

Artificial Intelligence in Laboratory Health Care: Machine Learning in Bacterial Identification

Likit Preeyanon^{1*}, Charlchalerm Isarankura Na Aayudha², Prachayasittikul, V³.,
Natasenamat, C⁴.

^{1,2,3,4}Faculty of Medical Technology, Mahidol University-Thailand.

*Corresponding author email: likit.pre@mahidol.ac.th

ABSTRACT

Accurately identifying bacterial species is essential in diagnostic microbiology, where biochemical testing remains a primary method for classification. These tests assess how microorganisms metabolize specific substrates, producing measurable reactions that indicate species-level traits. Although automated systems have improved interpretation through probability-based scoring, misidentification still occurs due to the natural variability of bacterial phenotypes. To address this challenge, the study introduces PyBact, an open-source Python-based software designed to simulate realistic bacterial biochemical profiles for research and educational purposes. PyBact generates binary datasets representing biochemical test results by using probability values derived from authoritative microbiology references, including the Manual of Clinical Microbiology. For each species, the program creates a user-defined number of simulated strains and assigns positive or negative test outcomes based on known frequencies. This probabilistic method faithfully reflects biological diversity rather than relying on fixed, idealized profiles. The resulting data matrices are compatible with machine learning workflows. To evaluate the tool, researchers simulated 100 strains for each species in two datasets: 12 *Vibrio* species (1,200 strains) and 134 Enterobacteriaceae species (13,400 strains). These datasets were analyzed using Weka with three machine learning models—Decision Tree (J48), Naive Bayes, and SVM—under 10-fold cross-validation. The classifiers performed exceptionally well, achieving up to 100% accuracy for *Vibrio* and above 91% for Enterobacteriaceae. Overall, the study demonstrates that PyBact effectively produces biologically meaningful simulated datasets that support highly accurate machine learning-based bacterial identification. The software also serves as a valuable educational resource and is freely available under the Open Software License.

Keywords: PyBact, biochemical simulation, machine learning, bacterial identification, Python software