Predicting wheat yield from genotypes and environmental data using four machine learning approaches

Hawlader Al-Mamun, Rob Dunne, Bill Bovill, Colin Cavanagh, Klara Verbyla

The ability to predict yield of untested genotypes has been a goal of plant breeding programs for many decades. In recent years, the volume of data with potential to improve prediction accuracies has increased substantially. Machine learning has been shown to be a valuable tool for many prediction problems involving Big Data. Using genotypes from multiparent advanced generation intercross (MAGIC) wheat populations and dense environmental data, we compared the predictive accuracy of deep neural networks (DNN) and three state-of-the-art decision tree-based methods: random forest (RF), gradient boosting machine (GBM) and model tree (MT) for yield prediction.

Methods

The data included genotypes from CSIRO's 4-way and 8-way MAGIC wheat populations and environmental data. MAGIC populations are created using multiple inbred founders that are intercrossed several times (Fig 1). This process creates highly diverse genotypes, each with a unique mosaic of founder alleles. These populations were grown and phenotyped in four consecutive years at three different locations in Australia.

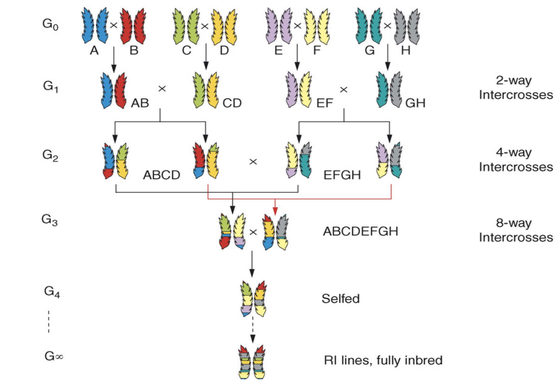

Figure 1: MAGIC populations are created by inter-crossing *n* lines for *n/2* generations until all founders are combined with equal proportions in the inter-crosses

The genomic dataset contained 1,481 4-way lines and 2,753 8-way lines with 30,686 SNP.

The environmental dataset contained rainfall (mm) and, maximum and minimum daily temperatures (°C). Features were constructed based upon the number of growing degree days (GDD) between the date of sowing and the eleven different growth stages. Features reflecting extreme weather conditions were created, for example, the number of GDD with a maximum temperature greater than 28^0C.

The phenotypic data used was yield (tonnes per hectare) after year and location were accounted for. The residuals were used as input phenotype after removing the outliers.

Approach: For each of the ML method, hyper-parameter tuning was performed via a nested grid search within 5-fold cross validation (CV) framework. Then the highest performing hyper-parameters were used for comparing the performance of ML methods in 10-fold CV framework.

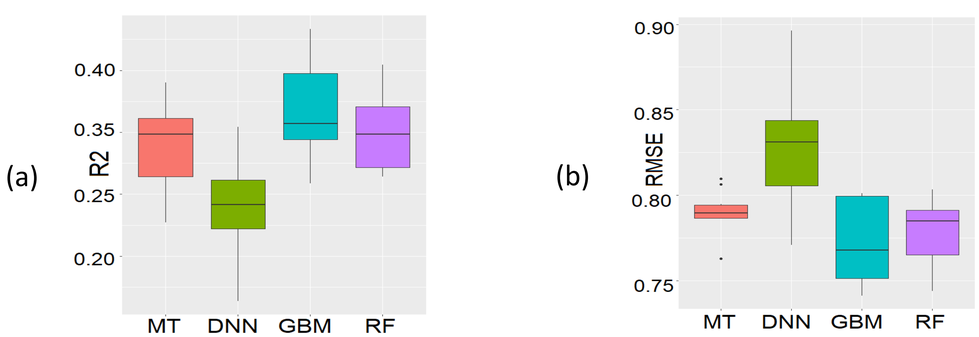

Figure 2: Comparisons of (a) squared correlation (r^2) and (b) RMSE of 10-fold cross validation for 4-way genotype plus the weather data

Results and discussion

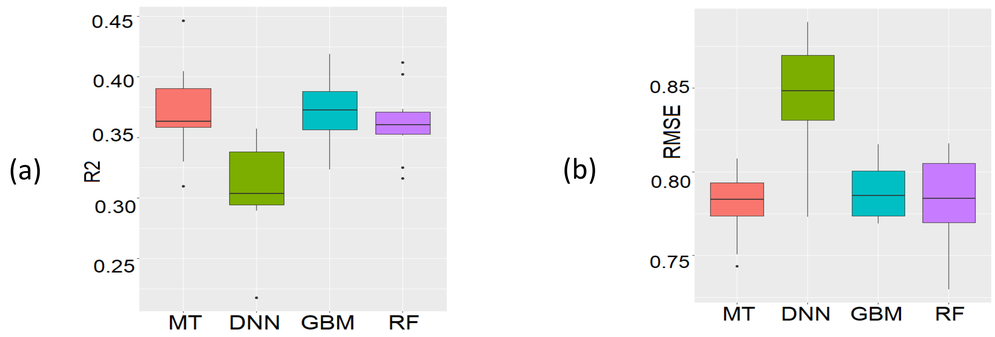

Figure 3: Comparisons of (a) squared correlation (r^2) and (b) RMSE of 10-fold cross validation for 8-way genotype plus the weather data

- **4-way genotypes and weather data (Fig 2):** the best predictive accuracy using squared correlation (r^2) was obtained with GBM (0.37), followed RF (0.35), MT (0.34), and DNN (0.31).
- **8-way genotypes and weather data (Fig 3)**: GBM and MT performed similarly with the squared correlation (r^2) of 0.37 followed by RF (0.36), and DNN (0.30).
- In both datasets, the performance of DNN was significantly worse than the other three methods tested; the accuracies obtained were significantly lower and root mean square error was higher (Fig 2 and Fig 3)
- The findings that GBM performed well for the prediction of yield is consistent with Abdollahi-Arpanahi et al.[1] where the authors reported that GBM is a robust method in genomic prediction of complex traits.
- Hyper-parameter tuning proved crucial. This tuning was challenging for DNN and the performance of DNN was found to be significantly affected by changes in hyper-parameters.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

FOR FURTHER INFORMATION
Hawlader Al-Mamun
hawlader.almamun@csiro.au
https://data61.csiro.au/

REFERENCES
[1] Abdollahi-Arpanahi, R., Gianola, D. & Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol 52, 12 (2020)