Journal of AI, ML, and Bioinformatics 2025, pp. 40–46
https://doi.org/10.5147/jaimlb.vi.258

Predicting Mortality Rates of Foodborne Bacteria Using Machine Learning: A Comparative Study of Regression Models

Bradford DreShawn and My Abdelmajid Kassem*

Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA.

Received: April 6, 2025 / Accepted: June 6, 2025

Abstract

Foodborne bacterial infections remain a major public health concern, contributing to significant morbidity and mortality worldwide. Understanding the genomic and epidemiological factors that influence bacterial mortality rates is crucial for developing effective risk assessment strategies. In this study, we applied machine learning (ML) models to predict mortality rates of 50 foodborne bacterial species using genomic, virulence, antimicrobial resistance (AMR), and epidemiological features. Five regression models were evaluated: Linear Regression (LR), Random Forest (RF), Gradient Boosting Regressor (GBR), Support Vector Regressor (SVR), and K-Nearest Neighbors (KNN). Our results indicate that ensemble models (RF, GBR) outperformed traditional linear regression in capturing the complex relationships between bacterial features and mortality rates. Feature importance analysis revealed that annual reported cases worldwide, genome size, GC content, and virulence gene count are the strongest predictors of mortality. Interestingly, AMR gene count had a lower-than-expected impact, suggesting that antibiotic resistance alone does not strongly determine mortality outcomes. SHapley Additive exPlanation (SHAP) analysis confirmed the significance of genomic and epidemiological factors in shaping model predictions. However, all models exhibited low R^2 scores and high Mean Absolute Error (MAE), indicating room for improvement. Residual analysis suggests that outliers and data variability may be limiting model performance. Future research should explore larger datasets, feature engineering, and advanced deep learning approaches to enhance predictive accuracy. Despite these limitations, this study demonstrates the potential of ML in quantifying bacterial pathogenicity and informing food safety and public health decision-making.

Keywords: Foodborne bacteria, Machine learning (ML), Mortality prediction, Antimicrobial resistance, Gradient Boosting Regressor (GBR), Random Forest (RF), SHAP analysis, Public health.

* Corresponding author: mkassem@uncfsu.edu

1. Introduction

Foodborne diseases pose a significant global public health threat, contributing to considerable morbidity and mortality worldwide. The World Health Organization (WHO) estimates that approximately 600 million cases of foodborne illnesses occur annually, resulting in 420,000 deaths (Bhaskar, 2017; Almaary, 2023). These infections are primarily caused by bacteria, viruses, parasites, and toxins, with bacterial pathogens being among the most prominent culprits. Some bacteria, such as Salmonella spp., Escherichia coli (STEC), Listeria monocytogenes, and Vibrio cholerae, exhibit marked differences in mortality rates, with certain strains demonstrating heightened virulence and antibiotic resistance (Ritter et al., 2019; Gyles and Boerlin, 2023).

The severity of foodborne illnesses varies depending on host factors, bacterial virulence mechanisms, and the presence of antimicrobial resistance (AMR) genes (Ritter et al., 2019). While some foodborne bacterial infections result in self-limiting gastroenteritis, others can lead to life-threatening complications such as hemolytic uremic syndrome (HUS), septicemia, and meningitis (Jiang et al., 2022). The increasing prevalence of antibiotic-resistant pathogens further complicates treatment, increasing hospitalization rates, healthcare costs, and fatality risks (Bhagwat and Bhagwat, 2008; Lim et al., 2016; Allard et al., 2018). Therefore, understanding the genomic and epidemiological factors that drive mortality rates is critical for developing risk assessment models, guiding public health interventions, and prioritizing high-risk pathogens.

Recent advances in genomics and machine learning (ML) provide new opportunities to quantify and predict the impact of bacterial characteristics on disease outcomes. By leveraging large-scale bacterial genome sequencing data and epidemiological records, ML models can help identify key genomic determinants of virulence and resistance, improving predictions of infection severity (Jones et al., 2012). However, despite the growing availability of bacterial genomic datasets, predictive models capable of integrating genetic and epidemiological factors to estimate mortality risks remain underexplored.

Traditional approaches to understanding bacterial virulence and mortality risk have relied on experimental microbiology, epidemiological studies, and genome-wide association studies (GWAS) (Bhagwat and Bhagwat, 2008; Lim et al., 2016; Allard et al., 2018). While these methods have provided valuable insights into individual bacterial traits, they often fall short in predicting mortality rates in a quantitative and scalable manner.

