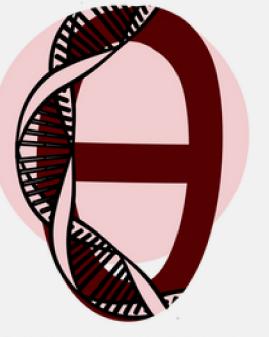


Classification of mastitis in cows using deep learning approach with model regularization

K. Kotlarz, M. Mielczarek, P. Biecek J. Szyda

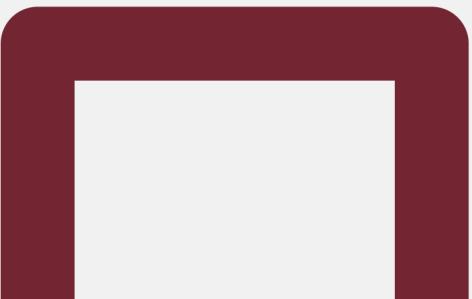


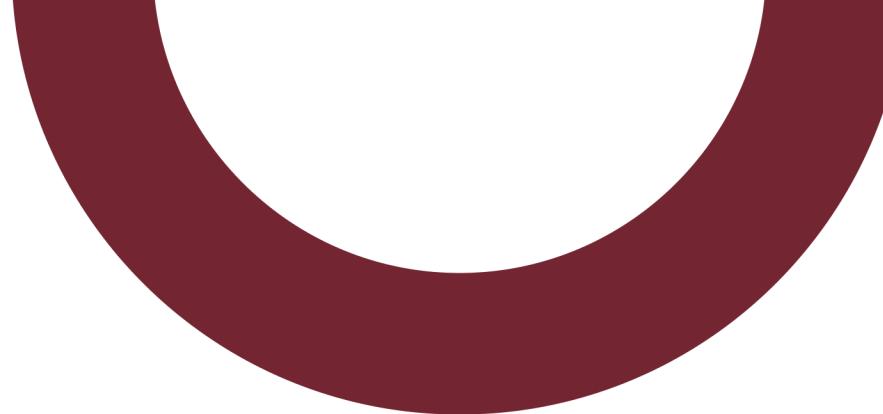
**WROCŁAW UNIVERSITY
OF ENVIRONMENTAL
AND LIFE SCIENCES**



BIOSTATISTICS GROUP
WROCŁAW, POLAND

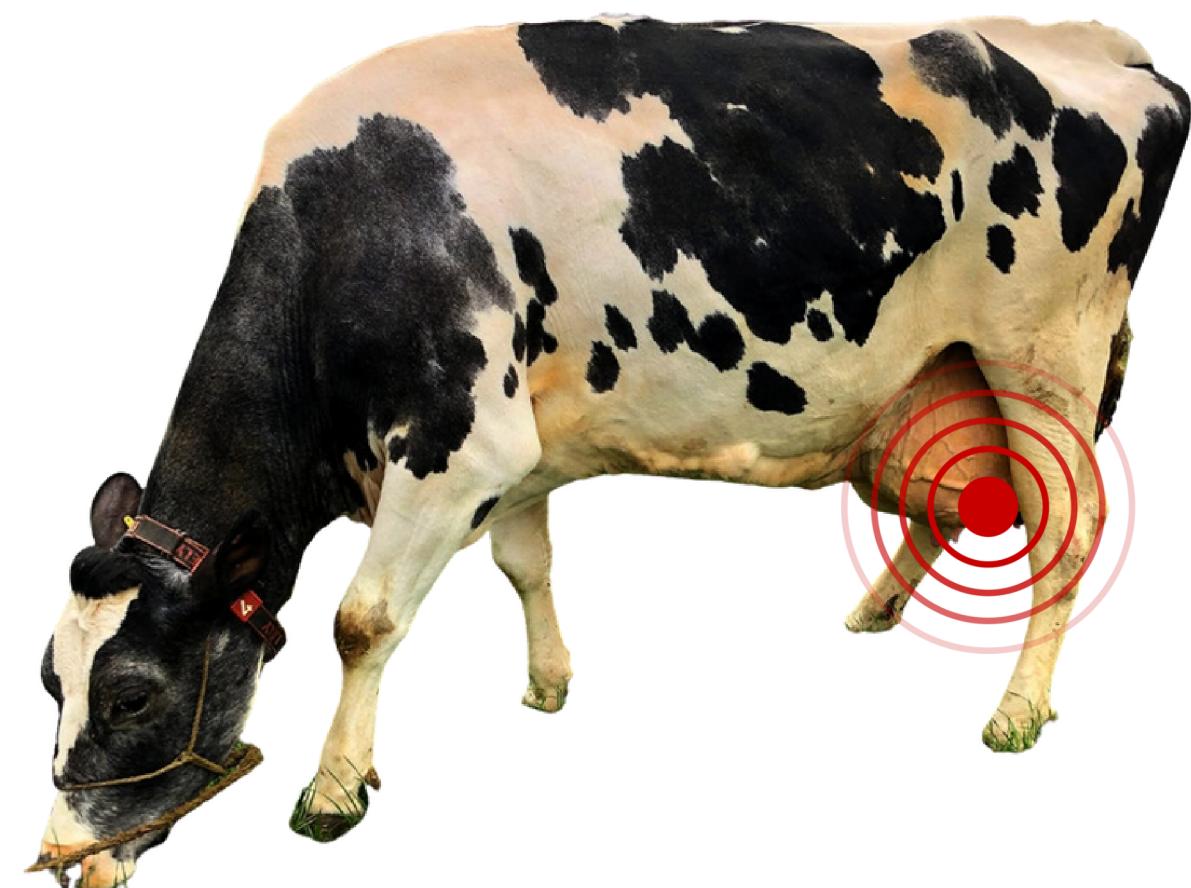
**Warsaw University
of Technology**





Clinical mastitis

- Bovine mastitis: most common disorders in dairy
- Animal welfare problems, economic losses
- Mainly caused by the bacteria



Content Index



MATERIALS

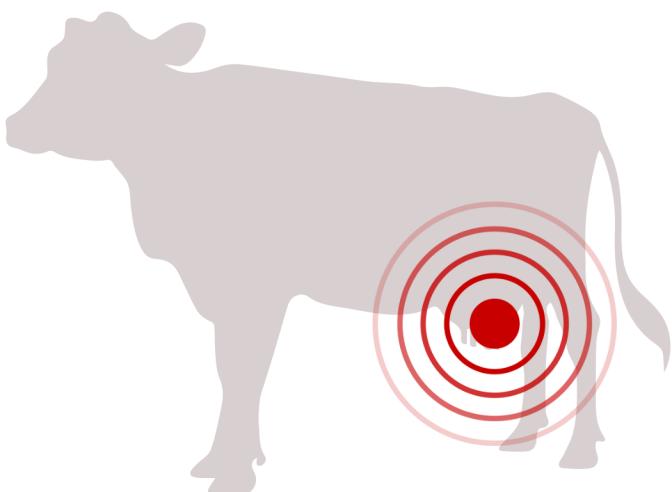
INDIVIDUALS

Whole genome samples of Polish Holstein-Friesian cows:

- Half-sibs matched for age, age at calving, year and season of the lactation start
- 31 individuals as train dataset (T)
- 20 individuals as test dataset (t)

SEQUENCING

- Illumina HiSeq2000 NGS platform
- 16 618 983 SNPs genotypes



$x 16(T) / 10(t)$



$15(T) / 10(t)x$



BIOINFORMATIC PIPELINE



1.Quality control / filtering

- FastQC
- Trimmomatic



2.Alignment

- BWA-MEM



3.Post-alignment processes

- SAMTools



4.SNP Calling

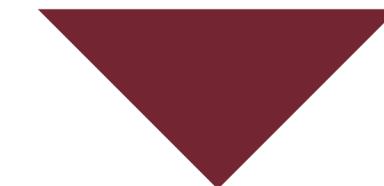
- GATK
- VCFtools

FEATURE (SNP) SELECTION

1. GENTYPE ORDER ENCODING

- 0, 1, 2

Individual/ SNP	SNP 1	SNP 2	...
Individual 1	0	1	...
Individual 2	1	0	...
...



