UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Random subsampling techniques for sea bass mortality prediction

Giovanni Gaio, Simone Moretti

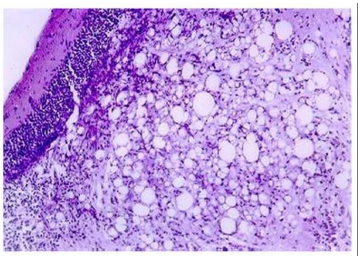
July 24, 2025

Overview

1. Introduction
2. Methodology
   - Datasets description
   - Subsampling techniques
3. Results
   - Uniform subsampling
   - Uniform over chromosomes
   - Annotated subsampling
4. Conclusions

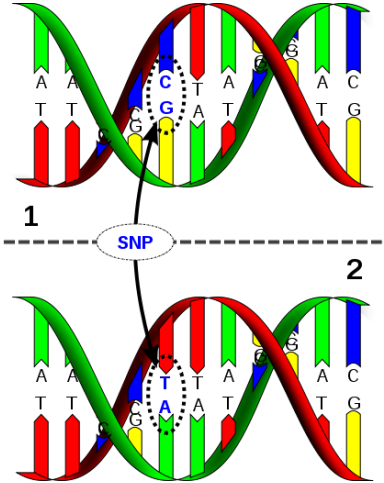
VNN and Sea Basses

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.

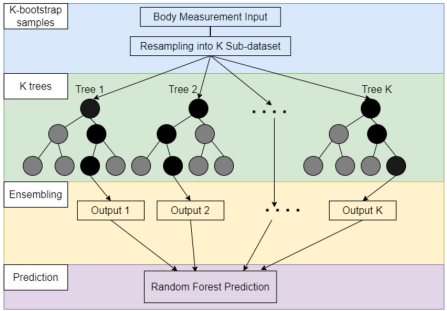
SNPs: Single nucleotide polymorphisms.

Machine Learning Approach

dmlc XGBoost

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

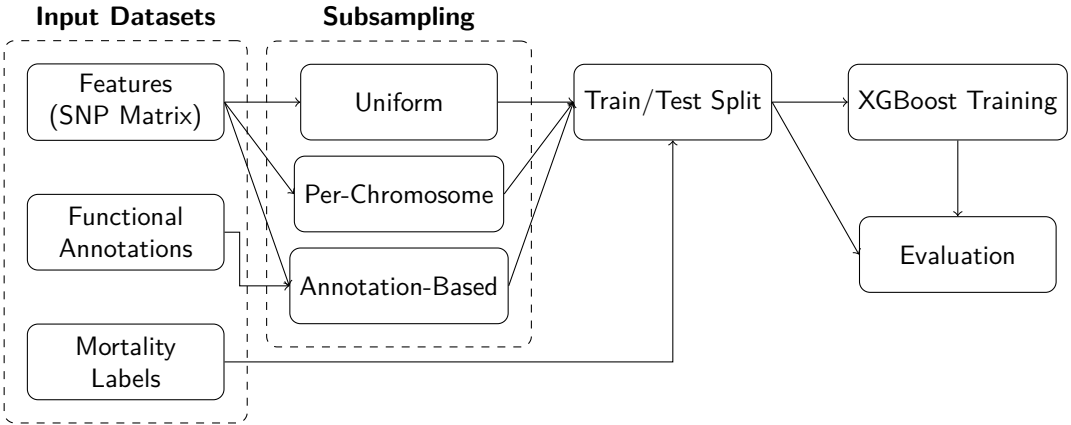
In particular, we use the **XGBoost** classifier.