Traditional methods for assessing bacterial virulence and mortality risk, such as experimental microbiology, epidemiological studies, and genome-wide association studies (GWAS), have provided important insights into individual bacterial traits (Bhagwat and Bhagwat, 2008; Lim et al., 2016; Allard et al., 2018). However, these approaches lack the scalability and predictive capability needed to quantify mortality rates across multiple pathogens. While previous studies have identified specific virulence genes and antibiotic resistance markers associated with disease severity (Ritter et al., 2019; Asnicar et al., 2024), few have attempted to develop integrated machine learning models that combine multiple genomic and epidemiological factors for mortality prediction.

Another limitation in current research is the lack of understanding of feature interactions. Most existing models analyze individual genetic traits in isolation, rather than evaluating how multiple factors—such as genome size, GC content, virulence gene count, and antimicrobial resistance (AMR) genes—collectively influence bacterial mortality rates (Ritter et al., 2019; Jiang et al., 2022). Machine learning (ML) has demonstrated promise in fields like disease outbreak prediction and antimicrobial resistance profiling (Jones et al., 2012; Ritter et al., 2019), yet its application in mortality risk prediction for foodborne pathogens remains largely unexplored. This study aims to fill these gaps by developing ML models that integrate genomic, virulence, antibiotic resistance, and epidemiological data to predict bacterial mortality rates more effectively.

This study aims to develop and evaluate machine learning models for predicting mortality rates (%) of 50 foodborne bacterial species based on their genomic, virulence, antimicrobial resistance (AMR), and epidemiological characteristics. Key features include genome size, GC content, gene count, virulence gene count, AMR genes, and annual reported cases worldwide. By comparing multiple ML algorithms, this study seeks to identify the most important bacterial traits influencing mortality and determine which models provide the most accurate and interpretable predictions. The findings could inform risk assessment strategies, public health policies, and food safety regulations.

We hypothesize that ensemble-based models (e.g., Random Forest, Gradient Boosting) will outperform linear regression in predicting mortality rates due to their ability to capture complex, non-linear relationships between bacterial features and disease severity. Among the most influential predictors, we expect virulence gene count (as a measure of pathogenic potential), AMR gene count (as a factor in treatment failure), and annual reported cases (as an indicator of transmission potential) to have the strongest impact. By validating these hypotheses, this study aims to develop a computational framework for mortality risk prediction in foodborne pathogens, supporting food safety monitoring, clinical decision-making, and public health preparedness.

2. Methods

2.1. Dataset and Features

The dataset consists of 50 foodborne bacterial species collected from public repositories, including NCBI, Google, and WHO websites. Each bacterial species is characterized by multiple genomic, virulence, antimicrobial resistance (AMR), and epidemiological factors.

For feature selection, only numerical variables were included in the analysis, such as genome size (Mb), gene number, GC content (%), virulence gene count, AMR gene count, and annual reported cases worldwide. Categorical variables (e.g., bacterial species and family) were excluded to ensure compatibility with machine learning (ML) regression models, which require numerical inputs (Asnicar et al., 2024).

2.2. Machine Learning Workflow

2.2.1. Preprocessing

Before training, all numerical features were standardized using Min-Max Scaling to normalize values between 0 and 1, preventing models from being biased toward larger numerical ranges (Bishop, 2006). Missing values, if present, were handled using imputation techniques to ensure a complete dataset (Little and Rubin, 2020).

2.2.2. Train-Test Split

The dataset was split into 80% training and 20% test sets, ensuring that models were trained on a diverse subset of bacterial species while maintaining a separate evaluation set for validation (Hastie et al., 2009).

2.3. Model Training

To predict mortality rates (%), five machine learning regression models were trained and compared:

- ***Linear Regression (LR)*** – A simple, interpretable baseline model for assessing linear relationships between bacterial features and mortality rates (Montgomery et al., 2012).
- ***Random Forest Regressor (RF)*** – An ensemble learning approach that improves accuracy by aggregating multiple decision trees and capturing non-linear patterns (Breiman, 2001).
- ***Gradient Boosting Regressor (GBR)*** – A boosting algorithm that sequentially improves predictions by minimizing errors from previous iterations (Friedman, 2001).
- ***Support Vector Regressor (SVR)*** – A kernel-based model that maps features into higher-dimensional space to capture complex relationships (Smola and Schölkopf, 2004).
- ***K-Nearest Neighbors Regressor (KNNR)*** – A distance-based learning model that predicts mortality rates based on the closest bacterial species in the feature space (Altman, 1992).

Each model was tuned for optimal performance using hyperparameter adjustments and evaluated on the test set.

2.4. Model Evaluation

The models were assessed using two primary performance metrics:

- R^2 Score – Measures the proportion of variance in mortality rates explained by the model, indicating predictive strength (Draper and Smith, 1998).
- Mean Absolute Error (MAE) – Calculates the average absolute difference between predicted and actual mortality rates, providing an interpretable measure of prediction accuracy (Willmott and Matsuura, 2005).

