medical sciences forum

MDPI

Proceeding Paper

A Systematic Implementation of Machine Learning Algorithms for Multifaceted Antimicrobial Screening of Lead Compounds [†]

Justin Shen [1,*] and Davesh Valagolam [2]

[1] School of Engineering, Stanford University
[2] School of Engineering and Applied Science, University of Pennsylvania; daveshv@seas.upenn.edu
[*] Correspondence: jshen3@stanford.edu

[†] Presented at the 2nd International Electronic Conference on Antibiotics—Drugs for Superbugs: Antibiotic Discovery, Modes of Action And Mechanisms of Resistance, 15–30 June 2022; Available online: https://eca2022.sciforum.net/.

Citation: Shen, J.; Valagolam, D.A Systematic Implementation of Machine Learning Algorithms for Multifaceted Antimicrobial Screening of Lead Compounds. *Med. Sci. Forum* **2022**, *2*, x. https://doi.org/10.3390/xxxxx

Academic Editor:

Published: date

Abstract: This study employed machine learning algorithms to identify lead compounds that inhibit the antibiotic targets, DNA gyrase and Dihydrofolate reductase in *Escherichia coli*, and identified new, multifaceted antimicrobial compounds. This study used three separate datasets: 1) 326 Escherichia coli DNA gyrase inhibitors and 132 non-inhibitors, 2) 346 Escherichia coli Dihydrofolate reductase inhibitors and 176 non-inhibitors, and 3) 18387 non-specific drug-like chemicals. All datasets were then processed using ECFP-4 fingerprints and split into train, test, and validation datasets according to a 70-15-15 train-test-validation split. We explored the potential of six different classification algorithms, all optimized with Bayesian optimization. Our results indicate that the Gradient Boosting Classifier (GBC) performed the best at identifying a compound's efficacy towards DNA gyrase with an accuracy, precision, recall, F1-score, and AUC of 0.91, 0.92, 0.86, 0.88, and 0.933, respectively. The Random Forest Classifier (RFC) performed optimally for identifying a compound's effectiveness towards Dihydrofolate reductase with an accuracy, precision, recall, F1-score, and AUC of 0.86, 0.83, 0.85, 0.84, and 0.944, respectively. As a result, the GBC and RFC were used to search for compounds that inhibited both DNA gyrase and Dihydrofolate reductase. Out of 18387 compounds, we identified 5 novel compounds that have a predicted probability greater than 95% to inhibit both DNA gyrase and Dihydrofolate reductase, suggesting a high antimicrobial potential. The models evaluated in this study, particularly the GBC and RFC models, hold tremendous promise in computationally screening large libraries of compounds for antimicrobial potential.

Keywords: Data Science, Machine Learning, Antimicrobial, Antibiotic, Superbugs, Classification

1. Introduction

1.1. Background on Antibiotics and Antibiotic Resistance

Amidst the recent explosion of antibiotic use in both humans and agriculture, antibiotic resistance in bacterial strains has begun to spike. This has led to the advent of "superbugs", bacteria that have developed resistance to multiple antibiotics [1]. As a result, there have been numerous research efforts in recent years aiming to identify new antibiotics.

1.2. Recent Advances in Computational Drug Discovery: Applications to Antimicrobial Compounds

Recent advances in machine learning and computational biology have demonstrated the potential to accelerate computational drug discovery by filtering the chemical space for target molecules [2,3]. Previous researchers have demonstrated the efficacy of random forest classification models when predicting for life-extending chemicals [4]. Furthermore, most recently, researchers have demonstrated the potential of deep learning models particularly in identifying novel antibiotics that are successful in vivo against a wide range of bacterial infections, indicating the potential for computational methods to revolutionize antibiotic bacteria [5].

1.3. DNA gyrase and Dihydrofolate reductase as Antimicrobial Targets

Many current antimicrobial compounds operate by inhibiting the function of key proteins that are vital to bacterial function [6]. This study focused on two such proteins proposed by prior literature as antimicrobial targets, DNA gyrase and Dihydrofolate reductase. DNA gyrase functions as topoisomerase in bacteria that aids in the process of ATP-dependent negative supercoiling of DNA in bacteria [7]. Previous successful antibiotic classes like Coumarins and Quinolones have modulated the function of DNA gyrase, leading to antimicrobial function via the breakdown of bacterial function [8]. Similarly, Dihydrofolate reductase has also been a popular target for antimicrobial agents due to its crucial role in nucleotide synthesis [9].

