



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Random subsampling techniques for sea bass mortality prediction

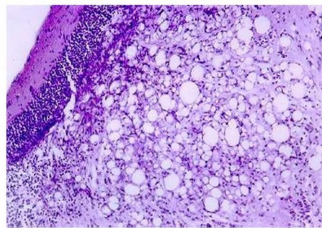
Giovanni Gaio, Simone Moretti

July 22, 2025

- 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

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

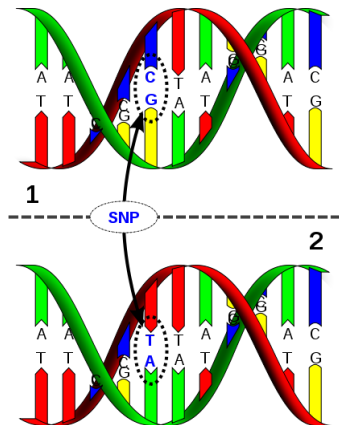
We concentrate our efforts on predicting the mortality of a population of **sea basses** affected by VNN.



SNPs for predicting mortality



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



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

SNPs: Single nucleotide polymorphisms.

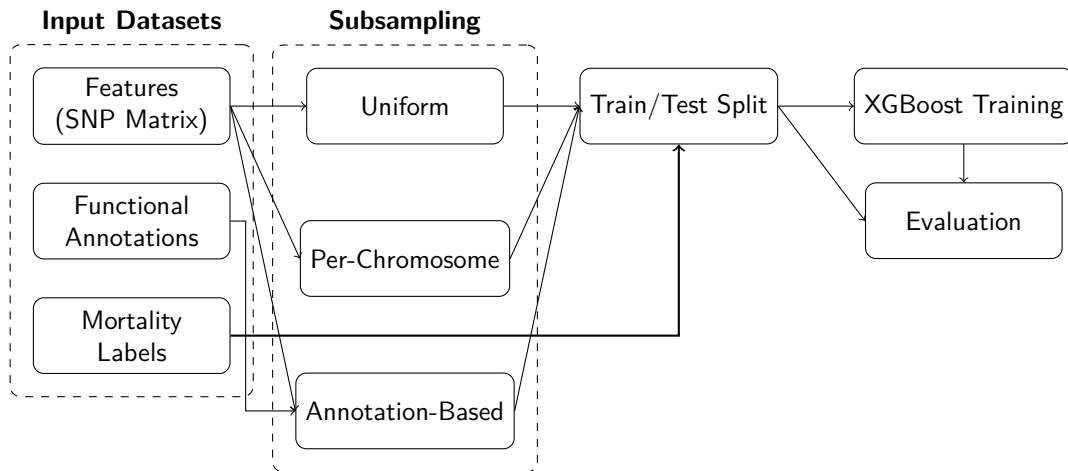
- Each fish: over 6 million SNP positions.
- Sample size: only 990 sea bass individuals.
- Traditional models overfit due to data dimensionality.

- Use XGBoost classifier for mortality prediction.
- Need to reduce feature space: apply subsampling.
- Evaluate performance on subsampled datasets.

Pipeline Overview



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



- 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.

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

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

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

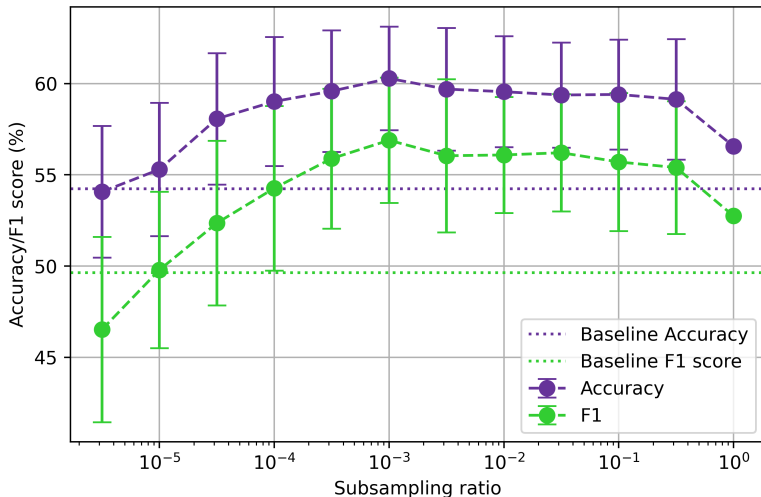
- Filter SNPs by biological annotation.
- Then apply uniform subsampling to relevant regions.

