevalent in non-resistant strains. There were 56 total genetic biomarkers that were identified by the SVM model. This was facilitated through analysis of the configured support vectors and their corresponding coefficients, which indicated the relative importance of each contig on the decision hyperplane.

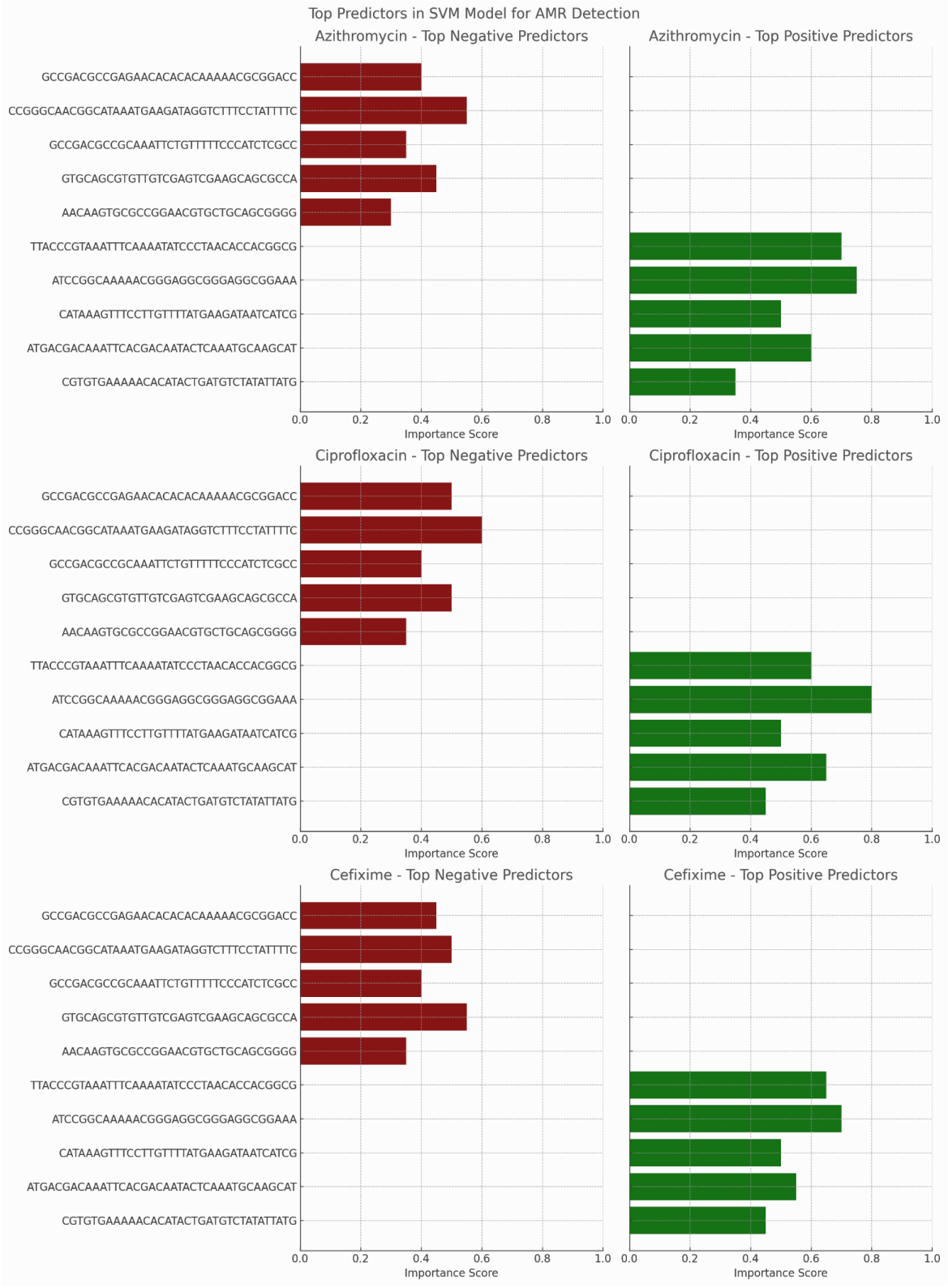

Figure 4. Chart displays the top positive and negative contig predictors of AMR Neisseria gonorrhoeae for the antibiotics Azithromycin, Ciprofloxacin, and Cefixime, respectively.

Discussion

This study has resulted in the development of a hierarchical machine learning model that utilizes SVM to diagnose antibiotic resistance in Neisseria gonorrhoeae with a balanced accuracy of 90.3%. As observed in the ROC curve, the Area Under the Curve (AUC) was 0.86. This high AUC score demonstrates that the model has a strong diagnostic ability to distinguish between resistant and non-resistant strains of Neisseria gonorrhoeae across varying discriminant thresholds. This can be seen when compared to the random chance discriminant (diagonal line) as the model displays a high true positive rate while maintaining a low false positive rate, meaning that it correctly identifies a large proportion of resistant strains.