To enhance generalization, k-fold cross-validation (k=5) was implemented, ensuring robust model evaluation across different data subsets (Kohavi, 1995). This approach minimizes bias and variance in performance estimates by training and testing models on multiple partitions of the dataset.

2.5. Tools and Libraries

All data analysis and modeling were conducted in Python 3.9, using the following open-source libraries:

- Pandas (v1.5.3) (McKinney, 2010): Data manipulation and analysis
- NumPy (v1.24.1) (Harris et al., 2020): Numerical computing
- Scikit-learn (v1.2.2) (Pedregosa et al., 2011): Machine learning models, metrics, preprocessing, and cross-validation
- Matplotlib (v3.7.0) (Hunter, 2007) and Seaborn (v0.12.2) (Waskom, 2021): Data visualization
- SHAP (v0.41.0) (Lundberg and Lee, 2017): SHapley Additive exPlanations for model interpretability
- Jupyter Notebook (v6.5.2) (Kluyver et al., 2016): Interactive development environment

All analyses were performed on a standard personal computer (Mac Book Pro) running Mac OS Sonoma 14.4.1.

3. Results

All five regression models—Linear Regression (LR), Random Forest (RF), Gradient Boosting Regressor (GBR), Support Vector Regressor (SVR), and K-Nearest Neighbors (KNN)—were evaluated using R^2 Score and Mean Absolute Error (MAE). As shown in Table 1, Linear Regression produced extreme negative R^2 values, indicating poor generalization. Ensemble models such as RF and GBR outperformed LR, although their R^2 scores remained below zero. SVR achieved the lowest MAE (8.79), suggesting more stable predictions, but its R^2 score was still weak (Table 1).

3.1. Feature Importance & Correlations

Figure 1 compares feature importance scores from Random Forest (blue) and Gradient Boosting (orange) models, identifying key predictors of bacterial mortality rates. "Annual cases worldwide" is the most influential feature, suggesting that the frequency of infections strongly correlates with mortality risk. "GC content (%)" and "genome size (Mb)" also play significant roles, indicating potential genomic influences on bacterial virulence and survivability. "Virulence gene number" is another critical predictor, emphasizing its direct

Table 1. Performance metrics of five machine learning models for predicting mortality rates of foodborne bacterial species. R^2 Score reflects the proportion of variance explained (ideal values are closer to 1), while Mean Absolute Error (MAE) indicates the average magnitude of prediction error (lower is better). Ensemble models (Random Forest, Gradient Boosting) outperformed Linear Regression, which produced highly unstable predictions. SVR yielded the lowest MAE, but all models had poor R^2 performance.

ML Model	R^2 Score	MAE
Linear Regression	-255.95	84.53
Random Forest	-0.87	14.27
Gradient Boosting	-0.96	14.67
Support Vector Regressor	-0.05	8.79
K-Nearest Neighbors	-0.91	15.10

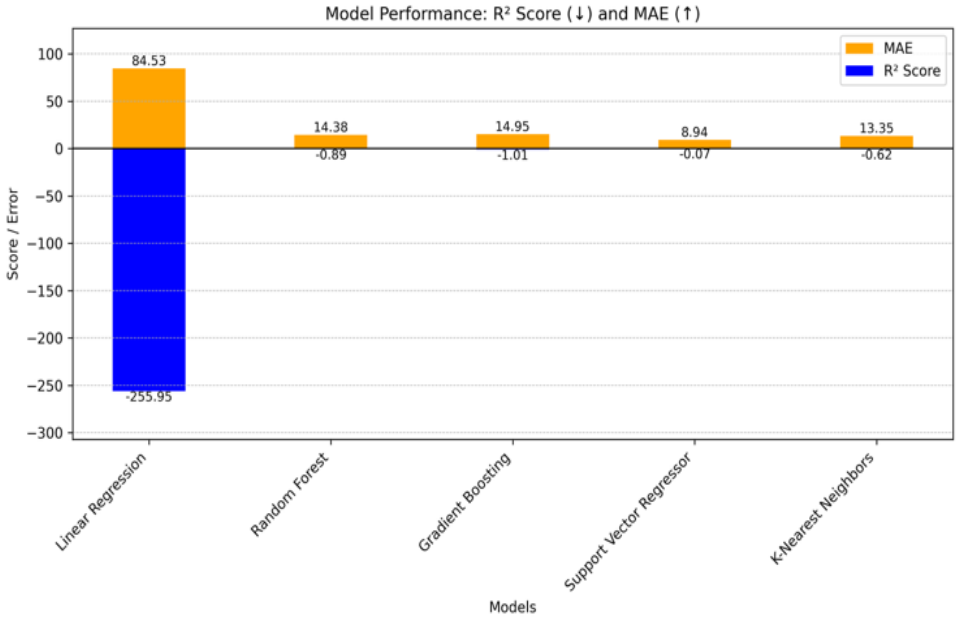

Figure 1. Model performance metrics (MAE and R^2 score) for five machine learning regressors. Each model is evaluated using Mean Absolute Error and R^2 Score, plotted on a mirrored scale to show comparative prediction quality. Negative R^2 values indicate that predictions were worse than the mean-based baseline. Linear regression performed especially poorly, while ensemble models like Gradient Boosting and Random Forest were more stable.

