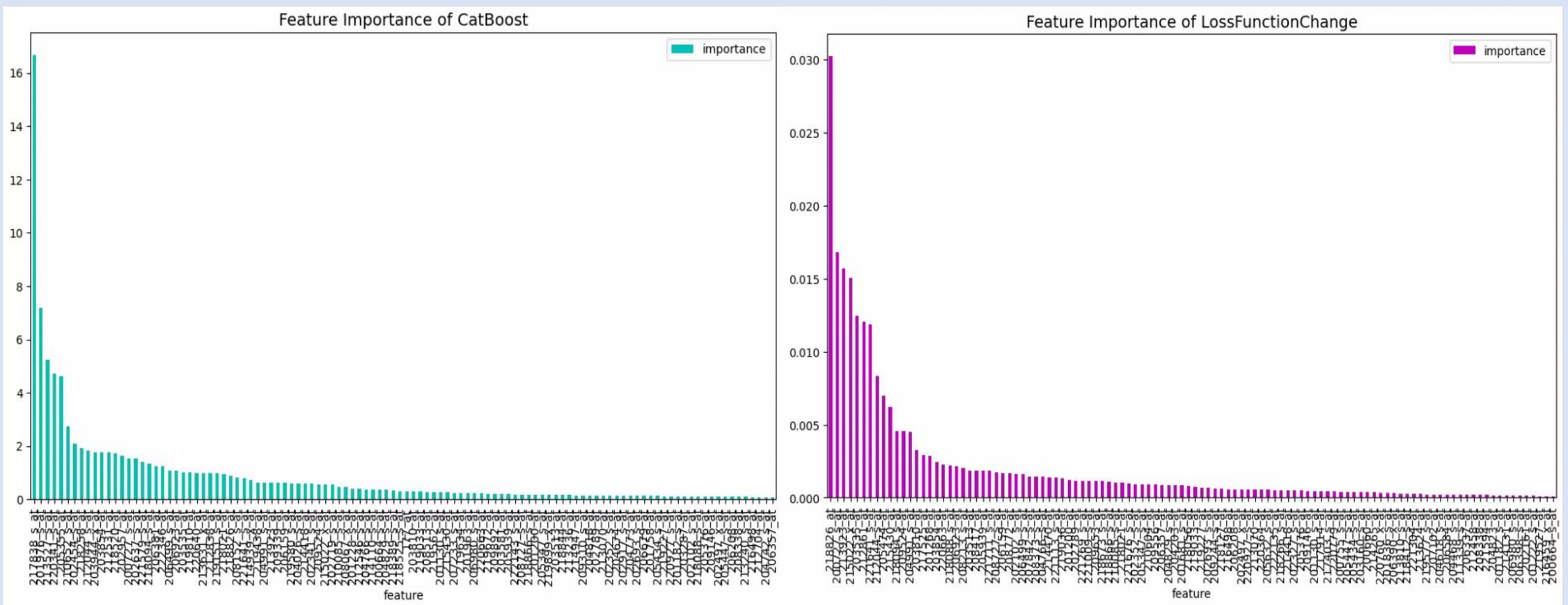
Biomachine Learning: the Dimensionality Reduction and Feature Selection Using Categorical Gradient Boosted Trees Paradigm

Athanasios Angelakis [1, 2, 3]

1) Amsterdam University Medical Centers; 2) Amsterdam Public Health Research Institute; 3) University of Amsterdam Data Science Center
a.angelakis@amsterdamumc.nl twitter: @AthAngelakis https://www.linkedin.com/in/athanasios-angelakis-116723103/

Utilizing a novel dimensionality reduction and feature selection methodology, using explainable artificial intelligence, predictability and loss error feature importance, we have achieved superior diagnostic results in breast cancer transcriptomics regarding estrogen receptor.

Even with a relatively small dataset of 344 data instances and 13092 features, we highlight the potential of machine learning in revolutionizing conventional statistical approaches in computational biology.



Models	Spec Test	Sens Test	AUC Test	F1 Test	Spec 10CV	Sens 10CV	AUC 10CV	F1 10CV
CatBoost 13092	0.7500	0.8947	0.8333	0.8500	0.8609	0.8866	0.8665	0.9005
CatBoost 1095	0.8125	0.9474	0.8929	0.9000	0.8860	0.8974	0.8828	0.9138
CatBoost 50	0.7778	1.0000	0.9048	0.8947	0.8947	0.9026	0.8896	0.9191

1. Robust Performance Across Cancer Types: Applying the methodology defined by A. Angelakis et al. [1], [2] previously tested on acute myeloid leukaemia and acute lymphoblastic leukemia yielded promising results for breast cancer, indicating its potential adaptability for various cancer types.

2. Significant Improvement with Reduced Genes: By retaining only 50 genes from the original 13,092 probesets after two rounds of dimensionality reduction, we achieved a test ROC-AUC score of 0.9048, surpassing the accuracy achieved using all genes. Reaching a number of less than 20 genes could make our approach clinically applicable.

3. Potential for New Discoveries: Our methodology not only promises better diagnostic accuracy but also paves the way for the identification of new disease-related genes and gene profile signatures [1], [2], even from small datasets. This could revolutionize gene-based medical research and drug discovery.

4. Overcoming Limitations of Conventional Approaches: Traditional statistical methods for dimensionality reduction/feature selection on gene expression microarray datasets face challenges, especially with linearities. Our novel methodology, however, demonstrates robust performance and adaptability even with small datasets comprising as few as 56 data instances, showcasing its potential to outperform and revolutionize these conventional techniques.

[1]. A. Angelakis et al. *Diagnosis of Acute Myeloid Leukaemia Using Machine Learning*, preprint, arXiv:2108.07396
[2]. A. Angelakis et al. *Uncovering the Gene Profile of Acute Myeloid Leukaemia Using Gradient Boosted Trees,* preprint, https://doi.org/10.21203/rs.3.rs-2374064/v1

Amsterdam UMC Universitair Medische Centra

EDS

University of Amsterdam Data Science Centre