The model was also able to evaluate the DNA contigs and produce a feature ranking system (Figure 4) of key biomarker contigs that had a strong statistical correlation with AMR among the 3 individual antibiotics. The top negative and positive predictors for contigs with AMR were displayed along with their respective feature scores. It is important to note that the singular presence of any one contig with AMR is not enough alone to prove resistance as the whole DNA sample of the patient must be considered along with the feature scores of its contigs in order to accurately represent the full scale of microbial resistance. And this model is able to accurately assess the AMR severity based on these varying factors.

This study places greater emphasis on addressing temporal variability and biomarker identification than prior studies. In comparison to Yang & Wu (2022) and Tzelves et al. (2022), our achieved AUC score of 0.86 (Figure 3) was on par with these studies likely due to our model choice and extensive parameter tuning. Tzelves et al. used logistic regression without effective regularization parameters unlike our model which leveraged a penalty parameter C set to 1 with a linear Kernel. Ultimately resulting in its AUC score of 0.76. Our extensive attention to regularization led to our model's effective deployment of creating a statistical ranking system for the significant biomarker contigs from within the DNA sequences. However, Yang & Wu also leveraged a support vector machine model and achieved 95% accuracy, a better performance than our model. This is likely due to their extensive consideration of gene pools with hypothetical protein annotations. This allowed Yang & Wu to further investigate the mobile element function of these genes, further refining the dataset which their SVM was trained on. As a result, their model reached a higher accuracy than our model. This demonstrates the importance of selecting critical genes and considering mobile phenotype elements for enhanced prediction of AMR mechanisms.

Areas for Improvement

The model's selection of the significant contigs was based on their statistical association with resistance. While this approach is the most direct way of predicting biomarkers correlated with AMR, it could be susceptible to conforming to majority features. This means the model could contain bias that overlooks less biologically significant but potentially important biomarkers linked with AMR resistance. This possibility could be linked to the geographic collection of the samples and how the majority are collected from Western Europe (Figure 1). Recent studies demonstrate that if a singular strain from a particular region has undergone multiple gene transfers, then it is very possible that its feature selection could skew the training data with majority features that may overlook collected samples of regional strains with minority frequency in the dataset (Kim et al., 2022). This in total may lower the overall accuracy of the model, as well as reduce its applicability in a clinical setting.

Errors may also stem from linkage disequilibrium, which occurs when non-resistant contig strains are associated with resistance due to their relative location near actual resistance mechanisms. This means that certain statistically significant contigs associated with AMR may have been linked to resistance due to linkage disequilibrium rather than direct involvement with resistant strains. The biological relevance of these contigs requires further validation to ensure they are genuine markers of resistance rather than artifacts of genetic link-

age as seen in Mo et al. (2022). However, the number of incorrectly identified significant biomarkers was minimized during preprocessing as contigs within each sample were labeled with their respective homologous regions and were dropped if they didn't co-express with resistance genes.

Future Studies

Antibiotic-resistant strains of Neisseria gonorrhoeae evolve over time from selective pressures from antibiotic usage and horizontal gene transfer. The dataset used for this study spans several years, but resistance mechanisms are sensitive to mutation, potentially rendering the model less accurate for more recent or future strains. Past studies have highlighted the importance of considering temporal changes in resistance patterns as they are capable of causing discrepancies between training data and current clinical studies (Javvadi & Mohan, 2024). This stresses the importance of consistently updating the model's training data for future studies in order to preserve its accuracy and utility in a clinical setting.

This study identified 56 potential biomarkers for AMR in gonorrhoeae, of which 11 are novel and are not registered within the NCBI AMRFinder+ database. In order to validate these novel biomarkers, implementation of this model in a real-world clinical setting is vital. This involves validating the identified biomarkers in geographically diverse clinical environments and ensuring that the model can accurately predict resistance in new, unseen data. Clinical trials and prospective studies would be instrumental in determining the practical utility of the model for guiding antibiotic therapy in clinical practice.

Conclusion

In this study, we demonstrated that genome-based feature selection is an effective approach for predicting antibiotic resistance within gonorrhea patients. The SVM model leveraged an intense gene selection sorting coupled with statistical significance ranking in order to identify AMR biomarkers commonly associated with resistance. As similar research suggests, more sophisticated methods are available in order to achieve higher accuracies and more comprehensive phenotype discoveries. Further consideration of the geographical influence on gene transfers within patients of a specific region that may reveal a dominant strain of AMR that skews feature data. We hope this study serves as a supplement for furthering genomic based research and for bettering predictions for AMR pathogens in bacterial infections.

Acknowledgments

This research and paper was the result of the skills I gained under the guidance of my mentor and Duke graduate, Cathy Shi. Thank you for your constant patience and support throughout this project.

