Using convolutional neural networks for image-based genomic prediction in mice

B.C. Perez[1*], A. Savchuk[1], P. Duenk[2], M.P.L. Calus[2], M.C.A.M. Bink[1]

[1] Hendrix Genetics, P.O. Box 114, 5830 AC Boxmeer, the Netherlands; [2] Wageningen University & Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [*]bruno.perez@hendrix-genetics.com

Abstract
Convolutional neural networks (CNN) are well suited image recognition tools for their ability to recognize latent patterns from images. Here, we investigated whether CNN can be used for genomic prediction. We created genomic images from genotype data and used them to predict phenotypes in mice. This approach was compared with traditional GBLUP and gradient boosting machine (GBM) models. For the two traits analysed, CNN was competitive in terms of predictive performance. The resolution of genomic images impacted model performance where, for this dataset, optimum results were obtained with 100x100 pixels. These first results demonstrate the potential of genomic images for genomic prediction using CNNs and merit investigation on adding layers of information to further increase accuracy of prediction.

Introduction
In recent years, machine learning algorithms have appeared as promising tools for genomic prediction. These algorithms do not require prior assumptions about the underlying genetic architecture of traits, being capable of capturing these patterns from training data (Pérez-Enciso and Zingaretti 2019). For these reasons, machine learning algorithms could yield better predictive performance than traditional linear models like GBLUP. Convolutional Neural Networks (CNN; fully connected layers and convolutional/pooling filters) are algorithms that are widely used in image recognition tasks for their ability to identify latent patterns and features from images (Trevisan *et al*. 2020). Visual representations of genotype data as predictors could help to improve performance of CNN models for genomic prediction of complex phenotypes (Galli *et al*. 2021). The objective of this research was to investigate the predictive performance of CNN models using genomic information transformed into images, benchmarked with a parametric linear method and an alternative machine learning method.
This project is part of EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global Functional Annotation of ANimal Genomes (FAANG) objectives.

Materials & Methods

Diversity Outbred (DO) Mouse dataset
The DO mice data comprising 835 animals were obtained from The Jackson Laboratory (Bar Harbor, ME). The animals originated from 6 non-overlapping generations (4, 5, 7, 8, 9 and 11) in which males and females were represented equally. Based on anticipated differences in genetic architecture of traits, we selected circulating cholesterol at 12 weeks (CHOL) and fat percentage at 19 weeks (FATP, Table 1). Prior to the analyses, phenotypes were pre-corrected for fixed effects of generation, sex, cage, and diet. All animals had genotype data available for 50,122 SNP markers.

Table 1. Trait abbreviation, number of records, heritability estimate, and anticipated genetic architecture of circulating cholesterol and fat percentage.

Trait	No. records	Heritability	Genetic architecture
CHOL	832	0.29	Evidence of epistasis
FATP	834	0.37	Highly polygenic

Genotype data formats

We transformed genotypes from tabular format into a genomic image using the DeepInsight algorithm (Python implementation, available at http://www.alok-ai-lab.com) proposed by Sharma *et al*. (2019). This algorithm applies a similarity measuring technique, here we used t-SNE, to obtain a "topology graph" based on the similarity between genomic markers. In this resulting graph, each point represents a marker. If two or more markers are strongly related due to linkage disequilibrium patterns, these will be placed at similar coordinates. After that, this graph is transformed into an image, one uniquely per individual, and the genomic marker information, e.g., 0, 1, or 2, is mapped to its corresponding pixel position (Figure 1). At this stage, each pixel in the genomic image may contain information from one or more SNP markers (in case these were collapsed due to high similarity) and colours for each pixel indicate A1A1 (light grey, 0), A1A2 (medium grey, 1) and A2A2 (dark grey, 2) genotypes, or no information (black). We compared three different image resolutions: 50x50 pixels (CNN50), 100x100 pixels (CNN100 and 150x150 pixels (CNN150) (Figure 1).

Figure 1. Illustration of genomic images for a single individual using resolutions of 50x50 (*left*), 100x100 (*middle*) and 150x150 (*right*) pixels, reshaped to a common format for visual comparison.

