BIOSTATISTICS GROUP
WROCLAW, POLAND

Enhancing bovine genome SNP call accuracy with autoencoder analysis of nucleotide impact with AI

Kotlarz K.[1], Mielczarek M.[1,2] , Biecek P.[4,5], Guldbrandtsen B.[3] and Szyda J.[1,2]

1. Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
2. National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland
3. Department of Veterinary and Animal Sciences, University of Copenhagen, Grønnegårdsvej 8, 1870 Frederiksberg C, Denmark
4. Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
5. Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland

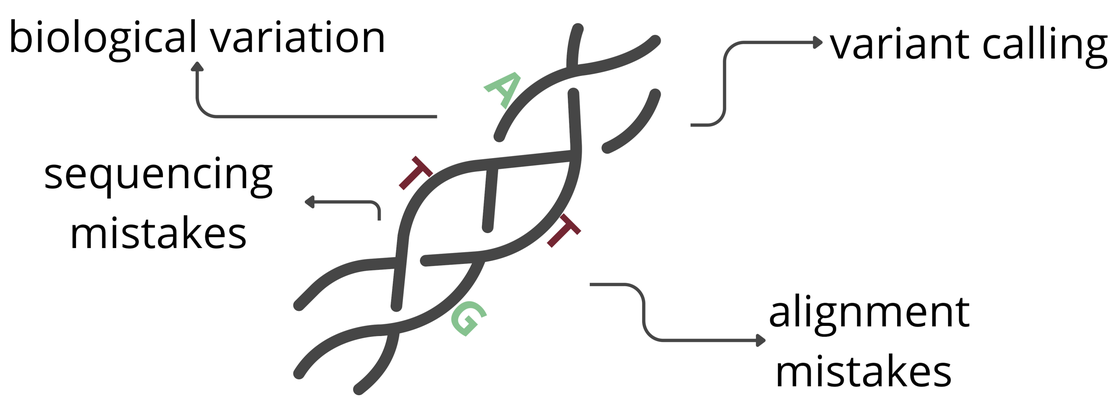
Objective

- Variant calling: critical step in the analysis of NGS data
- Potential mistakes for a number of reasons
- **<1% Incorrect calls: anomaly detection procedure implemented via autoencoder (AE) model**

Conclusions

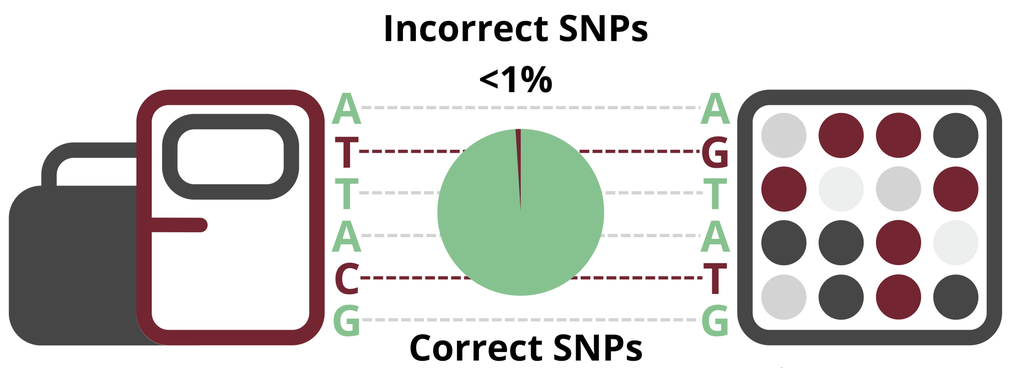
- **Autoencoders show promise in detecting anomalies and efficiently capturing complex patterns**
- Data with limited covariates variability can be represented by principal components
- The number of principal components and neural network architecture profoundly influence anomaly detection efficacy