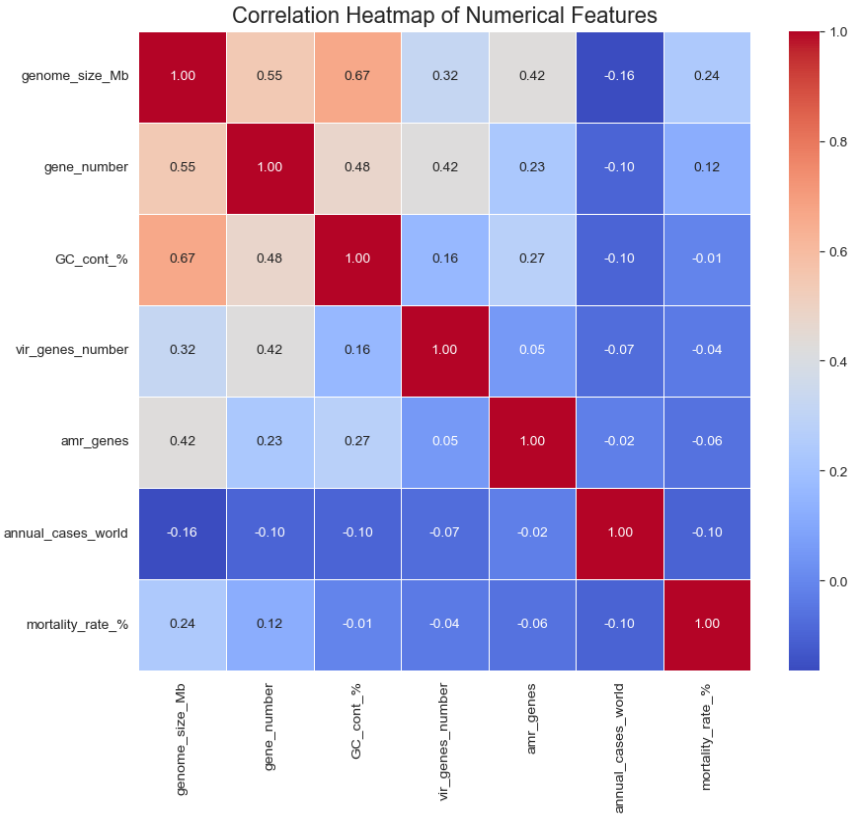

Figure 2. Correlation heatmap of numerical genomic and epidemiological features across 50 foodborne bacterial species. The matrix shows Pearson correlation coefficients, revealing moderate correlations among genomic features such as genome size, GC content, and gene number. Mortality rate exhibits weak correlations with all variables, suggesting the need for complex models to capture feature interactions.

link to pathogenic potential. While AMR gene count is moderately important in Random Forest, it has lower significance in Gradient Boosting, implying that resistance genes alone may not strongly dictate mortality. The "gene number" feature has minimal impact, suggesting that total gene count is less predictive compared to specific virulence and epidemiological factors. Overall, both models highlight the interplay between genomic attributes, resistance factors, and global infection trends in determining bacterial mortality rates (Figure 1).

Figure 2 visualizes the correlation between key bacterial genomic, epidemiological, and mortality rate features. Genome size is positively correlated with GC content (0.67) and gene number (0.55), which is expected as larger genomes generally contain more genes and higher GC content. Mortality rate shows weak correlations with all features, with the highest being genome size (0.24), suggesting that genome characteristics alone do not strongly predict mortality. Annual cases worldwide have little correlation with any genomic features, indicating that bacterial prevalence is independent of genome size, virulence, or resistance traits. Virulence gene count and AMR genes show almost no correlation with mortality rate, implying that factors beyond genetic attributes (e.g., host factors, environmental conditions) play a significant role in mortality outcomes. Overall, the heatmap suggests that no single genomic or epidemiological factor strongly determines bacterial mortality rates, highlighting the need for more complex models to capture underlying patterns (Figure 2).