Genomic best linear unbiased prediction (GBLUP)

For its well stablished and wide use in routine genomic prediction, GBLUP was used as a benchmark for comparisons. The implementation of this model was performed using the BGLR package (Perez and de los Campos, 2014).

Gradient Boosting Machine (GBM)

The GBM model is an ensemble learning technique that combines gradient-based optimization and boosting techniques (Friedman, 2001). It applies an iterative process of assembling "weak learners" and therefore, can be expressed as a linear combination of multiple models. The algorithm is also able to prioritize informative predictors while ignoring uninformative ones. Optimal hyperparameters were defined using a grid search (described in Perez *et al*. 2021) on an inner validation comprised of a random sample of 20% observations from the original training set. Results reported come from the best performing model for each trait analysed. The GBM model was implemented using the h2o.ai R package (Click *et al*. 2016).

Convolutional Neural Networks (CNN)

The input layer for the CNNs consisted of the pixel matrix of the genomic image for an individual. A basic network architecture was proposed (described below) and kept the same for both traits. At trait level, a simple hyperparameter search was performed on an inner validation set (20% of the training set) on one replicate, and the best combination of hyperparameters (lowest mean squared error) from this one search was used for all other tested resolutions in 10 replicates. For the network architecture chosen, the input data was first passed through three convolutional layers, followed by a maximum pooling layer, a dropout layer, a flattening layer, a dense (*i.e.*, fully connected) layer and finally to the output layer containing one node with the predicted trait value. We used the rectified linear activation unit (ReLU) as the activation function in the hidden layers. To reduce the time and memory requirements, models were trained using a fixed batch size (32) and run for a maximum of 400 epochs. The CNN were implemented in Python 3.6 using Tensorflow 2.0 and Keras libraries (Gulli and Pal, 2017).

Predictive performance

Predictive performance was determined with a forward-prediction validation scheme, in which models were trained using data from older individuals (generations 4,5,7,8 and 9; N=643) and prediction was carried out for younger animals (generation 11; N = 192). The predictive performance was evaluated by the correlation between model solutions and pre-corrected phenotypes on validation animals (hereafter termed accuracy). For GBM and CNN, for each trait results were reported for the models that obtained best results in the model-specific grid search analysis previously mentioned.

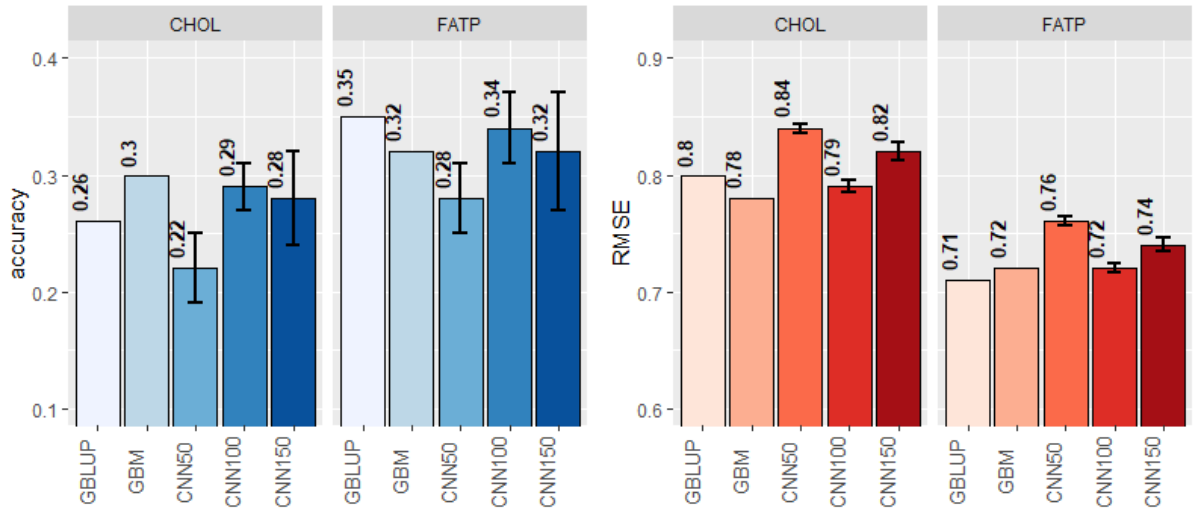

Figure 2. Predictive accuracy & RMSE for GBLUP, GBM and CNN models on genomic images of 50x50 (CNN50), 100x100 (CNN100) and 150x150 (CNN150) pixels.

