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: Professor Alexey Zaytsev

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 17-12-2024**

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

The overall quality and structure of the thesis are satisfactory. A significant amount of work was performed during its preparation, resulting in a substantial volume of content. Given the broad scope of the topic, the material presented is indeed relevant. Each chapter focuses on a specific application of machine learning in genomics and is supported by at least one publication. While the technical work in each chapter meets appropriate standards, the thesis lacks cohesion and does not sufficiently relate the individual studies to one another. It would be beneficial for the candidate to strengthen these connections, particularly in the introduction and conclusion. However, achieving a fully coherent narrative would likely require a substantial rewrite of the thesis.

The scientific significance of the results varies. Some chapters demonstrate state-of-the-art performance using relevant methodologies, while others apply standard methods to new datasets with limited novelty. For example, the core methods in Chapter 3—Federated Learning on genomic data—primarily rely on the FedAvg strategy and the P-STACK approach proposed in 2017 and 2021, respectively.

I had anticipated a deeper investigation of the proposed methods, but found limited evidence of this in the thesis. For instance:

- In Chapter 2, the author introduces an SNP selection scheme but does not compare it to other potentially simpler baselines. A straightforward alternative might be to train a single large model and extract SNP feature importance from XGBoost. Repeating this estimation multiple times with different data subsamples and averaging the results would be another common and sensible approach. The proposed window-based strategy warrants justification: Is this approach guided by domain intuition or theoretical underpinnings? These alternative methods also avoid the known pitfalls of GWAS-based SNP selection and would align more closely with standard community practices.
- Chapter 3 adopts a common GWAP procedure without leveraging the SNPnet framework introduced in the previous chapter.
- Additional references would enhance the thesis. For example, the author does not mention [1], which provides a consistent approach to LASSO regression in a federated setting, instead opting for a more rudimentary method.

The text would benefit from further detail and clarification in several places. For example, in Chapter 3 the term “meta-GWAS” is mentioned on page 56, but is not defined there or elsewhere in the methods section, and the same issue applies to “local GWAS.”

On a positive note, the open-source code accompanying the research makes the work accessible to the broader scientific community. The publication requirements are met through two shared first-author papers by Satyarth, one of which appears in a Q1 journal.

Because I am not an expert in the specific applications of machine learning presented in this thesis, I may have overlooked certain significant contributions or issues.

[1] Smith, Virginia, et al. "CoCoA: A general framework for communication-efficient distributed optimization." Journal of Machine Learning Research 18.230 (2018): 1-49.

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*