Msc Project Proposal

Interpretable Deep-Learning and Ensemble Models for Predicting Multidrug Resistance in *Klebsiella pneumoniae*

Antimicrobial-resistant **Klebsiella pneumoniae** is a WHO "critical" threat, driven by the rapid spread of carbapenemases, ESBLs and quinolone-target mutations. This project will develop an interpretable, genome-based pipeline that predicts phenotypic resistance to four key drug classes-carbapenems, third and fourth-generation cephalosporins, fluoroquinolones and

aminoglycosides-using public sequencing data.

Data. A balanced training set of 1000 short-read genomes collected before 31 Dec 2023, plus a temporal hold-out of 100 genomes from 2024-25, will be downloaded from NCBI Pathogen Detection. Reads will be trimmed, down-sampled to 100×, screened for contamination and assembled. Gene presence and absence and core-genome SNPs will be extracted for classical models, raw reads will be tokenised for deep learning.

Methods. Four classifiers will be benchmarked: (1) XGBoost, (2) LightGBM, (3) a 1-D convolutional neural network on k-mer spectra and (4) a sequence-language deep model fine-tuned on 250 bp read tokens. Five-fold nested cross-validation guides hyper-parameter search. Training runs on a MacBook Pro (M3 Pro, 18Gb RAM) for tree models and a short-term RTX A4000 GPU instance for deep learning.

Research questions:

1. How do the four models compare in F1 and balanced accuracy for each antibiotic?

2. Do pre-2024 models retain ≥0.85 F1 on 2024-25 isolates, and what drift occurs? 3. Which genes, SNPs or sequence motifs drive predictions, and how do these align with known resistance mechanisms?

Success criterion. Per-drug F1 ≥0.85; if balanced accuracy meets 0.85 but F1 does not, the discrepancy will be analysed.

Outputs. A containerised Snakemake workflow, performance benchmark, and motif-level interpretability report will be released under the MIT licence on Github, providing a reproducible template for future AMR-genomics studies.