Wrocław University of Environmental and Life Sciences

K. Kotlarz, M. Mielczarek, T. Suchocki, B. Czech, B. Guldbrandtsen, J. Szyda

DIG (DEEP)ER

Deep learning algorithms for the imbalanced classification of correct and incorrect SNP genotypes from WGS pipelines

Wrocław University of Environmental and Life Science, University of Bonn

THETA
Statistical Genetic Group
Institute of Animal Genetics

#1 MATERIALS

Whole-genome DNA sequence of four traditional Danish Red Dairy Cattle bulls:

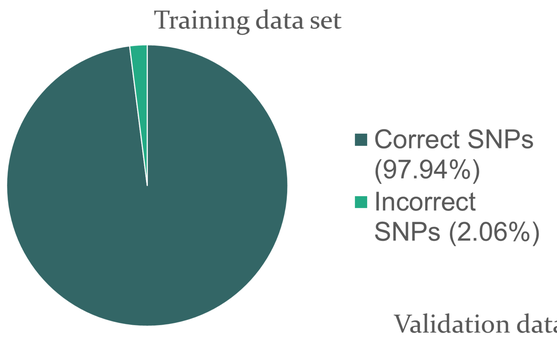
1) The training data set—**three animals**,
2) The validation data set—**the fourth animal**.

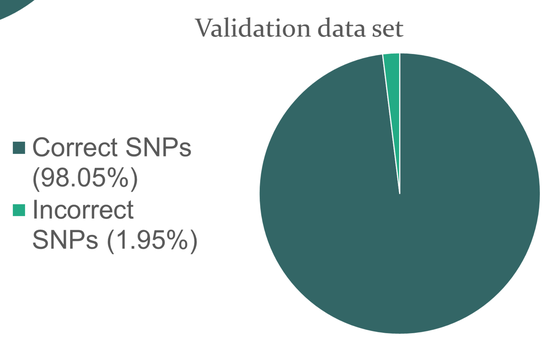
Correct SNPs (concordant WGS—Chip):

1) Training data set: 2 227 995 SNPs,
2) Validation data set: 749 506 SNPs.

Incorrect SNPs (discordant WGS—Chip):

1) Training data set: 46 920 SNPs,
2) Validation data set: 14 940 SNPs.



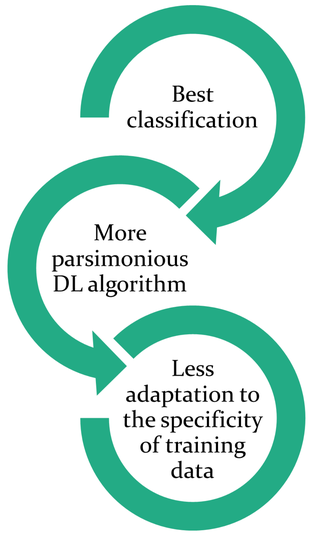


Contact:

Department of Genetics
Wrocław University
of Environmental
and Life Sciences

7 Kożuchowska Street
51-631 Wrocław
Poland
krzysztof.kotlarz@upwr.edu.pl

#4 CONCLUSIONS



#2 METHODS

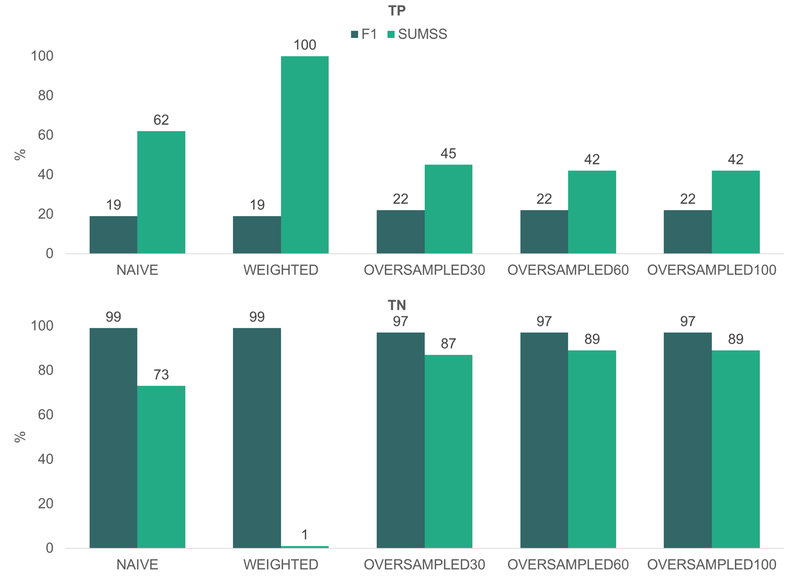
Deep Learning algorithms

1) Naïve algorithm
2) Weighted algorithm
3) Oversampled algorithm:
   - 30%
   - 60%
   - 100%

Cutoff points

1) The estimated cutoff points for each model by:
   - $F1 = \frac{2TP}{2TP+FN+FP}$
   - $SUMSS = \frac{TN}{TN+FP} + \frac{TP}{TP+FN}$

#3 RESULTS



Classification of **validation data** by the algorithms, based on the cutoff thresholds for the **F1** or **SUMSS** metrics.

1) **True positive (TP)**—an incorrect SNP classified as incorrect,
2) **False negative (FN)**—an incorrect SNP classified as correct,
3) **True negative (TN)**—a correct SNP classified as correct,
4) **False positive (FP)**—a correct SNP classified as incorrect,
5) **F1**—values of the F1 metric.