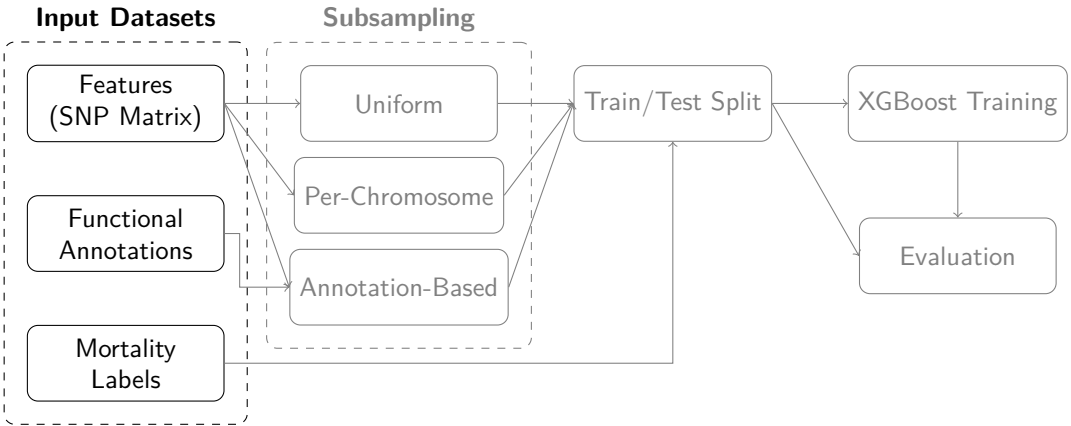
Challenges with Genomic Data

- 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



Input Datasets



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.

id	mortality
PL06-B12	1
PL06-B06	1
PL06-E06	1
PL08n-B05	0
PL08n-G09	0
⋮	⋮

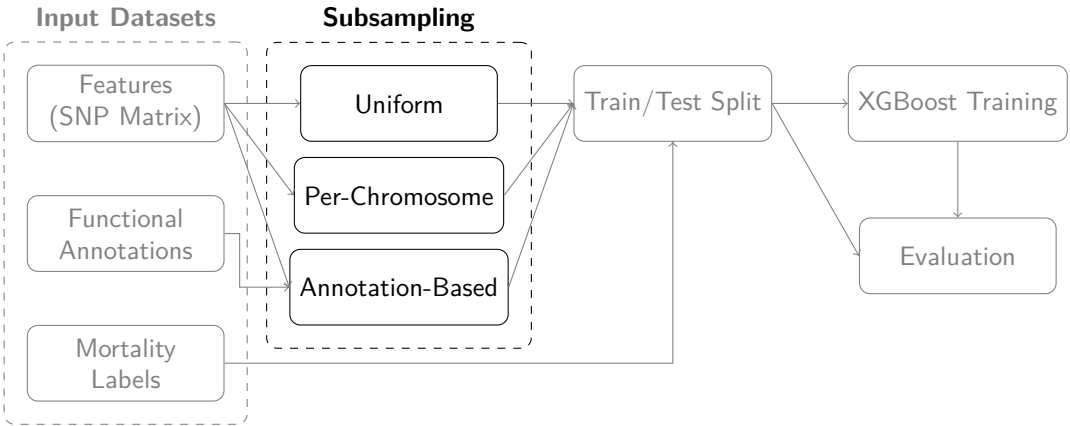
	CAJNNU010000001.1:299	CAJNNU010000001.1:903	CAJNNU010000001.1:986	...
PL04-A06	1	0	0	...
PL04-A08	0	0	1	...
PL04-A09	0	1	1	...
PL04-A10	0	0	0	...
PL04-A11	2	2	1	...
⋮	⋮	⋮	⋮	⋱

Annotation Metadata

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

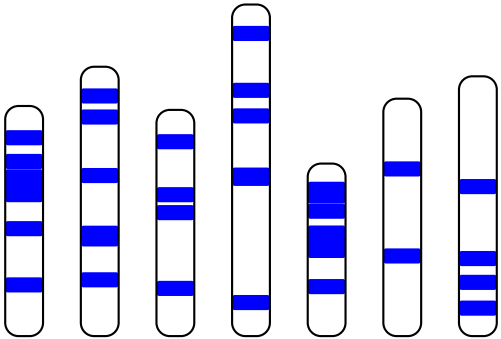
snp_id	funct	n_tissue
CAJNNU010000001.1:7825	Open_chromatin	10
CAJNNU010000001.1:7865	Open_chromatin	4
CAJNNU010000001.1:8046	Enhancer	21
CAJNNU010000001.1:8084	Open_chromatin	5
CAJNNU010000001.1:8116	Promoter	12
⋮	⋮	⋮