1.4. Purpose

In an effort to speed up antibiotic discovery, this study demonstrated the promise of machine learning classification models in multifaceted antimicrobial compound screening and identification.

2. Materials and Methods

2.1. Datasets and Dataset Preprocessing

The breakdown of the three datasets used in this research are displayed in Table 1. The datasets consisting of *Escherichia coli* DNA gyrase and Dihydrofolate reductase inhibitors were sourced from ChEMBL [10]. The dataset consisting of 18387 non-specific drug-like chemicals was sourced from Zinc 15. [11]

Table 1. Dataset Breakdown.

Data Type	DNA Gyrase	Dihydrofolate reductase	Unspecific
Inhibitor	326	346	0
Non-inhibitor	132	176	0
Unspecific	0	0	18387

All compounds in datasets were characterized using ECFP-4 fingerprints. All datasets were then split into train, test, and validation datasets according to a 70-15-15 train-test-validation split.

2.2. Machine Learning Models

This study employed six classification models in total, logistic regressions (LR), support vector machines (SVM), random forests (RFC), k-nearest neighbors (K-NN), AdaBoost (ADA), and Gradient Boosting (GBC). All models were evaluated using accuracy, prevision, recall, F1-score, and area under curve (AUC).

2.3. Bayesian Optimization

This study also implemented Bayesian optimization in order to optimize all six classification models. For each of the classification algorithms, bayesian optimization was run to optimize the parameters. The models were optimized using the validation dataset in order to minimize overfitting when evaluating model metrics with the test dataset.

3. Results

3.1. Machine Learning Model Evaluation

3.1.1 DNA gyrase Machine Learning Model Evaluation

All six optimized machine learning models were trained on DNA gyrase inhibitors and evaluated using accuracy, prevision, recall, F1-score, and AUC. The gradient boosting algorithm performed the best with an accuracy, precision, recall, F1-score, and AUC of 0.91, 0.92, 0.86, 0.88, and 0.933, respectively (**Table 2**, **Figure 1**, **Figure 2**).

Table 2. DNA gyrase Machine Learning Model Accuracy Metrics.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.88	0.88	0.82	0.84	0.919
Support Vector Machine	0.86	0.92	0.74	0.78	0.921
Random Forest	0.87	0.92	0.76	0.80	0.898
K-Nearest Neighbor	0.78	0.74	0.67	0.69	0.754
AdaBoost	0.83	0.79	0.75	0.77	0.920
Gradient Boosting	0.91	0.92	0.86	0.88	0.933

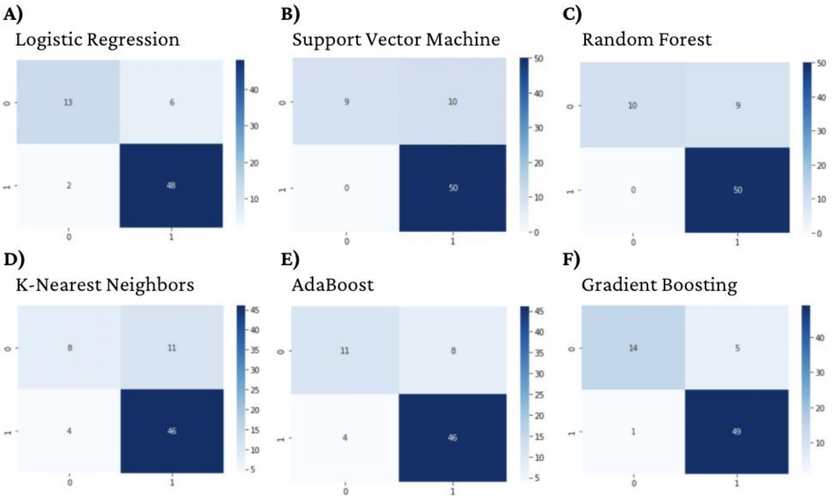

Figure 1. DNA gyrase Machine Learning Model Confusion Matrices.

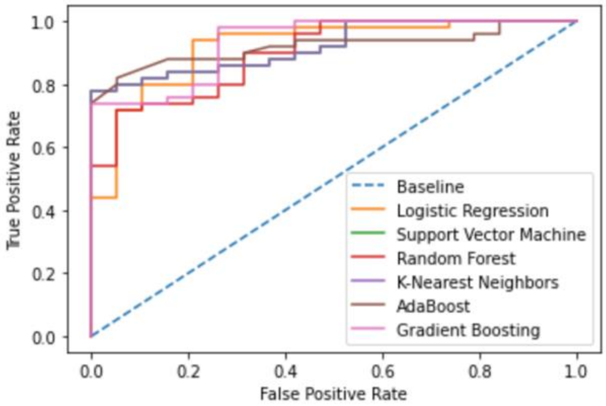

Figure 2. DNA gyrase Machine Learning Model receiver operating characteristic (ROC) curves.