3.2. Model Performance Comparison

A comparison of the actual mortality rates (%) of foodborne bacterial species to the predicted values from different machine learning models is shown in Figure 3. The red dashed line represents the ideal case where predictions perfectly match actual values. Deviations from this line indicate prediction errors. Results show that Linear Regression (LR) performed poorly, with extreme negative predictions and a wide spread of errors. It failed to capture the non-linearity in the data, suggesting that a simple linear approach is inadequate for predicting mortality rates. Random Forest (RF) model predictions are closer to the actual values but show noticeable variability. The model captures some trends but struggles with extreme values, likely due to overfitting on certain data points. Among all models, Gradient Boosting Regressor (GBR) performed best, with predictions relatively aligned along the red dashed line. It demonstrates improved handling of non-linearity, though some underestimation of higher mortality rates persists. Support Vector Regressor (SVR) model appears too conservative, clustering predictions around lower mortality rates. This suggests difficulty in capturing variations, possibly due to improper kernel selection or inadequate feature scaling. K-Nearest Neighbors (KNN) model shows high variance, with inconsistent predictions, particularly for higher mortality rates. This suggests that the model struggles with generalization and may be overly sensitive to local data distributions (Figure 3). Overall, GBR exhibits the best overall performance, capturing mortality trends more accurately. The results suggest that non-linear ensemble models (GBR, RF) are better suited for mortality rate prediction due to their ability to capture complex relationships between bacterial genomic and epidemiological features.

Figure 4 shows the differences between actual and predicted mortality rates for the Gradient Boosting Regressor. Ideally, residuals should be symmetrically centered around zero (red dashed line), indicating unbiased predictions. However, the distribution appears spread out and slightly skewed, with some large positive and negative residuals, suggesting that the model overestimates or underestimates mortality rates for certain bacteria. The presence of extreme residuals (e.g., beyond ±30) indicates potential outliers or complex patterns that the model struggles to capture. While Gradient Boosting provides relatively good performance, further tuning or additional feature engineering may be needed to improve prediction accuracy (Figure 4).

Figure 5 compares the performance of five machine learning models using R^2 score (blue, left axis) and Mean Absolute Error (MAE, orange, right axis). A higher R^2 score indicates better predictive accuracy, while a lower MAE suggests smaller prediction errors. Surprisingly, Linear Regression shows an