2. LASSO REGRESSION

- Penalized Logistic Regression was used for SNP pre-selection
- Penalty parameter: <0;1> with 0.1 step

$$\frac{\exp(\beta_0 + \sum_{i=1}^{N_{SNP}} \boldsymbol{\beta}_1 x_1 + \beta_2 x_2 + \boldsymbol{\beta}_3 x_3 + \dots + \beta_N x_N)}{1 + \exp(\beta_0 + \sum_{i=1}^{N_{SNP}} \boldsymbol{\beta}_1 x_1 + \beta_2 x_2 + \boldsymbol{\beta}_3 x_3 + \dots + \beta_N x_N)}$$



3. PARAMETERS FILTERING

- Only SNPs with nonzero estimates were used in the deep-learning classifier

$$\frac{\exp(\beta_0 + \sum_{i=1}^{N_{SNP}} \mathbf{0} x_1 + \beta_2 x_2 + \mathbf{0} x_3 + \dots + \beta_N x_N)}{1 + \exp(\beta_0 + \sum_{i=1}^{N_{SNP}} \mathbf{0} x_1 + \beta_2 x_2 + \mathbf{0} x_3 + \dots + \beta_N x_N)}$$



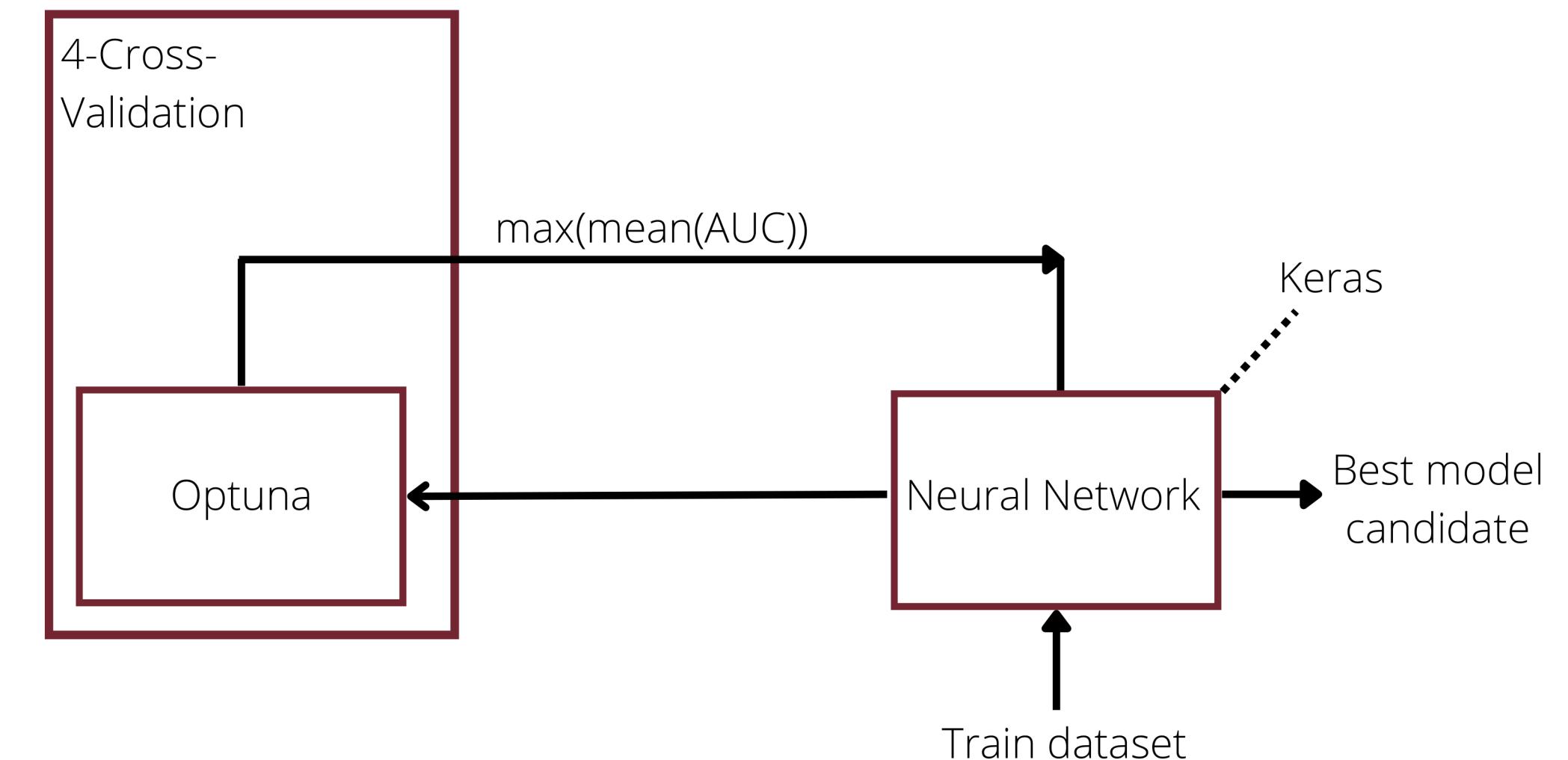
HYPERPARAMETERS TUNING

OPTUNA SOFTWARE

- Tree-structured Parzen Estimator
- 50 hyperparameters searching trials

HYPERPARAMETERS