3.1.2 Dihydrofolate reductase Machine Learning Model Evaluation

All six optimized machine learning models were trained on Dihydrofolate reductase inhibitors and evaluated using accuracy, prevision, recall, F1-score, and AUC. The random forest algorithm performed the best with an accuracy, precision, recall, F1-score, and AUC of 0.91, 0.92, 0.86, 0.88, and 0.933, respectively (**Table 3**, **Figure 3**, **Figure 4**).

Table 3. Dihydrofolate reductase Machine Learning Model Accuracy Metrics.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.85	0.82	0.82	0.82	0.949
Support Vector Machine	0.85	0.81	0.83	0.82	0.929
Random Forest	0.86	0.83	0.85	0.84	0.944
K-Nearest Neighbor	0.83	0.81	0.86	0.82	0.926
AdaBoost	0.82	0.78	0.80	0.79	0.866
Gradient Boosting	0.85	0.82	0.82	0.82	0.889

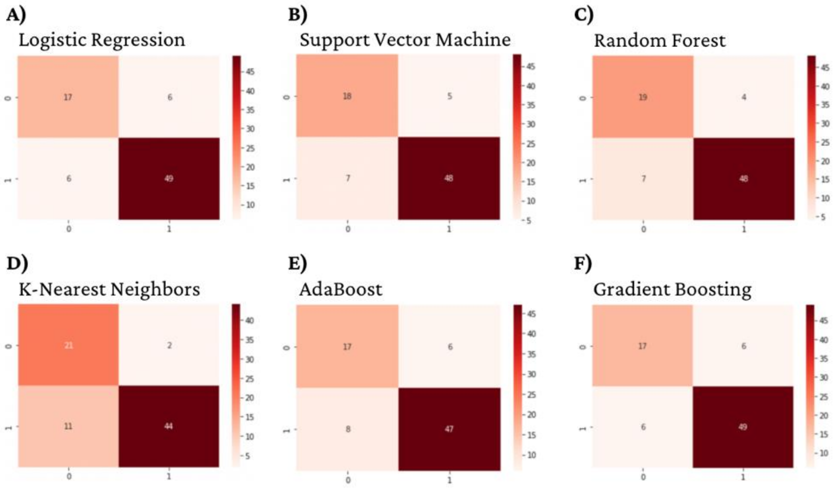

Figure 3. Dihydrofolate reductase Machine Learning Model Confusion Matrices.

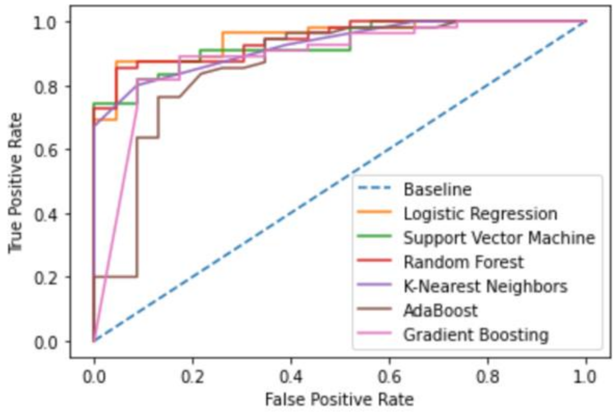

Figure 4. Dihydrofolate reductase Machine Learning Model ROC curves.

3.2. Identification and Analysis of Novel Antimicrobial Ligands

By implementing the best performing model for identifying DNA gyrase inhibitors (gradient boosting) and the best performing model for identifying Dihydrofolate reductase (random forest), this study used each model to identify novel compounds that are predicted to inhibit both DNA gyrase and Dihydrofolate reductase. Using both models, five compounds were identifed that had an average predicted probability greater than 0.97. The best performing compound is, CN(Cc1cnc2nc(N)nc(N)c2n1)c1ccc(C(=O)N[C@@H](CCC(=O)NO)C(=O)O)cc1, with a predicted probability of 0.9988515310206159 to inhibit DNA gyrase, a predicted probability of 0.9897304236200257 to inhibit Dihydrofolate reductase, and an averaged predicted probability of 0.9942909773203208.

Table 4. Novel Antimicrobial Compounds and Probabilistic Analyses.