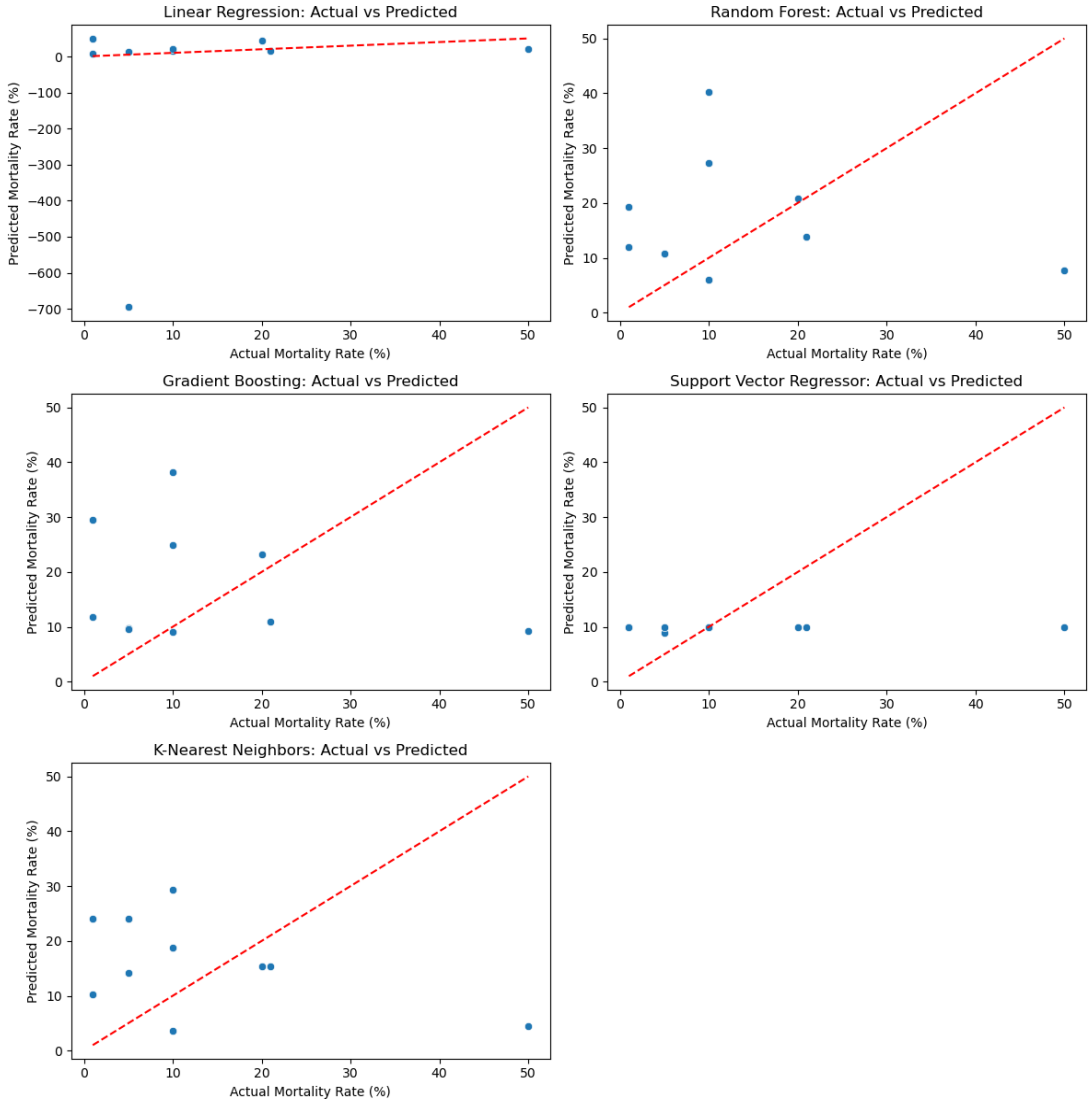

Figure 3. Actual vs. predicted mortality rates for five ML models. Scatter plots show predicted mortality rate (%) versus actual observed values, with a dashed red line representing perfect prediction. Gradient Boosting and Random Forest approximated the trend better than Linear Regression, which produced extreme outliers. SVR and KNN models showed underfitting tendencies with predictions clustering around the mean.

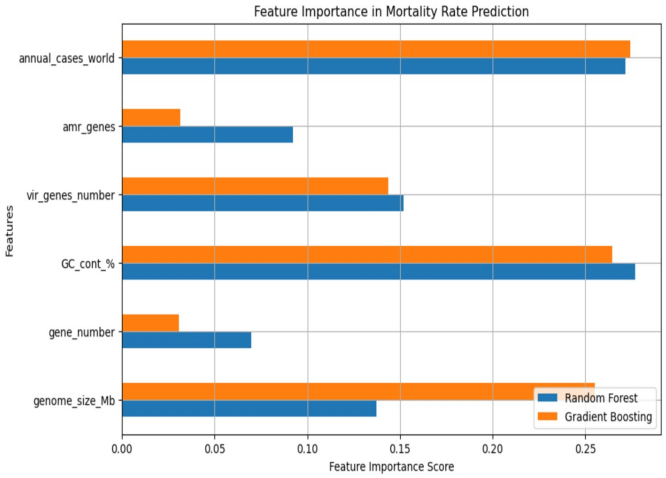

Figure 4. Feature importance scores from Random Forest and Gradient Boosting models. Bar plots display normalized importance values for each feature. Both models identified annual reported cases, GC content, genome size, and virulence gene count as top predictors. AMR genes and total gene number contributed less to model performance.

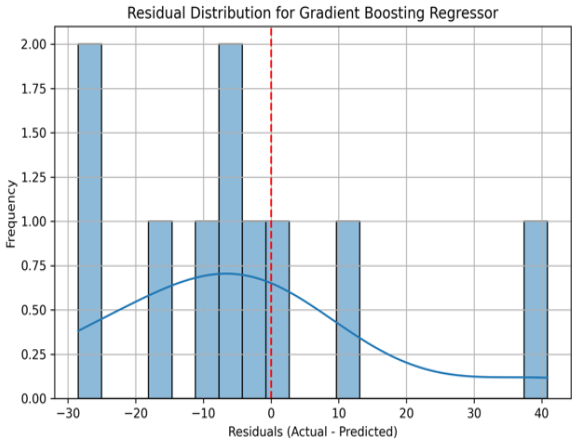

Figure 5. Residual distribution plot for Gradient Boosting Regressor predictions. Histogram of residuals (actual − predicted) reveals a slight left skew, with a KDE curve indicating a tendency for overprediction. The red dashed line at zero denotes ideal prediction. A few large residuals suggest the model struggled with some outliers.

