Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Satyarth Mishra Sharma

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Predicting phenotype from genotype with machine learning

Supervisor: Professor Dmitry Yarotsky

Name of the Reviewer:

I confirm the absence of any conflict of interest Prof. Elena Potokina	**Date: 11-12-2024**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications
- The summary of issues to be addressed before/during the thesis defense

- Brief evaluation of the thesis quality and overall structure of the dissertation.

The Doctoral thesis "Predicting Phenotype from Genotype with Machine Learning" evaluates quite diverse aspects of genotype-phenotype relationships, including even ethical aspects related to access to genomic data that have private and sensitive nature. The dissertation is written in a good, clear style, which allows even a non-expert in data science to understand the main idea of each chapter.

Chapter 1 provides some basics about the quantitative genetics, describes principles of classical GWAS approach, highlights the most direct application of phenotype prediction, such as genomic selection and precision medicine.

Chapter 2 describes original ML-based approaches for predicting human phenotypes from genotypes and compares them with predictive models that are based on multiple linear regression. Linear Regression is typically performed through genome-wide association study (GWAS) methods which identify SNPs most significantly associated with the phenotype. However, non-linear effects such as epistasis contribute significantly to some phenotypes and can be more effectively detected using machine learning methods that account for nonlinearity.

Using subset of the UK Biobank database combining genetic, environmental, and healthcare data, the author proposed a new approach for SNP selection, using a XGBoost models, which, compared to the conventional GWAS-based SNP selection, generated a more informative and less correlated set of SNPs. Furthermore, it was shown that when sufficiently many covariates are included as features for predictions, prediction accuracy is consistently improved using nonlinear models. For 5 phenotypes analyzed (height, eBMD, hypothyroidism, asthma, psoriasis), the most accurate predictive models were obtained by ensembling both linear (Snpnet) and nonlinear (XGBoost) models.

Chapter 3 introduces a Federated Learning approach for phenotype prediction that allows multiple clients to collaboratively train a model without sharing data directly, reducing the risk of disclosing sensitive information.

Chapter 4 describes some new ways to predict the consequences of mutations that alter protein stability and affect phenotype (e.g., health and disease). The neural network architecture was proposed and implemented in a software package called ABYSSAL.

Chapter 5 demonstrate how to use machine learning techniques to improve the accuracy of Fetal Variant Calling when performing genetic testing for prenatal diagnostics. Again, results obtained from the study, were implemented in a software package for genotyping correction in the presence of Maternal Cell Contamination (MCC).

- The relevance of the topic of dissertation work to its actual content

I believe that the topic of the dissertation corresponds to its actual content. In fact, the chapters of the dissertation reflect the successive stages of the classical study of the relationship between genotype and phenotype: clean the data from contamination (Chapter 5), find the best and most accurate predictive models (Chapter 2), and, upon detection of possible causal mutations, try to assess their dramatic impact on the encoded protein (Chapter 4).

- The scientific significance of the results obtained and their compliance with the international level and current state of the art

The dissertation highlights some interesting aspects of the approach to predicting phenotype from genotype. We typically include in our linear regression models a limited number of factors whose relevance to the phenotype under analysis is obvious (e.g. age and sex are obviously relevant for height variation). The author showed that including additional traits with uncertain effects on the phenotype can improve the predictive power if a nonlinear model is used.

Prediction of protein stability changes upon mutation using the newly developed software (ABYSSAL) is undoubtedly of interest to the scientific community, since prediction of the consequences of mutations is a key issue for many researchers working on the genetic basis of quantitative and qualitative traits.

- The relevance of the obtained results to applications (if applicable)

The most accurate ML models proposed for human genotype-to-phenotype prediction will mean more accurate and effective strategies for diagnosis, treatment and prevention in medicine. Furthermore, the proposed ML algorithms can be tested to predict complex plant phenotypic traits such as yield or disease resistance, since they capture nonlinear relationships between genetic variation and environmental factors extracted from datasets.

- The quality of publications

The results were published in three papers in highly ranked journals: European Journal of Human Genetics, PloS One, Frontiers in Big Data (the paper, titled "Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project" was mentioned in dissertation as preprint. It has now been published.)

- The summary of issues to be addressed before/during the thesis defense

I have some comments:

1) 1.3.1 GWAS & Summary Statistics.

Describing GWAS approach, the Author mentioned that "The loci are then filtered according to the statistical significance of the detected association, typically using a p-value threshold. These can then be analyzed downstream with the aim of identifying causal variants".

The "downstream" story is the major bottleneck for most GWAS-based research, at least in plants. It would be useful to mention at least some of the possible solutions for detecting causal variants.

2) page 18. "For example, microarray based genome-wide association studies (GWAS) can miss a large proportion of rare variants, or significant variants in populations differing from the ones considered during the design of the microarray".

I believe that missing a large proportion of rare variants is a problem inherent to the GWAS approach in general, regardless of the specific genotyping technology. Rare variants with MAF < 5% are usually excluded from analysis because they cannot be used for linear regression models. Thus, GWAS basically excludes rare (possibly valuable) variants.

3) page 35. "XGBoost starts to outperform Snpnet with a limited number of SNPs only when adding 19 more covariates." Unfortunately, I could not find any details about these 19 additional covariates. Meanwhile, it is interesting what type of features with a non-obvious impact on the phenotype can help to capture the nonlinear effect in the analysis.

Provisional Recommendation

☒ ***I recommend that the candidate should defend the thesis by means of a formal thesis defense***
☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*
☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*