- N-layers
- N-neurons
- Structure of Dropout layers
- Learning rate



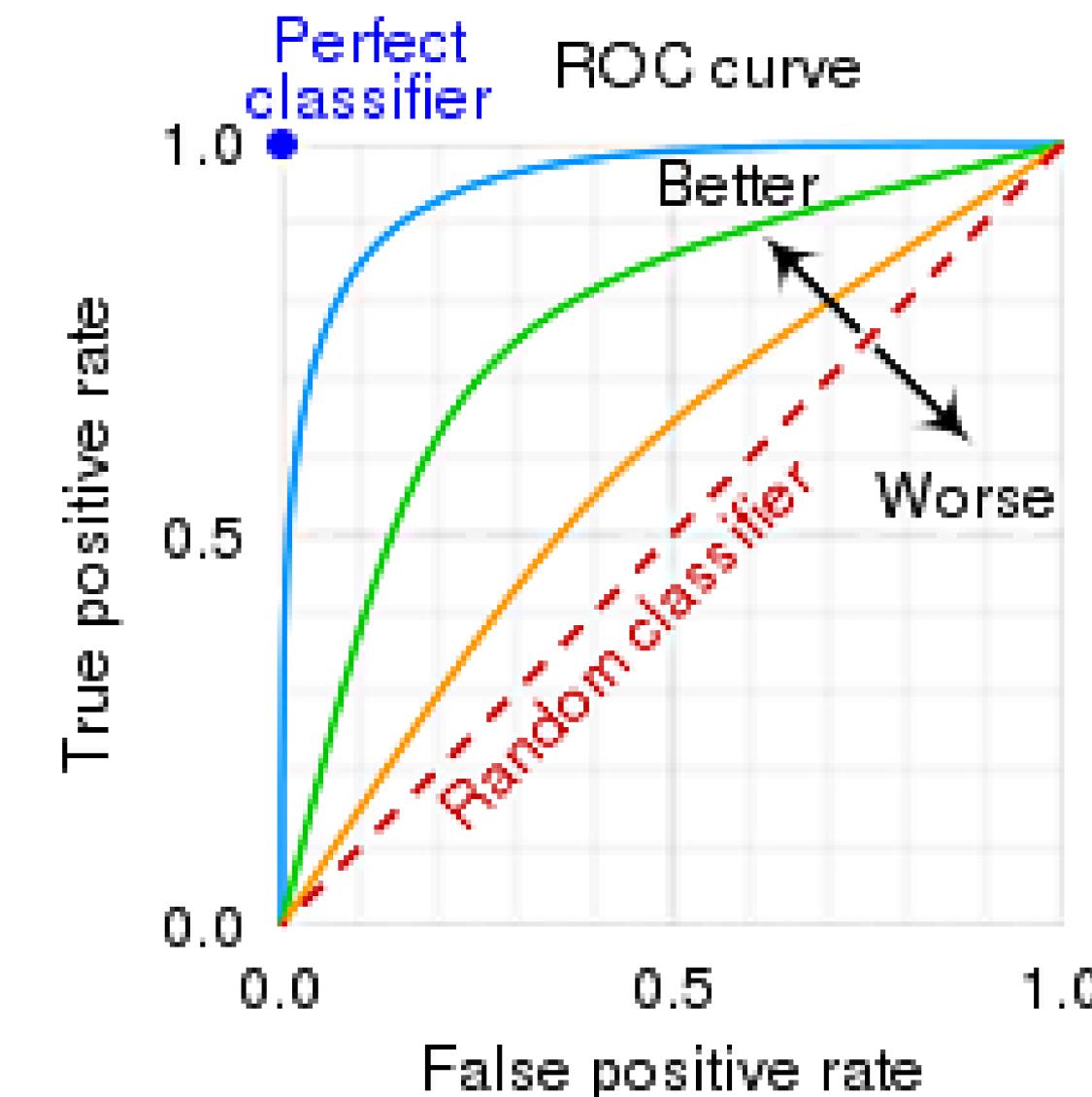
CUT-OFF POINTS ANALYSIS

CUTPOINT

- cutPointR
- Maximasing the accuracy of train dataset

VALIDATION

- 1 000 bootstrap samples of probability
- Out-of-bag



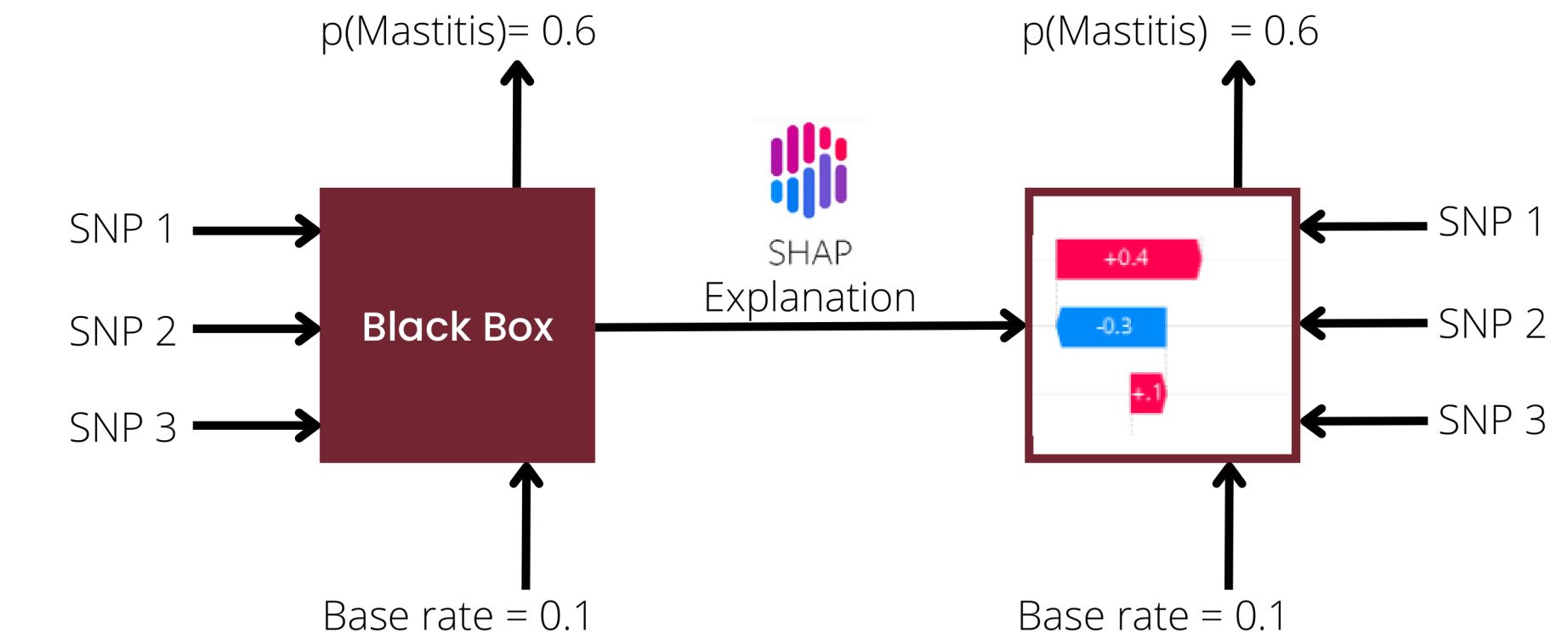