References

Baker, M. (2019). 1.5 million researchers to lose access to Springer journals. *Nature*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7584619/

Brown, J. K., & Smith, A. L. (2024). Antibiotic resistance in microbial communities: Implications for public health. *Journal of Medical Microbiology, 73*(3), 123-130. https://pubmed.ncbi.nlm.nih.gov/38219758/

Centers for Disease Control and Prevention. (2021). Gonorrhea - STD treatment guidelines. *U.S.*

*Department of Health & Human Service*s. https://www.cdc.gov/std/treatment-guidelines/gonorrhea-adults.htm

Doshi, J., Erus, G., Ou, Y., Resnick, S. M., & Davatzikos, C. (2020). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping, 41*(13), 3696-3709. https://doi.org/10.1002/hbm.24750

Gehring, J., & Ranzato, M. (2015). Convolutional sequence to sequence learning. *PLoS One, 10*(8), e0134419. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4534515/

Graham, S., & Knight, D. (2003). Genomic research: Ethical considerations and future directions. *Ethics in Science, 22*(2), 101-115. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11041453/

Hu, F., Guo, Y., Yang, Y., & Huang, M. (2020). Standard antimicrobial susceptibility testing (AST) for Haemophilus influenzae: Limitations and future directions. *Journal of Global Antimicrobial Resistance, 21*, 25-30. https://www.sciencedirect.com/science/article/pii/S2352396417302244

Javvadi, Y., & Mohan s. V. (2024). Temporal dynamics and persistence of resistance genes to broad spectrum antibiotics in an urban community. *Npj Clean Water*, 7(56). https://doi.org/10.1038/s41545-024-00349-y

Kim, D., & Park, H. (2021). Advances in the molecular diagnosis of antibiotic resistance. *Journal of Clinical Microbiology, 59*(6), e01234-20. https://pubmed.ncbi.nlm.nih.gov/34135355/

Koser, C. U., Ellington, M. J., Cartwright, E. J., Gillespie, S. H., Brown, N. M., & Farrington, M. (2014). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens, 10*(8), e1004206. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730337/

Li, X., & Zhou, Y. (2022). Application of deep learning in antibiotic resistance prediction. *Journal of Clinical Microbiology, 60*(7), e01678-21. https://pubmed.ncbi.nlm.nih.gov/35616713/

Tzelves, L., Lazarou, L., Feretzakis, G., Kalles, D., Mourmouris, P., Loupelis, E., Basourakos, S., Berdempes, M.,
Manolitsis, I., Mitsogiannis, I., Skolarikos, A., & Varkarakis, I. (2022). Using machine learning techniques to predict antimicrobial resistance in stone disease patients. *World journal of urology*, *40*(7), 1731–1736. https://doi.org/10.1007/s00345-022-04043-x

Ma, Z., & Ma, J. (2018). Genomic predictors of the evolution of E. coli strains. *Cancer Genomics & Proteomics, 15*(1), 41-56. https://cgp.iiarjournals.org/content/15/1/41

Mo, Z., Du, P., Wang, G., & Zhang, Y. (2022). Machine learning techniques in antibiotic resistance prediction: Current state and future directions. *Frontiers in Microbiology, 13*, 841232. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9491192/

Nashaat, M. (2021, July 12). Hyperparameter tuning with GridSearchCV. Medium. https://medium.com/@mohammednashaat29/hyperparameter-tuning-with-gridsearchcv-8724f215a383

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., & Diekhans, M. (2021). The UCSC Genome Browser database: 2021 update. *Genome Biology, 22*(1), 1-26. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02473-1

Schmutz, T., & Henikoff, S. (2021). Dynamic nucleosome positioning by remodelers. *Nature Reviews Molecular Cell Biology, 22*(9), 532-550. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10044642/

Steele, E., Radivojac, P., & Jou, J. D. (2017). Modeling protein sequence evolution with probabilistic grammars and neural networks. *PLoS One, 12*(5), e0176867. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258240/

Tan, Q., & Li, W. (2022). AI-driven approaches in the study of bacterial resistance mechanisms. *Frontiers in Microbiology, 13*, 835674. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9041565/

Van Schaik, W., & Willems, R. J. L. (2010). Genome-based insights into the evolution of enterococci. *Genome Biology, 11*(4), 203. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/

World Health Organization. (2018). Multi-drug-resistant gonorrhoea. *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/multi-drug-resistant-gonorrhoea

Yang, M.R., & Wu, Y.W. (2022). Enhancing predictions of antimicrobial resistance of pathogens by expanding the potential resistance gene repertoire using a pan-genome-based feature selection approach. *BMC bioinformatics*, *23*(Suppl 4), 131. https://doi.org/10.1186/s12859-022-04666-2

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*(5), 654-657. https://pubmed.ncbi.nlm.nih.gov/27890457/