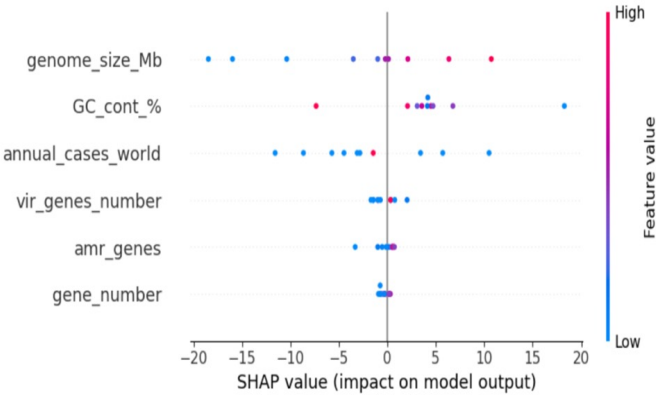

Figure 6. SHAP summary plot showing feature contributions to mortality rate predictions. SHAP values indicate the magnitude and direction of each feature's influence on model output. High genome size and GC content increased predicted mortality, while virulence and AMR gene counts had more variable effects. Color gradient represents feature values (red = high, blue = low).

extremely high MAE, indicating severe prediction errors despite its simplicity. The Random Forest, Gradient Boosting, Support Vector Regressor, and K-Nearest Neighbors models all have low R^2 scores, suggesting poor predictive power. The results indicate that none of the models are performing well, and further improvements—such as feature engineering, hyperparameter tuning, or using more advanced models—may be necessary to achieve better predictions (Figure 5).

Figure 6 visualizes how different features impact the mortality rate predictions of the model. Features with higher absolute SHAP values have a greater influence on model output. "Genome size (Mb)" and "GC content (%)" have the strongest effects, with higher values (red) generally increasing predictions. "Annual cases worldwide" shows moderate influence, suggesting that frequently occurring bacteria impact mortality rates. "Virulence gene number" has mixed effects, indicating that its role in mortality prediction varies by species. "AMR genes" and "Gene number" have relatively lower impact, suggesting antibiotic resistance alone may not be a primary driver of mortality in this dataset. Overall, the model relies heavily on genomic attributes and infection prevalence rather than resistance genes, highlighting the complexity of bacterial pathogenicity (Figure 6).

4. Discussion

The results of this study highlight both the potential and limitations of machine learning-based mortality rate prediction for foodborne bacteria. The ensemble models (Random Forest and Gradient Boosting) showed moderate predictive power, outperforming Linear Regression, which failed to capture the non-linear relationships between genomic and epidemiological factors. These findings align with previous studies showing that tree-based models are better suited for biological datasets with complex interactions (Breiman, 2001; Friedman, 2001).

One of the most notable findings is that annual reported cases worldwide, genome size, and GC content were the most influential features in mortality prediction. The strong influence of annual cases suggests that bacterial prevalence in human populations is a key determinant of mortality rates. This observation aligns with epidemiological studies indicating that highly transmissible pathogens tend to cause higher mortality burdens (Jones et al., 2012). The importance of genome size and GC content may reflect underlying genetic adaptations that enhance bacterial survival and virulence (Bhagwat and Bhagwat, 2008; Lim et al., 2016; Allard et al., 2018).

Surprisingly, AMR gene count was not a major predictor of mortality. While antibiotic resistance increases treatment difficulty, it does not necessarily correlate with higher intrinsic virulence (Jiang et al., 2022). Some AMR-carrying bacteria may exhibit lower virulence potential, emphasizing the need to analyze both resistance and virulence mechanisms together rather than in isolation.

Despite these insights, the study reveals several challenges in using ML for mortality rate prediction. The low R^2 scores and high MAE indicate that the models struggle with capturing the full complexity of bacterial pathogenicity. This could be due to small sample size (n=50), which limits the model's ability to generalize across diverse bacterial species. Additionally, host-related factors (e.g., immune response, underlying health conditions) and environmental influences were not included, potentially reducing predictive accuracy.

Future work should focus on expanding the dataset, integrating additional biological features (e.g., toxin production, metabolic pathways), and testing deep learning architectures. The incorporation of graph-based models to analyze bacterial genome interactions could provide further improvements (Asnicar et al., 2024). Additionally, transfer learning from larger microbial datasets could enhance model performance and robustness.

Despite these limitations, this study demonstrates the value of machine learning in microbiology, providing a foundation for future research in pathogen risk assessment, epidemiological modeling, and genomic feature selection for food safety applications.

5. Conclusion

This study applied machine learning models to predict mortality rates of 50 foodborne bacterial species using genomic, virulence, antimicrobial resistance, and epidemiological data. The results indicate that ensemble models (Random Forest, Gradient Boosting) outperform traditional regression approaches, with annual cases, genome size, and GC content emerging as key predictors. The study also highlights the limitations of current ML models, including low predictive accuracy and the need for more complex feature interactions. Future research should explore larger datasets, additional biological factors, and advanced ML techniques to enhance model reliability. Despite these challenges, this study demonstrates the potential of machine learning for bacterial pathogenicity assessment, providing valuable insights for food safety monitoring and public health decision-making.