Results

Predictive accuracies varied from 0.221 (CNN50) to 0.301 (GBM) for CHOL and from 0.285 (CNN50) to 0.352 (GBLUP) for FATP (Figure 2). For the root mean-squared error (RMSE), results varied from 0.78 (GBM) to 0.84 (CNN50) for CHOL and from 0.71 (GBLUP) to 0.76 (CNN50) for FATP. The resolution of the genomic image had an impact on both predictive accuracy and error, with CNN100 showing the best results for both traits. The CNN50 yielded poorest results, while CNN150 performed slightly worse than CNN100.

Discussion

In the present study our current CNN model was slightly outperformed by either GBM or GBLUP for CHOL and FATP, respectively (Figure 2). Nevertheless, our results show the potential of using genomic images for phenotype prediction using CNN models as prediction performance for CNN was better than GBLUP for CHOL and better than GBM for FATP. It must be noted here that for time/computational constraints, the network architecture was kept simple and the hyperparameter tuning for the CNN model was limited to a simple search. A more detailed procedure could help to improve performance of CNN models. In a case study in maize hybrids, Galli *et al*. (2021) observed superior predictive performance for CNN (using genomic images) over GBLUP (relationship matrix). A common limitation in both studies was the size of datasets (N = 904 in Galli *et al*. (2021); N = 835 in the present study). Future studies using larger datasets are required to assess performance of genomic prediction using images on routine of breeding programs or other commercial applications.

The resolution of genomic images had major impact on predictive performance. In the present study, the resolution of 100x100 pixels (CNN100) yielded best results and weakening results were found for CNN50 and CNN150. This indicates that there is an optimum image resolution, most likely dependent on the number of SNP genotypes available for prediction and the size and structure of the population studied. For a fixed number of variants, lower resolutions will merge too much information making it harder for the model to capture some patterns in data. Higher resolutions on the other hand can accommodate detailed and complex patterns but will require much bigger training sets to achieve a good predictive performance. These encouraging results and the flexibility of CNN models deserve further investigation to include additional layers of information, such as genome annotations, that can increase accuracy of prediction.

Acknowledgments

The GENE-SWitCH project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 817998.

References

Click C., Malohlava M., Candel A., Roark H., and Parmar V. (2016) *Gradient Boosted Models with H2O*. Available at: "http://h2o-release.s3.amazonaws.com/h2o/master/3568/docs-website/h2o-docs/booklets/GBMBooklet.pdf"

Galli G., Sabadin F., Yassue R.M., Souza C.G., Carvalho H.F., *et al*. (2021) Research Square. https://doi.org/10.21203/rs.3.rs-840380/v1

Gulli A. and Pal S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.

Friedman, J. H., (2001) Ann. Stat. 29:1189-1232.

Perez B.C., Bink M.C.A.M., Churchill G.A., Svenson K.L., and Calus M.P.L. (2021) bioRxiv 2021.08.02.454826. https://doi.org/10.1101/2021.08.02.454826

Pérez P. and de los Campos G. (2014) Genetics. 198:483-495. https://doi.org/10.1534/genetics.114.164442

Pérez-Enciso M. and Zingaretti L.M. (2019) Genes 10:1–19. https://doi.org/10.3390/genes10070553

Sharma A., Vans E., Shigemizu D., Boroevich K.A., and Tsunoda, T. (2019) Sci Rep 9:1–7. https://doi.org/10.1038/s41598-019-47765-6

Trevisan R.G., Pérez O., Schmitz N., Diers B., and Martin N. (2020) Remote Sensing 12:3617. https://doi.org/10.3390/genes10070553