Causes of mistakes



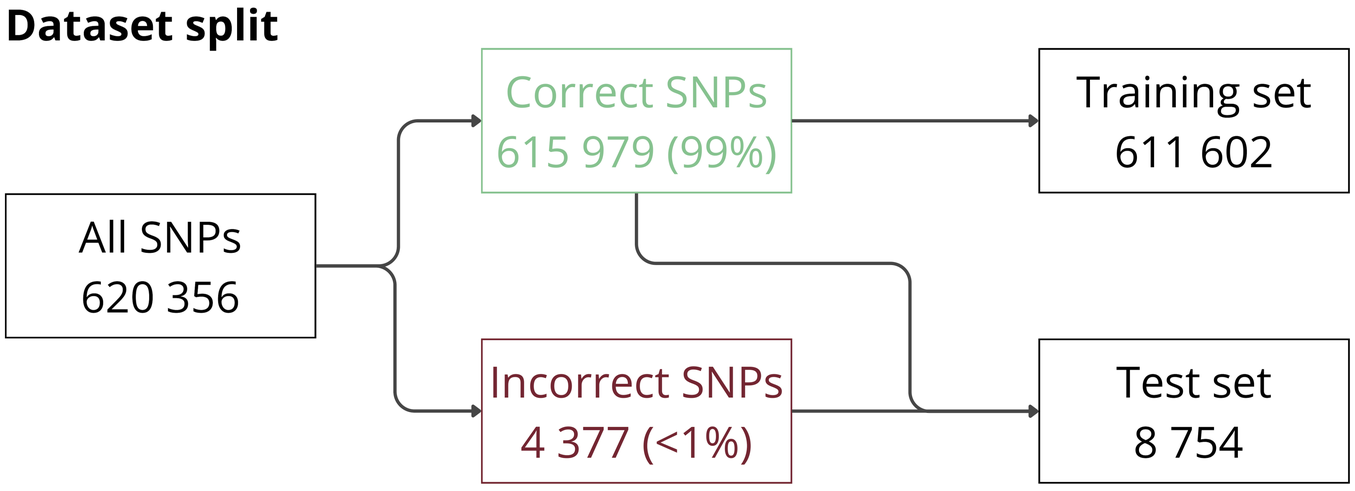
Incorrect calls

- 20 Polish Holstein-Friesian cows



Autoencoders

Dataset split

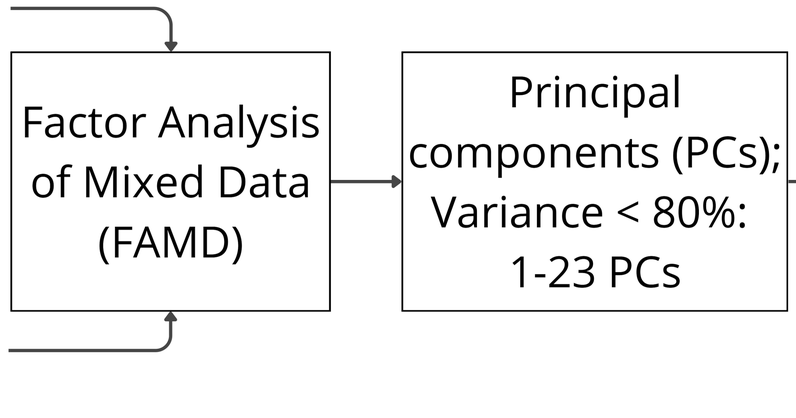


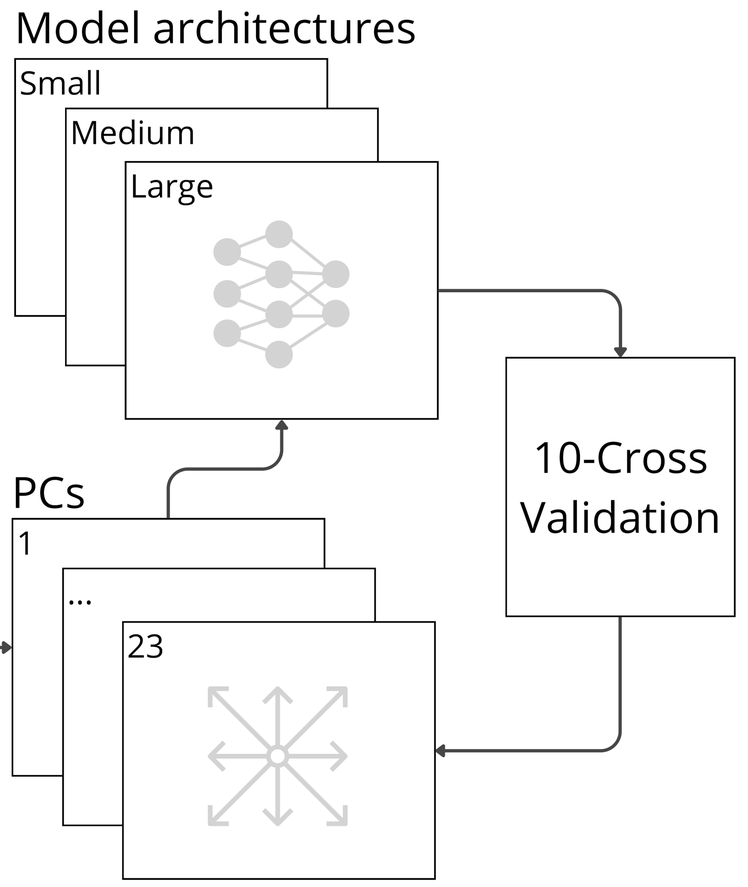
Model implementation

- Downstream: 1-4 bp
- Upstream: 1-4 bp

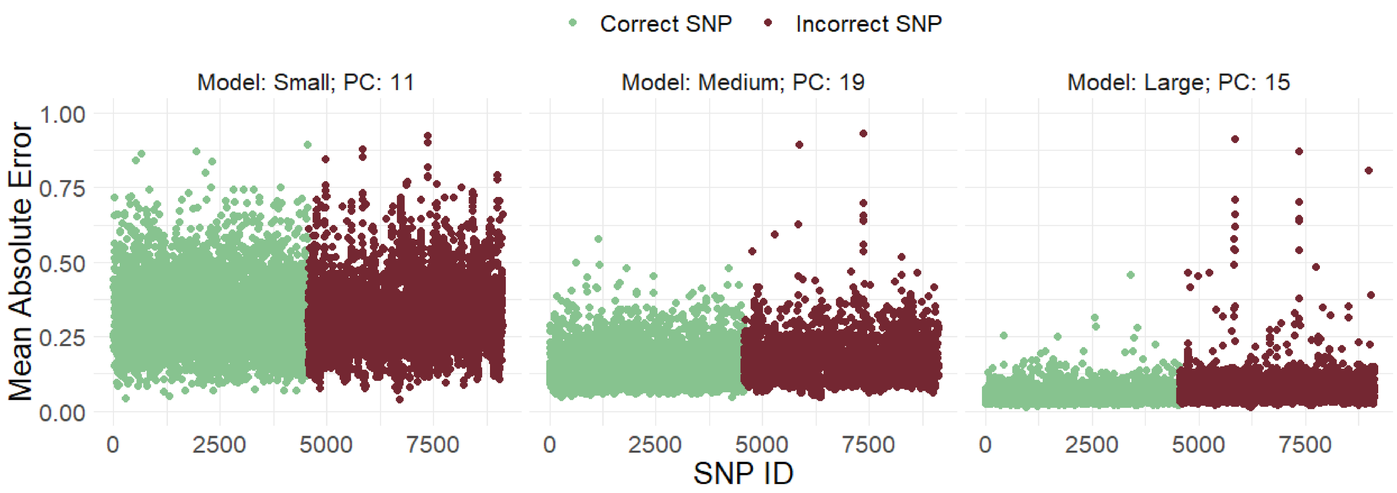
VCF

- Allele: A
- Allele: B
- Read depth: A
- Read depth: B
- Genotype quality





Results



Best model selection:

- Mann Whitney U test
- Sensitivity/Precision threshold

WROCŁAW UNIVERSITY OF ENVIRONMENTAL AND LIFE SCIENCES

Krzysztof Kotlarz
THETA Biostatistics Group
Wrocław University of Environmental and Life Sciences
krzysztof.kotlarz@upwr.edu.pl