Subsampling techniques



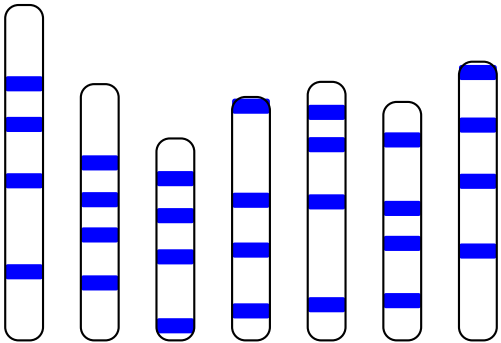
Uniform Subsampling

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



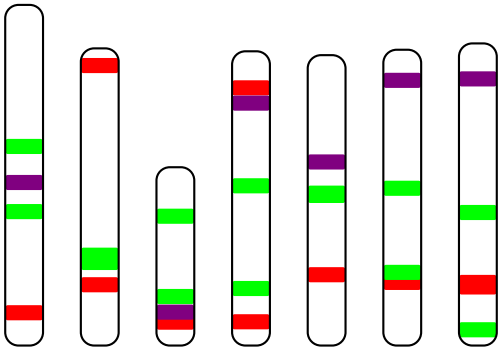
Per-Chromosome Subsampling

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

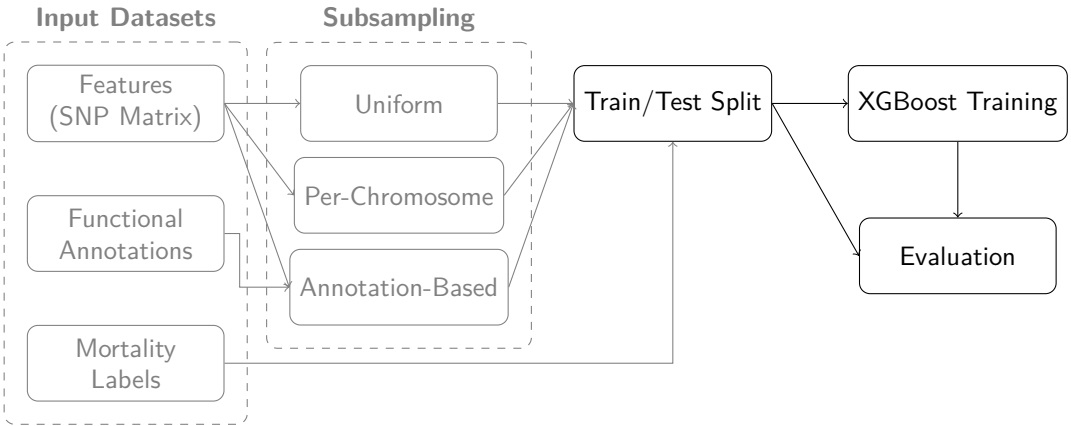


Annotation-Based Subsampling

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



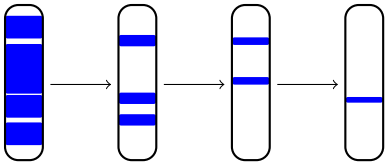
Parameter Evaluation



Let's get to the substance

Subsampling Ratios and baseline results

- Subsampled with multiple subasmpling ratio values: log-spaced varying from the whole genome to few SNPs.

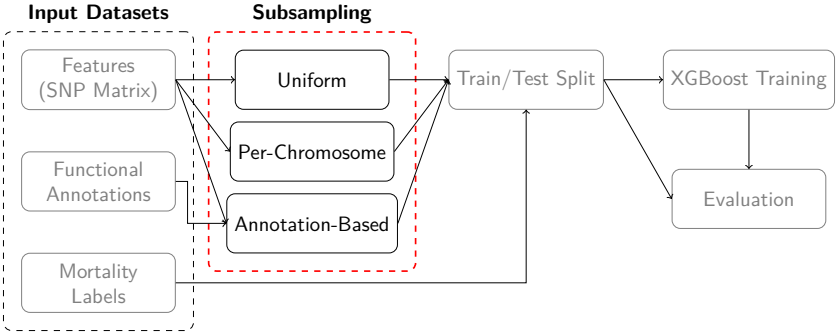


- Trained model for each combination of model and subsample rate.
- Multiple runs for each pair of parameters.
- Comparison with **baseline results** from a "dumb" classifier, always guessing the most common class.

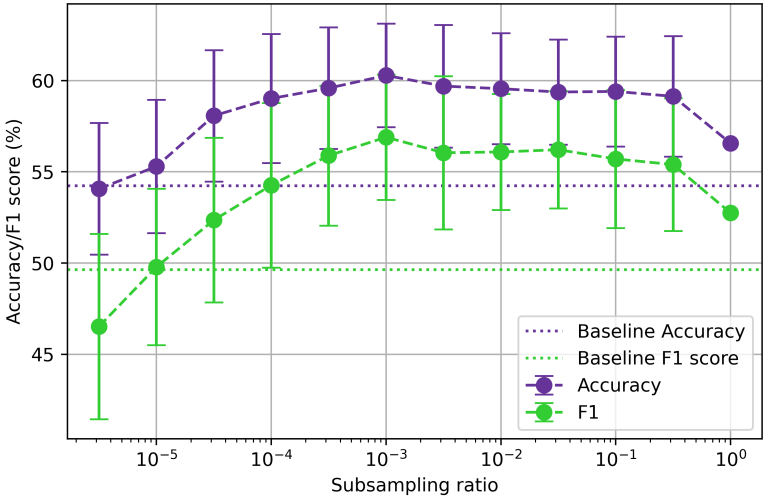
Control of Randomness

In order to limit the variance of results we impose:

