UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Random subsampling techniques for sea bass mortality prediction

Giovanni Gaio, Simone Moretti

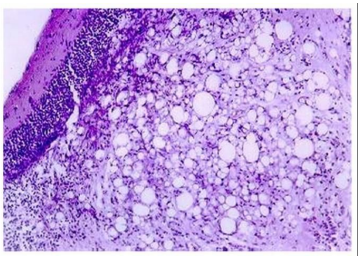
July 23, 2025

Overview

- Motivation: Identifying impactful SNPs in sea bass mortality
- Dataset: Genomic SNP data, mortality outcomes, and annotations
- Method: Subsampling techniques with XGBoost
- Results: Accuracy/F1 vs. subsampling rate
- Conclusion: Subsampling preserves predictive power

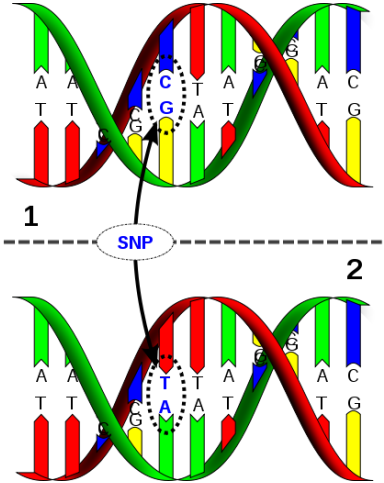
VNN and Sea Basses

Viral nervous necrosis (**VNN**) is a highly spread disease among sealife.



We concentrate our efforts on predicting the mortality of a population of **sea basses** affected by VNN.

SNPs for predicting mortality



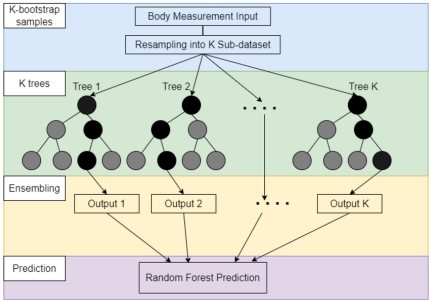
The **genome** might be useful to predict mortality.

SNPs: Single nucleotide polymorphisms.

Machine Learning Approach

Predict if a sea bass will die by watching its genome: **Machine Learning**.

In particular, we use the **XGBoost** classifier.



Challenges with Genomic Data

- Each fish: over **6 million** SNP positions.
- Sample size: only **990** sea bass individuals.
- Traditional models may **overfit** due to high dimensionality.
- We mitigate through **subsampling**.

Pipeline Overview

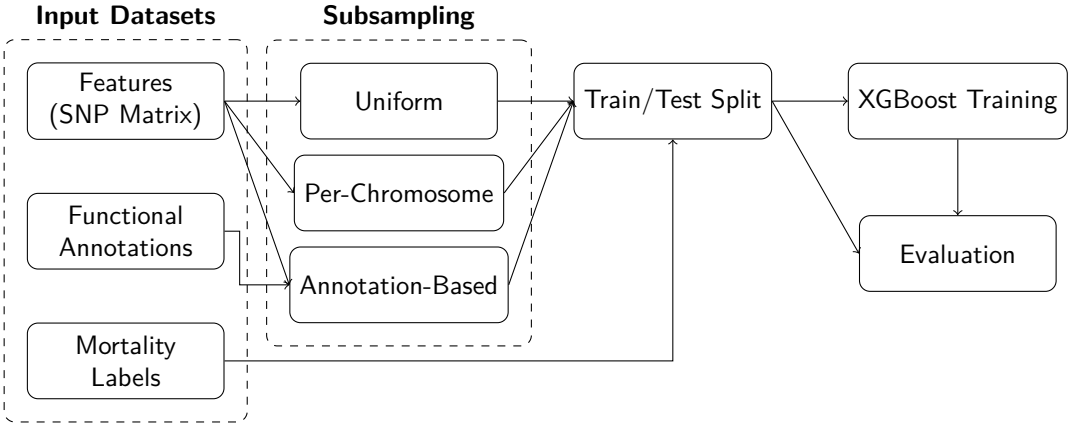


Figure: Pipeline of the model training after subsampling of the data

SNP Dataset Structure

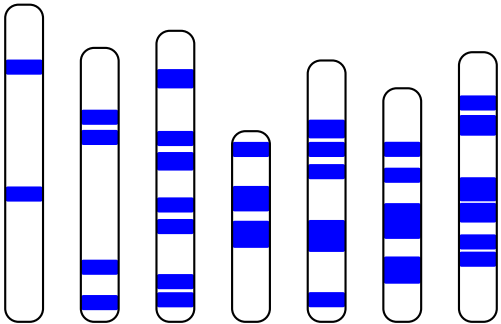
- 990 rows (fish), each with 6,072,853 SNP features.
- SNP values: 0 (no mutation), 1 (heterozygous), 2 (homozygous alt).
- Each fish is paired with a mortality label.

