Benchmarking predictive models: evaluating parametric, ensemble, and deep learning approaches for animal phenotype prediction from genotypes.

Eric Barrey, Blaise Hanczar, Julien Chiquet, Didier Boichard, Jocelyn de Goër de Herve, Anne Ricard, Thierry Tribout, Joon Kwon, Jean-Benoist Leger, Tristan Mary-Huard, et al.

AI and biology

12 – 15 March 2024 | EMBL Heidelberg and Virtual

46

Benchmarking predictive models: evaluating parametric, ensemble, and deep learning approaches for animal phenotype prediction from genotypes

Eric Barrey[2], Thierry Tribout[7], Blaise Hanczar[3], Ronny Tonatto[1], Fatima Shokor[7], Jianshu Zhu[8], Joon Kwon[4], François Victor[4], jean-benoist Leger[4], Tristan Mary-Huard[4], Beatriz Castro Dias Cuyabano[7], Pascal Croiseau[7], Anne Ricard[6], Jocelyn De Goer de Herve[5], Didier Boichard[7], Julien Chiquet[4], Sihan Xie[7]

[1] Université Paris-Saclay, AgroParisTech, INRAE, GABI UMR1313, Jouy-en-Josas, France., France
[2] GABI-UMR1313, INRAE, AgroParisTech, Université Paris-Saclay, France
[3] IBISC, UEVE, Université Paris-Saclay, France., France
[4] MIA-PS, Université Paris-Saclay, AgroParisTech, INRAE, France., France
[5] UMR EPIA, Centre de recherche Auvergne-Rhône-Alpes – Theix, France., France
[6] Université Paris-Saclay, AgroParisTech, INRAE, GABI UMR1313, Jouy-en-Josas and IFCE, Exmes, France., France
[7] Université Paris-Saclay, AgroParisTech, INRAE, GABI UMR1313, Jouy-en-Josas, France, France

Presenter: Sihan Xie

Over the past two decades, advancements in DNA genotyping have revolutionized the acquisition of extensive genomic data, essential for enhancing genomic selection in animal breeding across multiple traits. Concurrently, the rapid progress of artificial intelligence has piqued interest in its application to genomic selection. Despite its potential, the utilization of ensemble methods and deep learning models for predicting livestock animal phenotypes based on genotypes lags behind parametric GBLUP mixed models, which are regarded as the current state of the art. The primary objective of this study was to conduct a benchmark analysis, comparing various parametric, ensemble, and neural network methods for predicting animal phenotypes using extensive genotyping data.

The dataset employed for training and validating different models comprised 100,000 Holstein cows characterized for 33 quantitative traits (phenotypes: milk production, fertility and morphology category). Genotyping data (50K SNP/animal) were utilized to predict the 33 traits, employing parametric mixed model GBLUP, as well as gradient boosting (GB) and various deep learning models, including MultiLayer Perceptrons with various architectures (MLP), Convolutional Neural Networks (CNN), CNN Transformer, Value Imputation and Mask Estimation (VIME), and MLP Variational AutoEncoder (VAE). To assess predictive model performance, we employed the same data for training, validation, and evaluation sets, measuring accuracy through root mean square error (RMSE) and Pearson correlations (R) between actual and predicted traits.

According to the correlation coefficient (R), GBLUP (R=0.38) outperformed all other models in predicting most of the traits. However, on average, MLP performed slightly less effectively (R=0.38) than GBLUP, with no significant difference ($p>0.05$). The remaining models (GB, CNN, CNNT, VIME+MLP, VAE+MLP) demonstrated lower efficiency. Conversely, based on RMSE, certain traits (2 to 9 traits of milk production and fertility) were more accurately predicted by GB or MLP models, never by more complex models (CNN, CNNT, VIME+MLP, VAE+MLP). Further refinement of deep learning models specifically tailored to genomic data is essential for improved prediction of traits determined by polygenic gene effects.

Benchmarking Predictive Models: Evaluating Parametric, Ensemble and Deep Learning Models for Animal Phenotype Prediction from Genotypes

Sihan Xie [1], Thierry Tribout [1], Jianshu Zhu [2], Ronny Tonatto [1], François Victor [2], Fatima Shokor [1], Anne Ricard [1,5], Joon Kwon [2], Jean-Benoist Léger [2], Tristan Mary-Huard [2], Beatriz Castro Dias Cuyabano [1], Pascal Croiseau[1], Jocelyn De Goer de Herve [4], Didier Boichard [1], Julien Chiquet [2], Blaise Hanczar [2], Eric Barrey [1]

[1] Université Paris-Saclay, AgroParisTech, INRAE, GABI-UMR1313, 78350 Jouy-en-Josas, France
[2] MIA-PS, Université Paris-Saclay, AgroParisTech, INRAE, France
[3] IBISC, UEVE, Université Paris-Saclay, France
[4] UMR EPIA, Centre de recherche Auvergne-Rhône-Alpes – Theix, France
[5] IFCE, Recherche et Innovation, Exmes, France

Eric.barrey@inrae.fr

INTRODUCTION

Shortcut

Background

- **Revolution in Genomics**: the past two decades have witnessed significant advancements in DNA sequencing and high identity genotyping technologies.
- **AI in Genomics**: the rapid progress in artificial intelligence has sparked interest in its application to genomic selection (1,2).