DEEP LEARNING EXPLANATORY

SHAP VALUES

Break down a prediction
to show the impact of
each SNP

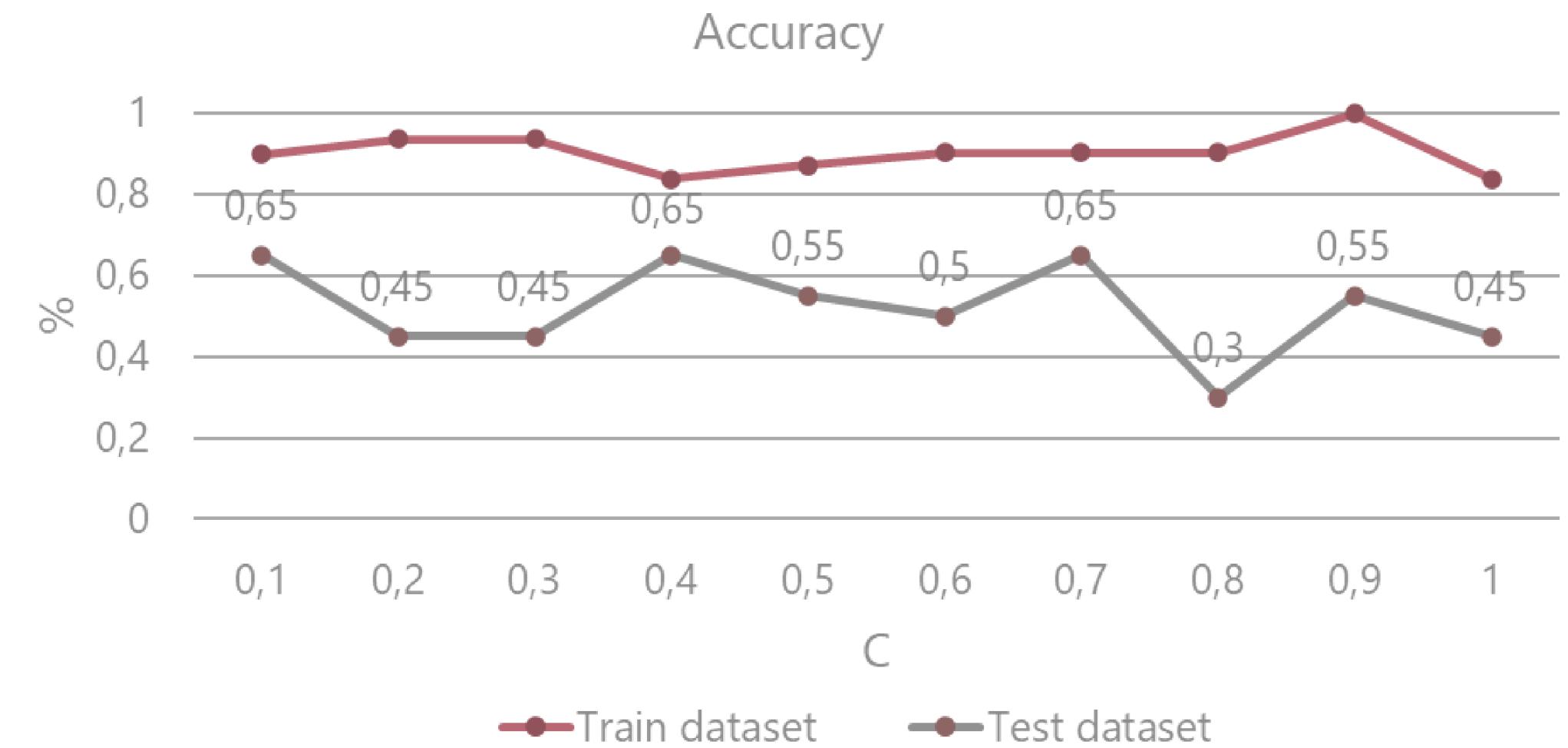
GLOBAL INTERPRETABILITY

Each SNPs effect on mastitis
susceptibility

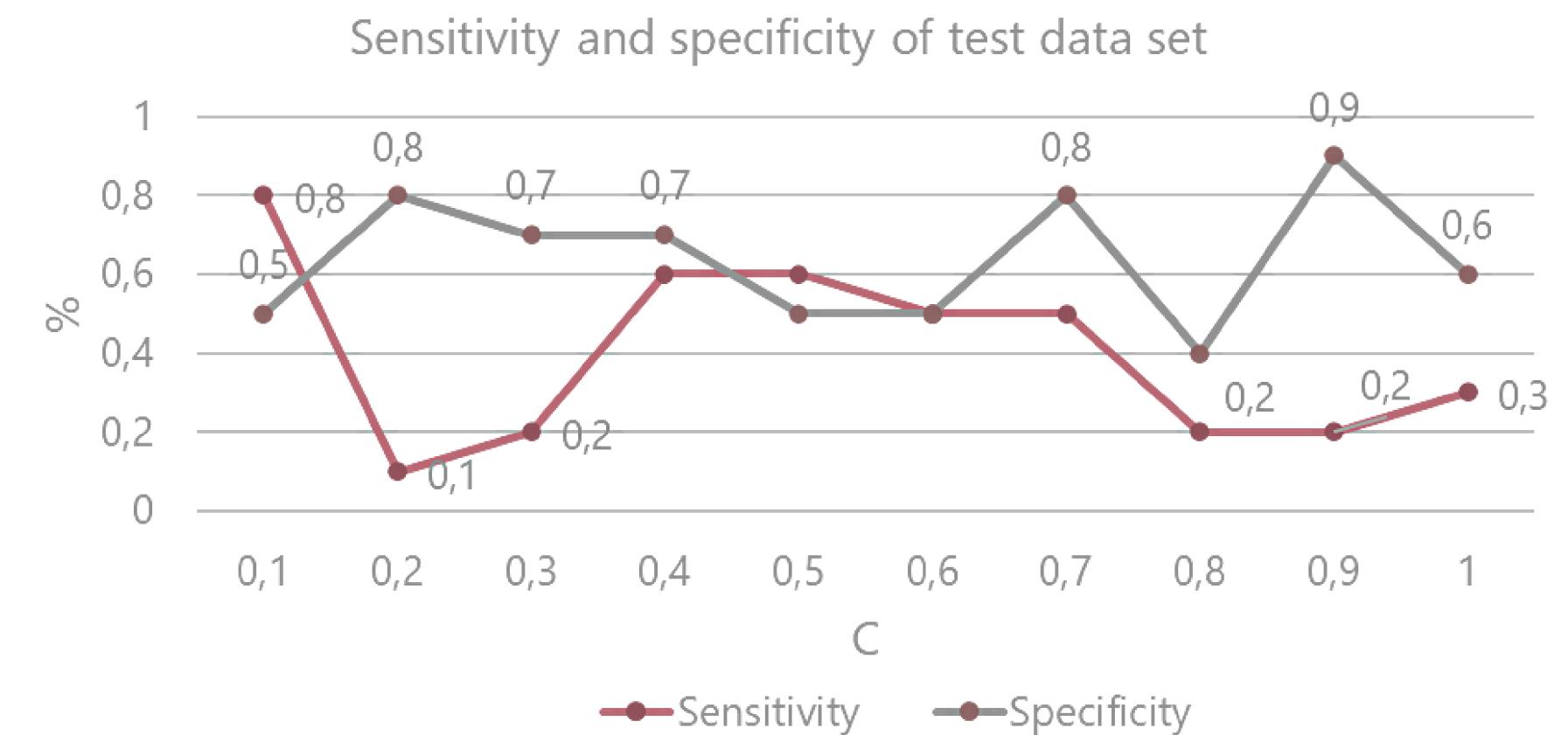
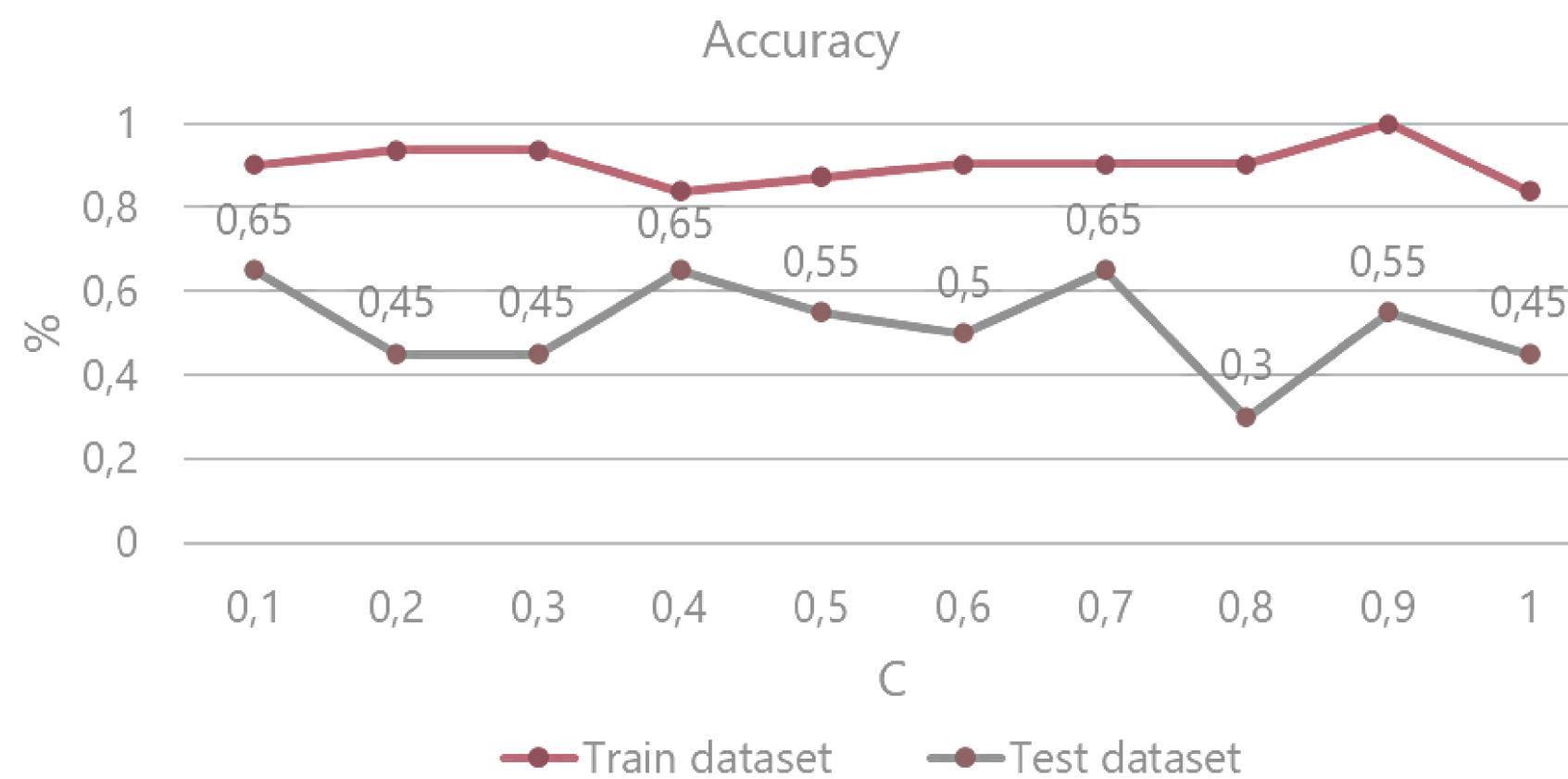