Compound Formulas and ZINC IDs	Predicted Probability: DNA Gyrase	Predicted Probability: Dihydrofolate reductase	Predicted Probability: Dihydrofolate reductase
$C_{20}H_{23}N_9O_5$ (ZINC27637231)	0.9988515310206159	0.9897304236200257	0.9942909773203208
$C_{25}H_{29}N_9O_8$ (ZINC5385827)	0.9910340430619817	0.9974326059050064	0. 994233324483494
$C_{25}H_{29}N_9O_8$ (ZINC4772545)	0.9995824679977368	0.9858793324775353	0.9927309002376361
$C_{24}H_{27}N_9O_8$ (ZINC5372693)	0.9908400010423145	0.9691912708600771	0.9800156359511958
$C_{25}H_{30}N_{10}O_6$ (ZINC28713649)	0.9993063830032061	0.959349593495935	0.9793279882495705

4. Discussion

Of the six models trained on DNA gyrase, gradient boosting was the most accurate and had the highest F1-score. Unlike the some of the other algorithms, the Bayesian optimization of gradient boosting models drastically changes the model performance and metrics (i.e., number of trees, learning rate, and maximum depth), greatly enhancing its performance. The use of gradient boosting is further enhanced as the DNA gyrase dataset consists of few outliers and overall computation time was not a major constraint [12].

For the Dihydrofolate reductase models, the random forest model nominally outperformed other models. With the optimal models from both DNA gyrase and Dihydrofolate reductase, the compounds identified had a probability of 0.97 to inhibit both DNA gyrase and Dihydrofolate reductase.

Since the precision in both models are either similar or higher than respective recall metrics, the false positivity rate will be comparable to the accuracy metrics of the models, ensuring that the chosen compounds maintain a high probability of being effective.

5. Conclusions

This study evaluated the efficacy of machine learning models at identifying novel antimicrobial compounds. The machine learning models evaluated in this study, particularly the gradient boosting and random forest models, performed very well and hold tremendous potential in computationally screening large libraries of compounds for antimicrobial potential. Furthermore, the compounds identified in this study hold promise as potential, novel antimicrobial compounds. Future investigations should explore alternative classification approaches to antimicrobial compound screening. The compounds identified in this study should also be researched further *in vivo* to identify additional antimicrobial potential.

Author Contributions: Conceptualization, J.S. and D.V.; methodology, J.S. and D.V.; software, J.S. and D.V.; validation, J.S. and D.V.; formal analysis, J.S.; investigation, J.S.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.S. and D.V.; writing—review and editing, J.S. and D.V.; visualization, J.S.; supervision, J.S.; project administration, J.S.

Funding:

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement: Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: The authors would like the acknowledge their friends, families, and mentors for their endless support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ventola C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P & T : a peer-reviewed journal for formulary management*, *40*(4), 277–283.
2. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, *18*(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5
3. Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine Learning in Drug Discovery: A Review. *Artificial intelligence review*, *55*(3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4
4. Kapsiani, S., & Howlin, B. J. (2021). Random forest classification for predicting lifespan-extending chemical compounds. *Scientific reports*, *11*(1), 13812. https://doi.org/10.1038/s41598-021-93070-6
5. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, *180*(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

6. Rahman, M., Browne, J. J., Van Crugten, J., Hasan, M. F., Liu, L., & Barkla, B. J. (2020). *In Silico*, Molecular Docking and *In Vitro* Antimicrobial Activity of the Major Rapeseed Seed Storage Proteins. *Frontiers in pharmacology*, *11*, 1340. https://doi.org/10.3389/fphar.2020.01340
7. Reece, R. J., & Maxwell, A. (1991). DNA gyrase: structure and function. *Critical reviews in biochemistry and molecular biology*, *26*(3-4), 335–375. https://doi.org/10.3109/10409239109114072
8. Maxwell A. (1997). DNA gyrase as a drug target. *Trends in microbiology*, *5*(3), 102–109. https://doi.org/10.1016/S0966-842X(96)10085-8
9. Zhang, Y., Chowdhury, S., Rodrigues, J. V., & Shakhnovich, E. (2021). Development of antibacterial compounds that constrain evolutionary pathways to resistance. *eLife*, *10*, e64518. https://doi.org/10.7554/eLife.64518
10. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic acids research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074
11. Sterling, T., & Irwin, J. J. (2015). ZINC 15--Ligand Discovery for Everyone. *Journal of chemical information and modeling*, *55*(11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559
12. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021