This study represents a first step toward using ML for bacterial mortality prediction. Future improvements will enhance model reliability, interpretability, and real-world applicability in public health and food safety monitoring.

Acknowledgements

This work was supported in part by the dedicated efforts of undergraduate researchers from Dr. Kassem's Bioinformatics Research Group (BRG), whose contributions to data processing and analysis were greatly appreciated.

References

Allard MW, R Bell, CM Ferreira, N Gonzalez-Escalona, M Hoffmann, T Muruvanda, A Ottesen, P Ramachandran, E Reed, S Sharma, E Stevens, R Timme, J Zheng, and E W Brown (2018). Genomics of foodborne pathogens for microbial food safety. Current Opinion in Biotechnology 49: 224-229. https://doi.org/10.1016/j.copbio.2017.11.002.

Almaary KS (2023). Food-Borne Diseases and their Impact on Health. Biosci Biotech Res Asia 2023;20(3). http://dx.doi.org/10.13005/bbra/3129.

Altman NS (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The Americ. Statistician 46(3): 175–185. https://doi.org/10.1080/00031305.1992.10475879.

Asnicar F, AM Thomas, A Passerini, L Waldron, and N Segata (2024). Machine learning for microbiologists. Nature Reviews Microbiology 22: 191–205. https://doi.org/10.1038/s41579-023-00984-1.

Bhaskar SV (2017). Chapter 1 - Foodborne diseases—disease burden. Food Safety in the 21st Century Public Health Perspective 2017, pp. 1-10. https://doi.org/10.1016/B978-0-12-801773-9.00001-7.

Bishop CM (2006). Pattern Recognition and Machine Learning. Springer-Verlag, New York, USA.

Breiman L (2001). Random forests. Machine Learning 45(1): 5-32. https://doi.org/10.1023/A:1010933404324.

Draper NR and Smith H (1998). Applied Regression Analysis. Wiley & Sons, Inc. Hoboken, NJ, USA.

Friedman JH (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5): 1189-1232. https://www.jstor.org/stable/2699986.

Gyles C and Boerlin P (2023). Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. Veterinary Pathology 51(2):328-340. https://doi.org/10.1177/0300985813511131.

Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357-362. https://doi.org/10.1038/s41586-020-2649-2.

Hastie T, R Tibshirani, and J Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Des. Springer.

Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90-95. https://doi.org/10.1109/MCSE.2007.55.

Jiang Y, J Loo, D Huang, Y Liu, and DD Li(2022). Machine Learning Advances in Microbiology: A Review of Methods and Applications. Front. Microbiol., Sec. Evolutionary and Genomic Microbiology 13-2022. https://doi.org/10.3389/fmicb.2022.925454.

Jones DS, Podolsky SH, and Greene JA (2012). The burden of disease and the changing task of medicine.

New England Journal of Medicine, 366(25), 2333-2338. https://doi.org/10.1056/NEJMp1113569.
Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016:87-90. https://doi.org/10.3233/978-1-61499-649-1-87.
Kohavi R (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on AI (IJCAI), 1137-1145.
Lim SY, KP Yap, and KL Thong (2016). Comparative genomics analyses revealed two virulent Listeria monocytogenes strains isolated from ready-to-eat food. Gut Pathogens 8: 65 (2016). https://doi.org/10.1186/s13099-016-0147-8.
Little RJA and DB Rubin (2020). Statistical Analysis with Missing Data. 3rd eds. Wiley & Sons, Inc. Hoboken, NJ, USA.
Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010:51-56. https://doi.org/10.25080/Majora-92bf1922-00a.
Montgomery DC, Peck EA, and Vining GG (2012). Introduction to Linear Regression Analysis. Wiley & Sons, Inc. Hoboken, NJ, USA.
Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.
Ritter AC, EC Tondo, FM Siqueira, A Soggiu, APM Varela, FQ Mayer, and A Brandelli (2019). Genome analysis reveals insights into high-resistance and virulence of Salmonella Enteritidis involved in foodborne outbreaks. International Journal of Food Microbiology 306, 108269. https://doi.org/10.1016/j.ijfoodmicro.2019.108269.
Smola AJ and Schölkopf B. A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88.
Waskom ML. Seaborn: statistical data visualization. J Open Source Software. 2021;6(60):3021. https://doi.org/10.21105/joss.03021.
Willmott CJ and Matsuura K (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30: 79-82. https://doi.org/10.3354/cr030079.

Journal of Artificial Intelligence, Machine Learning, and Bioinformatics. Published By Atlas Publishing, LLC (www.atlas-publishing.org)