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

- XGBoost random seed fixed.
- Train-test split fixed
- Subsampling is the only random step.

- 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.

Results: uniform subsampling

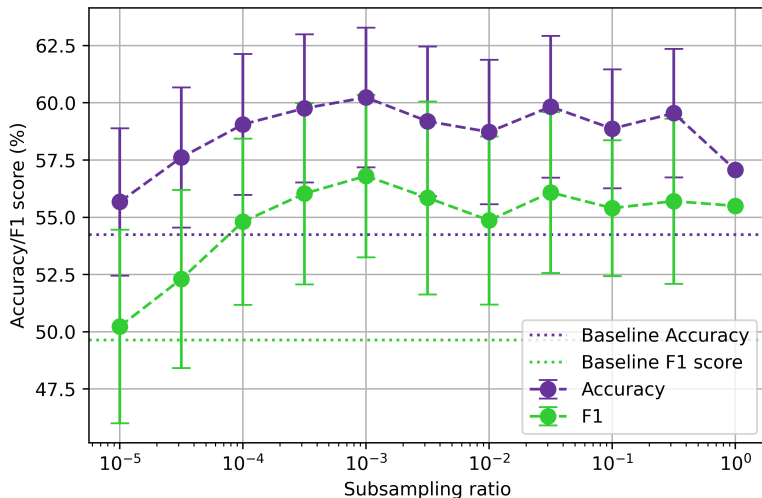


Results: observations



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Results: uniform subsampling on each chromosome



Results: observations



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Results: annotated subsampling (function)

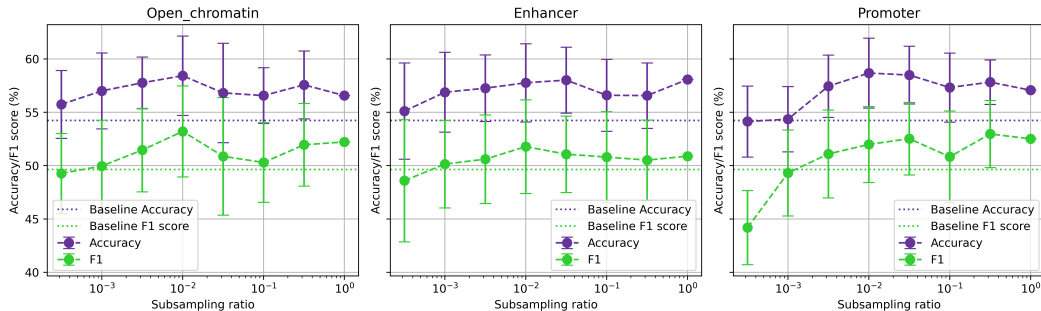


Figure: Plot of accuracy and F1 scores when subsampling uniformly on each chromosome.

Results: annotated subsampling (tissue number)

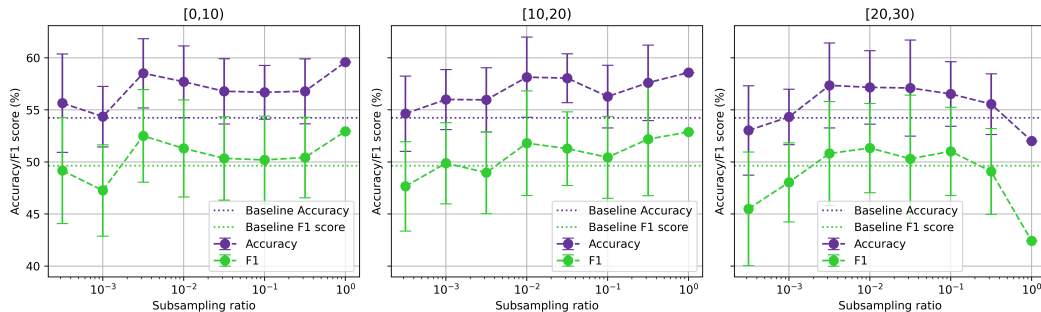


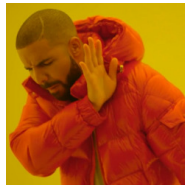
Figure: Plot of accuracy and F1 scores when subsampling uniformly on each chromosome.

Results: observations



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- 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.



The
results don't
show any
meaningful trend



We showed
that subsampling
allows
the use of more
complex models

End



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Thank You