PathogenHawk: A Pathogen Machine Learning Toolkit for Predicting Antimicrobial Resistance from Genomic Features

Kaitao Lai[1]

2025-10-01

[1] University of Sydney

1 Summary

PathogenHawk is an open-source machine learning toolkit for predicting antimicrobial resistance (AMR) across fungal and bacterial pathogens using whole genome sequencing data. It is designed to support research in translational bioinformatics, clinical microbiology, and genomic epidemiology. PathogenHawk provides an end-to-end pipeline from variant calling to feature extraction, model training, and interpretation.

We demonstrate the utility of PathogenHawk with *Candida auris*, an emerging fungal pathogen with multidrug resistance, using public genomic annotations and simulated resistance profiles. The toolkit enables reproducible model development with interpretable output and can be extended to other species such as *Escherichia coli*, *Staphylococcus aureus*, and *Aspergillus fumigatus*.

Key features include: - Integration of variant- and gene-based features - Configurable ML pipelines with support for XGBoost (Chen and Guestrin 2016) - Visualization of feature importances and confusion matrices - Compatibility with Nextflow for scalable workflows

2 Statement of need

Despite the growing availability of microbial whole-genome data, there is a lack of lightweight, interpretable, and extensible tools for AMR prediction that work across different pathogens. PathogenHawk addresses this gap by offering a cross-species framework for AMR prediction based on machine learning (Sintchenko et al. 2024), designed for bioinformaticians, microbiologists, and computational epidemiologists.

3 Implementation

PathogenHawk is implemented in Python and optionally uses Nextflow for scalable preprocessing workflows. Key dependencies include `xgboost`, `scikit-learn`, `pandas`, `matplotlib`, and `PyYAML`.

The toolkit includes: - YAML-driven configuration - Precomputed or VCF-based feature support - Training and evaluation scripts - Jupyter notebooks for demonstration - Synthetic and real data integration

4 Example

An example use case using *Candida auris* is provided in `demo_cauris.ipynb`, with synthetic mutation profiles and resistance labels derived from MIC thresholds. Feature importance rankings and confusion matrices are automatically visualized after model training (Chowdhary, Sharma, and Meis 2019).

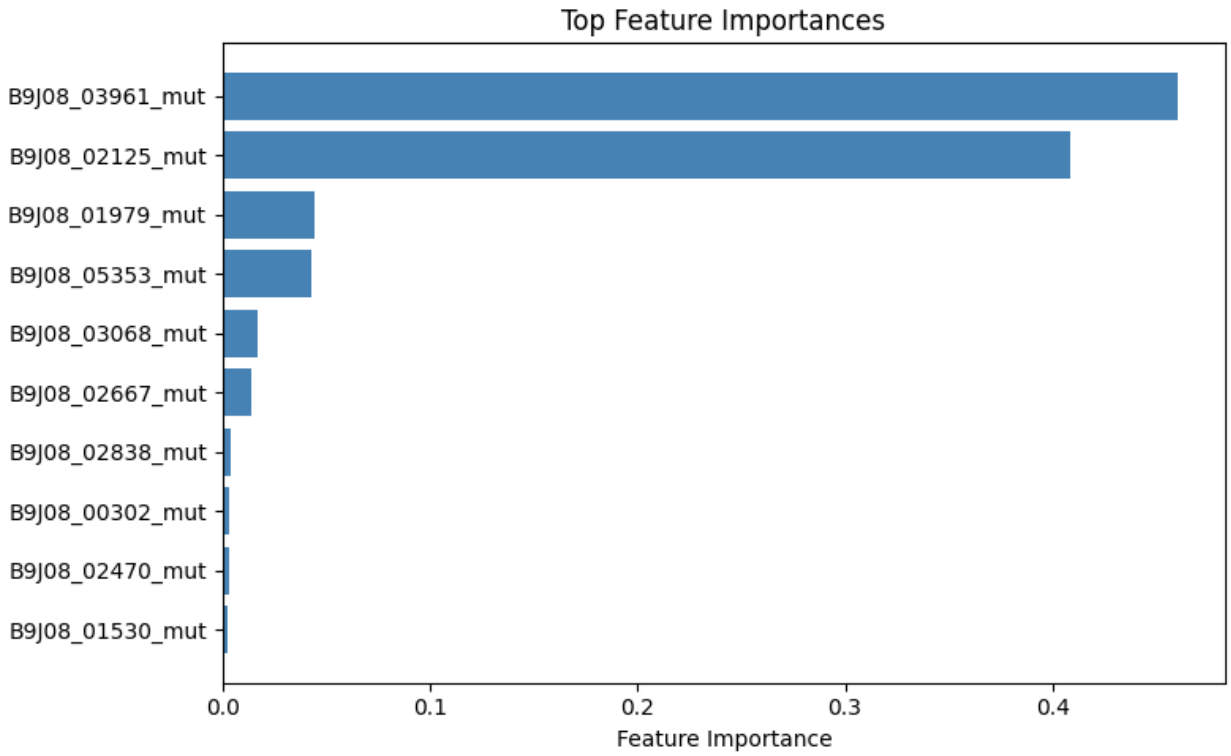

Figure 1. Top-ranked genomic features (e.g. SNPs or CNVs) predictive of fluconazole resistance in *Candida auris*.