Penalty parameter (C)	SNPs number
0,1	6 665
0,2	29 869
0,3	92 136
0,4	204 642
0,5	347 924
0,6	495 476
0,7	659 423
0,8	813 650
0,9	988 650
1	1 154 608

RESULTS



RESULTS



RESULTS

GO TERMS

- GO:0006952
 - Response to biotic stimulus
 - Reactions from the infection caused by the attack
- GO:0009607
 - Defense response
 - Change in state or activity of a cell or an organism

GO Terms			
rs136766061	STAT6	GO:0006952	-0,0034
rs43773324	<i>IFI44L</i>	GO:0006952 GO:0009607	-0,0018
rs384409113	<i>GBP4</i>		0,0027
rs134587020 rs134221127	<i>PLCG2</i>		-0,0013 0,0011
rs380635364	<i>LOC511531</i>		-0,0027
rs109030124	<i>ADCY8</i>		-0,0010
rs133398278	<i>FYN</i>		-0,0016
rs385794234 rs455109043 rs380173423 rs379846168	<i>ENSBTAG00000054018</i>		-0,0027 0,0017 0,0011 0,0002
rs134436759	<i>KLRD1</i>		-0,0016
rs41957485	<i>CCT5</i>		-0,0057

RESULTS

QTL CATTLE

- Milk beta-lactoglobulin percentage
 - Potential antibacterial activity against mastitis agents

QTLcattle			
rs134380074	<i>ENSBTAG00000052898</i>	Milk beta-lactoglobulin percentage	-0,0006

RESULTS

MOST INFLUENTIAL

- RAP1
 - GTPase activation as a biomarker of infection-induced immune activation
- HMBOX1
 - Could play a role in the progression of Ketosis (and further cost-intensive diseases, e.g. mastitis)

Most influent SNPs			
rs135813968	<i>RAP1</i>	GTPase-GDP	-0,0095
rs133091490	<i>HMBOX1</i>	Ketosis	0,0073

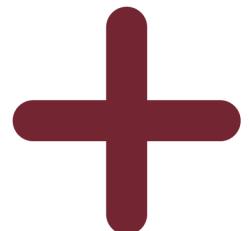
Take-Home messages

- Data with large numbers of features and a small number of observations ($p \gg n$) need pre-selection
- The varying classification accuracy demonstrates the importance of a proper feature
- Deep learning models allow not only good classification, but also the interpretation of complex biological processes
- Metrics like sensitivity and specificity should be taken under consideration in classification, especially in diseases classification

Thank You

 Wroclaw University of
Environmental and Life Sciences

 Email
krzysztof.kotlarz@upwr.edu.pl



THETA
Statistical Genetics Group
Institute of Animal Genetics