Annotation Metadata

- Annotations include function: Promoter, Enhancer, Open Chromatin.
- Tissue number (0–25) indicates location-specific relevance.

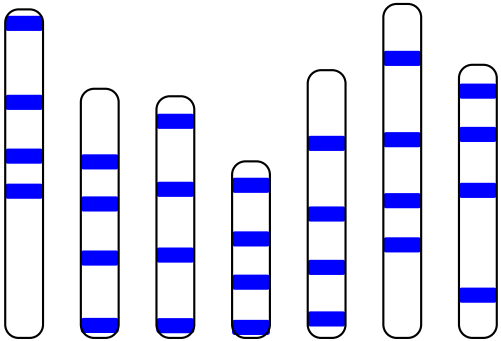
Uniform Subsampling

- Randomly sample a fixed proportion p of all SNPs.
- Simple but may cause imbalance across chromosomes.



Per-Chromosome Subsampling

- Ensures balanced representation from each chromosome.
- Randomly sample same number of SNPs per chromosome.



Annotation-Based Subsampling

- Filter SNPs by biological annotation.
- Then apply uniform subsampling to relevant regions.

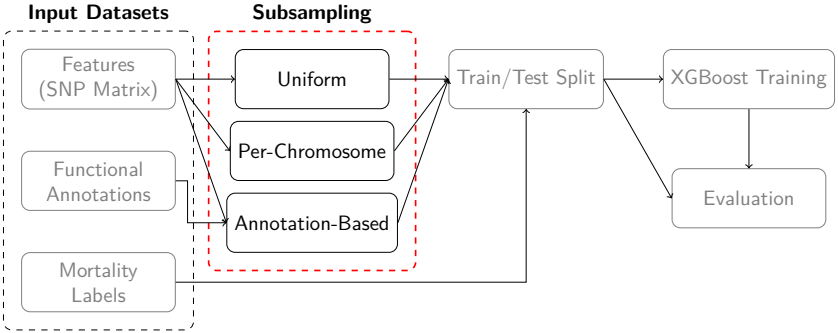
Subsampling Strategy Comparison

- Trade-offs in simplicity, biological interpretability, and balance.
- Aim: maximize predictive power while reducing dimensionality.

Control of Randomness

In order to limit the variance of results we impose:

- XGBoost random seed, train-test split fixed
- Subsampling is the only random component varying between experiments.



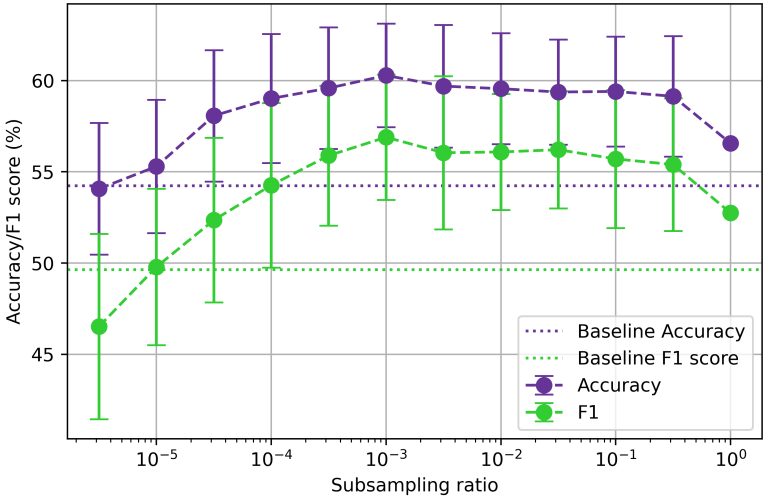
Subsampling Ratios and baseline results

- Subsampled with multiple p values: log-spaced varying from the whole genome to few SNPs.
- Trained model for each combination of model and subsample rate.
- Multiple runs for each pair of parameters.
- Comparison with **baseline results** from a "dumb" classifier, always guessing the most common class.