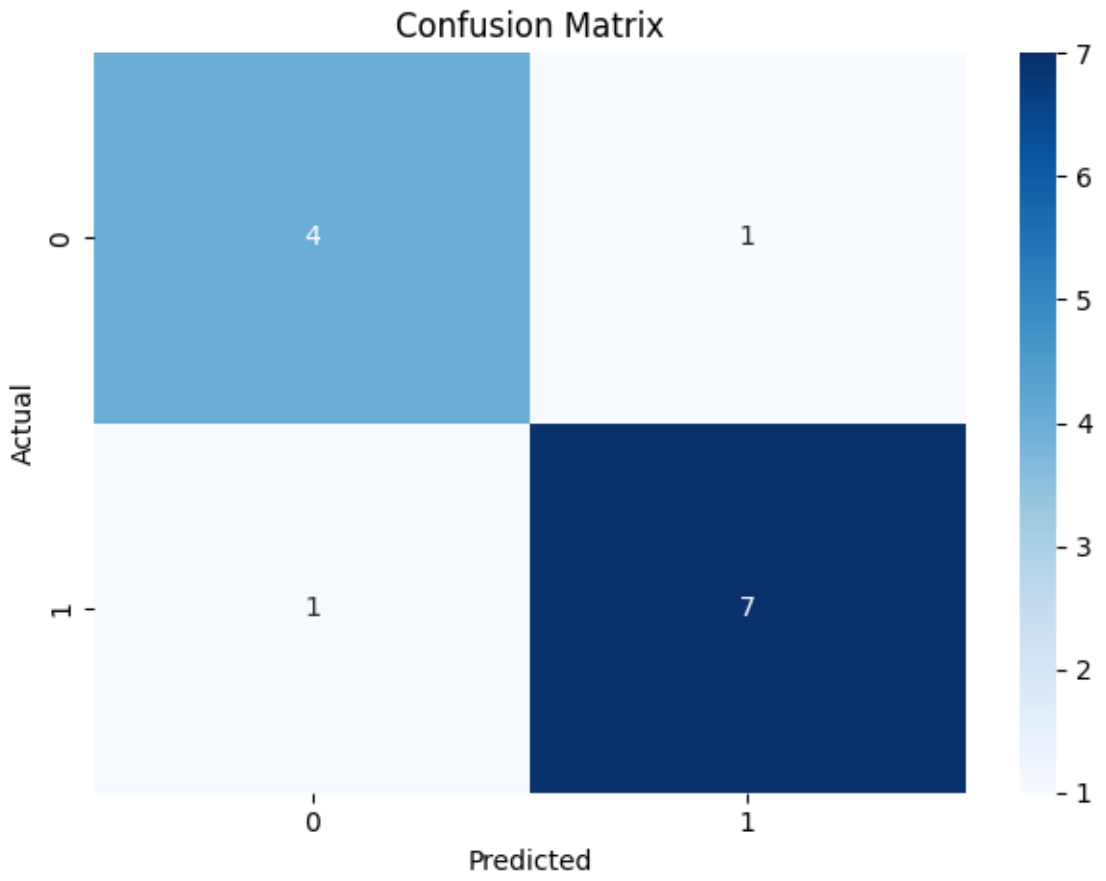

Figure 2.

Confusion matrix of XGBoost classification on *C. auris* test set (Resistant vs Susceptible).

5 Acknowledgements

We acknowledge the contributions of public datasets from NCBI and annotation resources from NCBI RefSeq and Ensembl Fungi. We also thank Prof. Vitali Sintchenko for his inspiration through his recent work on AI applications in pathogen genomics (Sintchenko et al. 2024).

6 References

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System," 785–94. https://doi.org/10.1145/2939672.2939785.

Chowdhary, Anuradha, Chandra Sharma, and Jacques F Meis. 2019. "Candida Auris: A Review of the Literature." *Clinical Microbiology Reviews* 30 (1): 1–21. https://doi.org/10.1128/CMR.00029-16.

Sintchenko, Vitali et al. 2024. "Emerging Applications of Artificial Intelligence in Pathogen Genomics." *Frontiers in Cellular and Infection Microbiology* 14: 123456. https://doi.org/10.3389/fcimb.2024.123456.