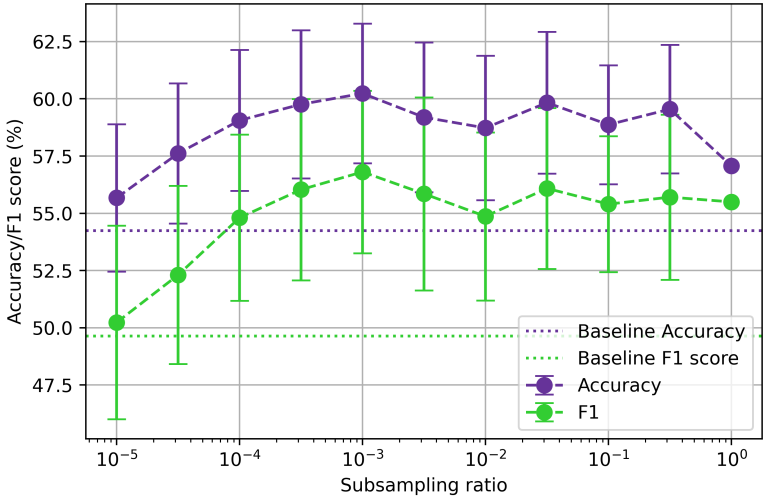
- XGBoost random seed and train-test split are fixed.
- Subsampling is the only random component varying between experiments.



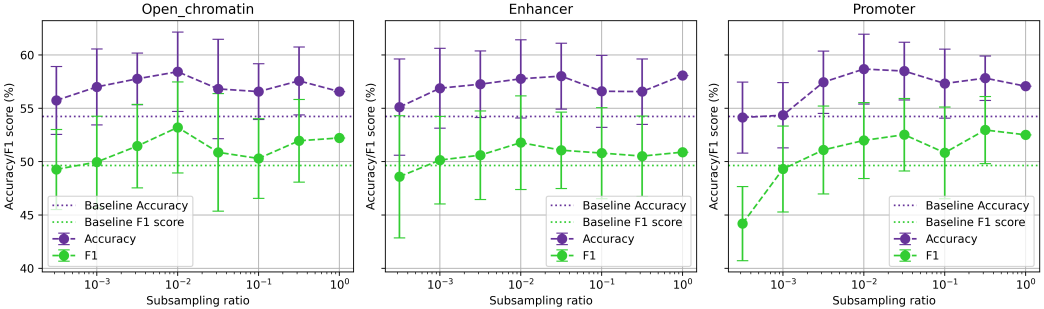
Results: uniform subsampling



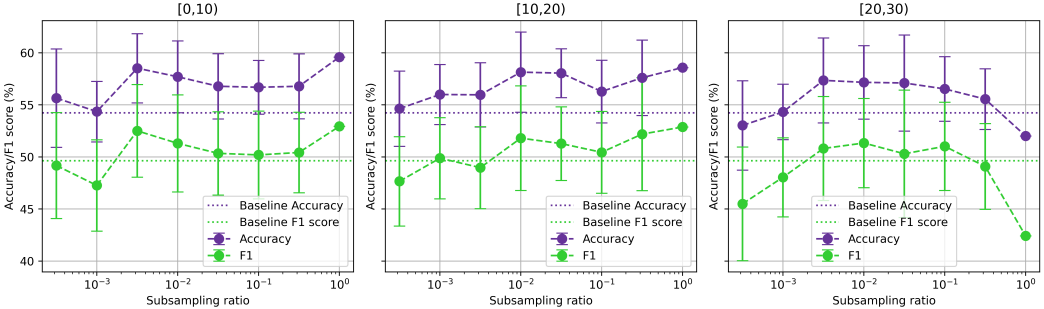
Results: subsampling uniformly on chromosomes



Results: annotated subsampling (function)



Results: annotated subsampling (tissue number)



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!