Objective

Our study aims to perform a **model benchmark analysis**.
We compare the effectiveness of various modeling approaches—mixed-effect model (GBLUP), ensemble algorithms (Gradient Boosting), and neural networks (MLP, CNN, VIME+MLP, Transformer+MLP, VAE+MLP)—in predicting animal phenotypes based on extensive genotyping data.

CONCLUSION

- **GBLUP Dominance:** the traditional GBLUP model remains the most effective for predicting livestock genetic value of the traits from genomic data.
- **MLP's Potential**: the Multi-Layer Perceptron (MLP) model, with an optimized architecture, closely approaches the effectiveness of GBLUP in predicting certain traits.
- **Performance of Other Models**: while Gradient Boosting (GB) and MLP models showed superior performance for a small subset of traits, more complex deep learning models did not outperform the GBLUP.

References

- [1] Bellot, P., de los Campos, G., & Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics*, *210*(3), 809–819. https://doi.org/10.1534/genetics.118.301298
- [2] Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(1), 12. https://doi.org/10.1186/s12711-020-00531-z

MATERIAL & METHODS

Data

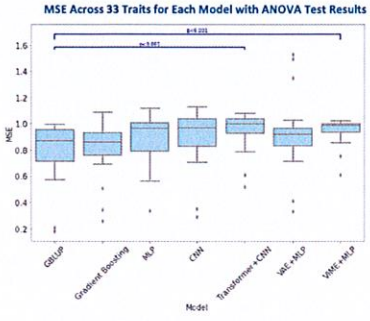
- **Animals & Phenotypic traits**
  We used a real dataset consisting of 113,599 French Holstein cows, with records for 33 quantitative traits associated with milk production, measured by the Yield Deviation.
- **Genotypes**
  Genome-wide SNP data were obtained using Illumina SNP-chip. 53,469 biallelic SNPs were selected and they were coded as 0 for homozygous major allele, 1 for heterozygous, and 2 for homozygous minor allele.

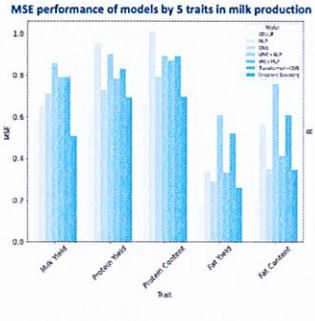
Training & Evaluation

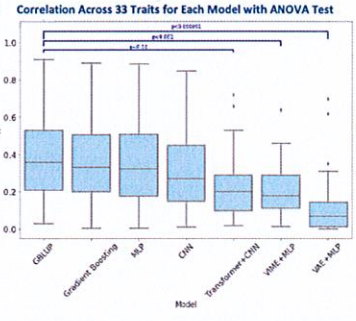
- **Data Splitting**
  We divided the data into 3 sets: training (93,484 cows), validation (10,086 cows), and test (10,015 cows), ensuring no half-sisters were shared between sets. The test set includes the latest generation of offspring.
- **Model Evaluation**
  We calculated the Pearson Correlation Coefficient (PCC) and the Mean Squared Error (MSE) between the actual and predicted Genomic Estimated Breeding Values (GEBV).

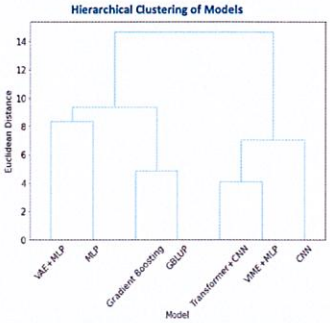
Key words: IA, Deep Learning, SNP, genomics, cow, milk production

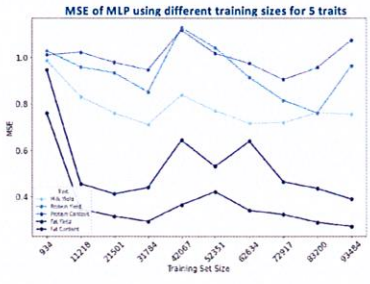
RESULTS











Results in a nutshell

- There are significant differences among the models. The models that perform better in terms of MSE are not necessarily better in terms of correlation. In terms of overall performance, **GBLUP > Gradient Boosting ≥ MLP > Other models**, especially for traits with high heritability (i.e. large additive genetic effect).
- MLP stands out for its efficiency, as it has a simple architecture and can perform multi-trait predictions simultaneously.
- By performing feature selection to eliminate less important SNPs, or by augmenting the dataset with additional SNP data in the training set, the model's performance can be enhanced.
- Utilizing neural networks to predict phenotypes solely from SNP data presents challenges. Enhancing performance will necessitate refining models and incorporating biological pedigree or genomic information.