Name and surname:	**Joanna Szyda**
Academic Degree:	prof. dr hab. inż. (Prof.)
Institute/Department:	Department of Genetics
e-mail address:	joanna.szyda@upwr.edu.pl
ORCID:	0000-0001-9688-0193
UPWr Base of Knowledge - link:	https://bazawiedzy.upwr.edu.pl/info/author/UPWr0b20a261b6864df1b93c3a1feb84e5cc?tab=main&conversationPropagation=begin&sort=&title=Person%2Bprofile%2B%25E2%2580%2593%2BJoanna%2BSzyda%2B%25E2%2580%2593%2BWroc%25C5%2582aw%2BUniversity%2Bof%2BEnvironmental%2Band%2BLife%2BSciences&lang=en&pn=1
Researchgate:	https://www.researchgate.net/profile/Joanna-Szyda
Personal website / Working group website:	theta.edu.pl
Participation in projects in last 5 years (chronological; with distinction into PI (kierownik) and RF (wykonawca)):	Title in English: Bioinformatic modelling of the impact of probiotic supplementation on microbiomes of breeding ponds and of digestive tract of the Common carp (Cyprinus carpio) Registration number: 2021/41/B/NZ9/01409 Source(s) of funding: NCN Name of the call: OPUS-21 Amount of funding: 1 184 760 PLN Entity's name in Polish: Uniwersytet Przyrodniczy we Wrocławiu Start date (yyyy-mm-dd): 2022-04-01 End date (yyyy-mm-dd): 2026-03-31 Project in progress == Title: The application of deep learning methods in the analysis of livestock genomes Keywords: CNV; classification; DNA-seq; imputation; machine learning; neural networks; SNP; == Title: Biodiversity within and between European Red dairy breeds – conservation through utilization Registration number: 696231 Source(s) of funding: Horizon 2020 (ERANet - SusAn programme) Amount of funding: 1 790 000 EUR Entity's name: Kiel University, Kiel, Germany Start date (yyyy-mm-dd): 2017-09-01 End date (yyyy-mm-dd): 2021-09-30 == Title: European Network on Livestock Phenomics Registration number: European Network on Livestock Phenomics Source(s) of funding: CA22112 Amount of funding: 125 000 EUR Entity's name: University of Bologna
PhD topic:	Small Samples, Big Decisions: AI Algorithms vs. Conventional Statistical Models in Multidimensional Data Analysis
Research discipline in Doctoral School:	Biological Sciences
Short description of the research problem to be solved in the PhD (minimum 1000 characters):	Artificial Intelligence (AI) methods have recently gained a lot of attention, both within the research community and in the public. This is related to their flexibility to handle high-dimensional data structures that are often characterised by many, highly correlated features. Another group of methods that can be applied for the analysis of high dimensional data is the family of mixed models, in which the features' correlation is handled by the covariance matrix of random effects. Also, feature selection is a fast-growing field of research, which tackles the high-dimensionality problem by fitting multiple models, each considering only a subset of all available explanatory covariates. Such high dimensional data is typical also for the field of genomics in which, thanks to the development of high-throughput technologies, the past few decades have seen a considerable increase in the availability of genomic, metagenomic, and epigenetic data. Consequently, the most serious drawback underlying the statistical modelling of genomic data is the so-called p>>n problem, where the number of features (p) is much higher than the number of available individuals (n). Although AI-based data modelling does not technically suffer from the p>>n problem, the most spectacular performance of AI-based approaches has been observed for large data sets that contain sufficient information for training of the algorithm. While a small data set accompanied by a very large number of features poses a problem for the accurate estimation of numerous network parameters typically underlying the AI-based algorithms. Therefore, the goal of this project is to compare the performance of conventional statistical models with AI-based algorithms for classification and prediction purposes, for the situation of high-dimensional genomic data sets suffering from the p>>n problem of different degrees of severity. The comparison will be based on the following data sets: (1) 1,067 persons, each with 43,49928 SNP genotypes and phenotypes corresponding to classes representing different responses to COVID-19 infection, (2) 600 bulls each with 72,015,48 SNP genotypes, pedigree relationship data, and quantitative pseudophenotypes for protein yield together with their precision, (3) 232 persons representing 114 controls and 118 cases diagnosed with liver cirrhosis, each with the abundance of 185 genera from the gut microbiome, and (4) 88 persons each with 1,194,548 counts of methylated sequence reads at cytosine sites and phenotypes corresponding to classes of early and late stress as well as the quantitative cumulative life stress score. Each of the data sets will be analysed by three approaches: (1) AI-based approach, (2) mixed model-based approach, and (3) a series of models fitting a subset of available covariates at a time with a post-estimation feature selection. Model comparison will cover the accuracy of classification or prediction (depending on the data set) and computational efficiency, expressed by programme execution time and peak memory usage. In particular, AI algorithms will be implemented as deep learning algorithms. Various network structures, including sets of dense neural networks and convolutional neural networks, with an optimal architecture comprising the number of layers and neurons per layer, the dropout rate within each layer, and the learning rate for the optimisation algorithm will be tested. The mixed models will be applied either as linear models (i.e. without transformation of the input data) or as generalised mixed models. Both types of mixed models will imply a shrinkage of the multiple features by fitting them as random effects following a pre-imposed distribution with either estimated or assumed distribution parameters. In the Supervised Rank Aggregation approach, a number of models with a subset of the available covariates will be fitted, then the final model incorporates the goodness-of-fit of the submodels as a dependent variable and effects of all. Finally, covariate effects will be classified by 1-dimensional clustering. The team is made up of experts with complementary expertise in the bioinformatic analysis of whole-throughput molecular data, in the application of machine learning models, in statistical modelling of multi-omic data and the PI who is an expert in using mixed models as well as in the management of research projects. The team has conducted projects together, which is the prerequisite for an effective collaboration.
Professional skills for PhD candidate (e.g. master program, specializations, softwares, language, analytical techniques, minimum 500 characters):	The applicant is expected to be fluent in scripts like Python and R; in writing command line scripts on Linux using Bash; in writing deep learning workflows using the keras API or possibly directly using the Tensor Flow library; fluent in using English language in writing and speaking. Intermediate level of programming in either Fortran or C, as well as intermediate knowledge of mathematical statistics. Additional asset is the basic knowledge of script management systems like e.g. the Nextflow and using the parallelization techniques, in particular OpenMP.
a) Project title:	none
b) Agreement number:	none
c) Number of months in the project to support PhD student (in months; starting from 1st of October 2024):	
Project website:	