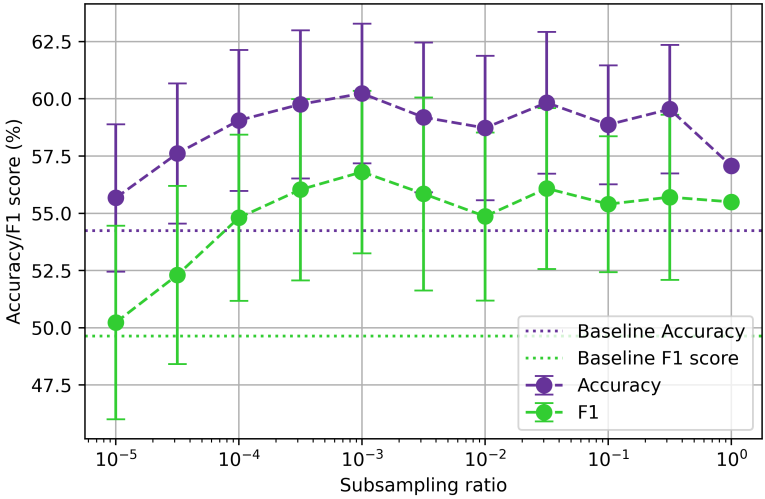
Results: testing setup

Baseline scores: In every plot we compared our scores with some "baseline scores". These baseline scores are the score of a "trivial" classifier, that always reports the most common class.
For different subsampling ratios and techniques we ran multiple instances.

Results: uniform subsampling



Results: uniform subsampling on each chromosome



Results: annotated subsampling (function)

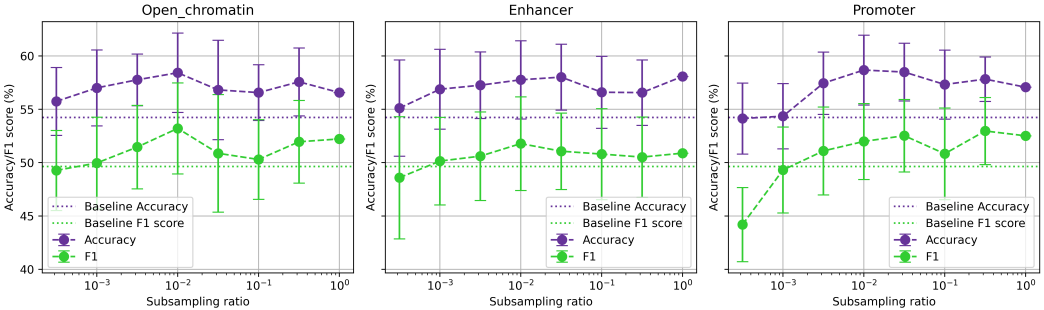


Figure: Plot of accuracy and F1 scores when subsampling uniformly on each chromosome.

Results: annotated subsampling (tissue number)

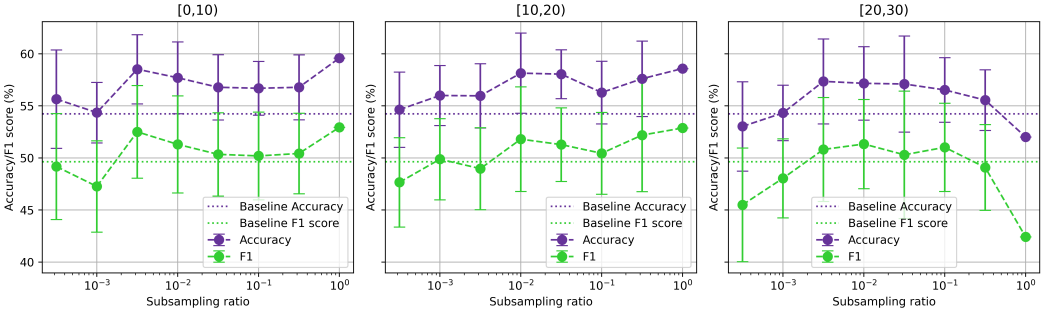


Figure: Plot of accuracy and F1 scores when subsampling uniformly on each chromosome.

Conclusions

- Random SNP subsampling retains model effectiveness.
- No strong trend between rate and accuracy (outside extremes): this may be good.
- There doesn't seem to be specific regions of the genome containing the information determining the disease effects.

Thank You!