World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:19, No:03, 2025

Comparison of Machine Learning-Based Models for Predicting Streptococcus pyogenes Virulence Factors and Antimicrobial Resistance

Authors : Fernanda Bravo Cornejo, Camilo Cerda Sarabia, Belén Díaz Díaz, Diego Santibañez Oyarce, Esteban Gómez Terán, Hugo Osses Prado, Raúl Caulier-Cisterna, Jorge Vergara-Quezada, Ana Moya-Beltrán

Abstract : Streptococcus pyogenes is a gram-positive bacteria involved in a wide range of diseases and is a major-human-specific bacterial pathogen. In Chile, this year the 'Ministerio de Salud' declared an alert due to the increase in strains throughout the year. This increase can be attributed to the multitude of factors including antimicrobial resistance (AMR) and Virulence Factors (VF). Understanding these VF and AMR is crucial for developing effective strategies and improving public health responses. Moreover, experimental identification and characterization of these pathogenic mechanisms are labor-intensive and time-consuming. Therefore, new computational methods are required to provide robust techniques for accelerating this identification. Advances in Machine Learning (ML) algorithms represent the opportunity to refine and accelerate the discovery of VF associated with Streptococcus pyogenes. In this work, we evaluate the accuracy of various machine learning models in predicting the virulence factors and antimicrobial resistance of Streptococcus pyogenes, with the objective of providing new methods for identifying the pathogenic mechanisms of this organism.Our comprehensive approach involved the download of 32,798 genbank files of S. pyogenes from NCBI dataset, coupled with the incorporation of data from Virulence Factor Database (VFDB) and Antibiotic Resistance Database (CARD) which contains sequences of AMR gene sequence and resistance profiles. These datasets provided labeled examples of both virulent and non-virulent genes, enabling a robust foundation for feature extraction and model training. We employed preprocessing, characterization and feature extraction techniques on primary nucleotide/amino acid sequences and selected the optimal more for model training. The feature set was constructed using sequence-based descriptors (e.g., k-mers and One-hot encoding), and functional annotations based on database prediction. The ML models compared are logistic regression, decision trees, support vector machines, neural networks among others. The results of this work show some differences in accuracy between the algorithms, these differences allow us to identify different aspects that represent unique opportunities for a more precise and efficient characterization and identification of VF and AMR. This comparative analysis underscores the value of integrating machine learning techniques in predicting S. pyogenes virulence and AMR, offering potential pathways for more effective diagnostic and therapeutic strategies. Future work will focus on incorporating additional omics data, such as transcriptomics, and exploring advanced deep learning models to further enhance predictive capabilities.

Keywords : antibiotic resistance, streptococcus pyogenes, virulence factors., machine learning
Conference Title : ICMCSSE 2025 : International Conference on Mathematical, Computational and Statistical Sciences and Engineering
Conference Location : Madrid, Spain
Conference Dates : March 17-18, 2025