CNN for Prediction of Yield Component Traits in Wheat from Transcriptome Data

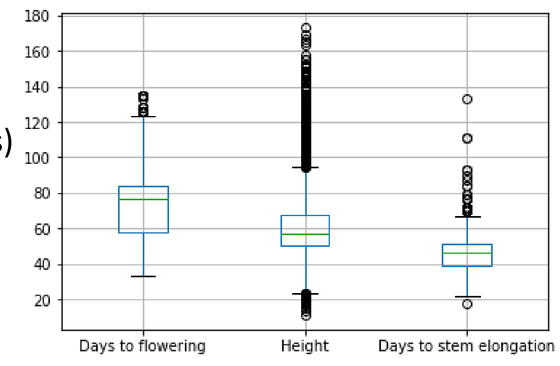
Md Zakir Hossain, Jessica Hyles, James A. Broadbent, and Shannon Dillon; CSIRO Agriculture & Food

Global food security is a critical challenge in agriculture today. Deep learning (DL) approaches have potential to provide improved avenues for accelerated genetic gain in crops including wheat, a major source of human nutrition. Strategies to extract value from the increasing availability of 'omic data need to be identified to support production and breeding opportunities. We implemented a DL method, Convolutional Neural Network (CNN), to predict important components of wheat yield, from transcriptome data collected on ~300 varieties across two controlled environments.

Dataset

The OzWheat dataset includes 44,566 transcripts abundances and 9 components of wheat yield (listed below), collected in two controlled environments (long days and short days).

1. Days to flowering (counts)
2. Height (mm)
3. Spike length (mm)
4. Days to stem elongation (counts)
5. Leaf number (counts)
6. Spikelets per spike (counts)
7. Empty spikelets (counts)
8. Tillers (counts)
9. Coarse grain mass (gm)



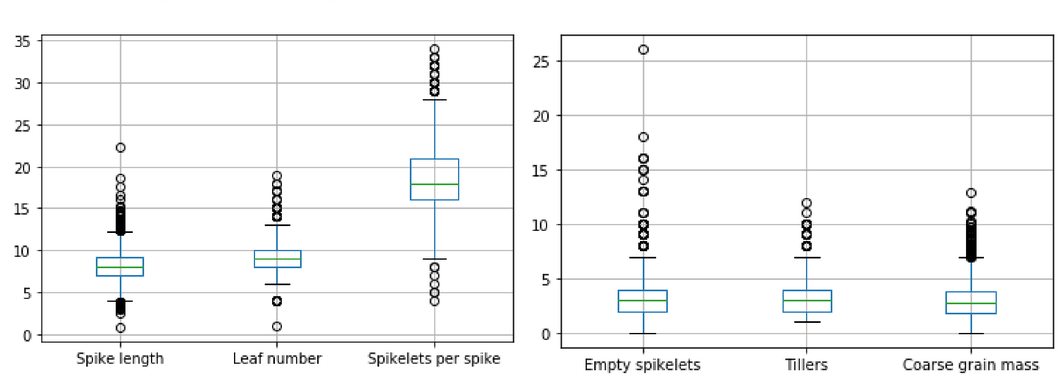


Figure 1: Read counts distribution of each wheat trait.

Pre-processing

A normalisation approach is introduced to transcripts abundances between [0, 1] and least square linear regression ($p \leq 0.01$) is considered to select a number of important transcripts (**Table 1**).

Table 1: [least square linear regression to select number of transcripts abundances

TRAITS	OBSERVATIONS	TRANSCRIPTS
Overall	**3,420**	**44,566**
Days to flowering	2,738	25,708
Height	2,990	19,574
Spike length	2,995	19,625
Days to stem elongation	2,521	23,130
Leaf number	2,615	9,811
Spikelets per spike	2,988	17,743
Empty spikelets	2,952	5,041
Tillers	3,023	16,287
Coarse grain mass	2,929	10,958

CNN Model

Convolutional neural network (CNN) is a special case of artificial neural networks where different layers (**Figure 2**) are connected together to perform the convolution operation along with an input of predefined width and strides. Here, a CNN model is implemented with a rectified linear unit activation function, Adam optimiser, and 10-fold cross validation strategy. A grid search approach is adapted to select hyperparameters, mainly number of epochs and batch size.

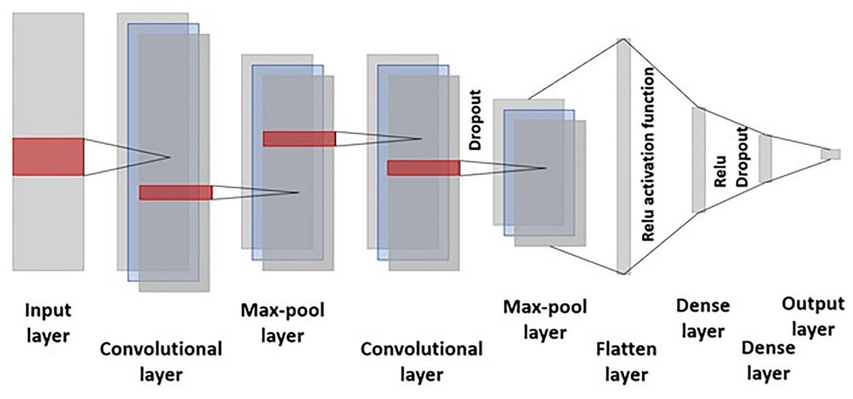


Figure 2: CNN Representation (adopted from Sandhu et. al. [1]).

Results

It has been seen that CNN reached 0.10 and 0.09 RMSEs (root mean square errors) for two important traits (anthesis and height) on original features with 64 batch size and 250 epochs (**Figure 3**). The performance is improved a bit when feature selection method is used. Mostly importantly, feature selection reduces computational time from 100.96 hrs to 43.42 hrs.

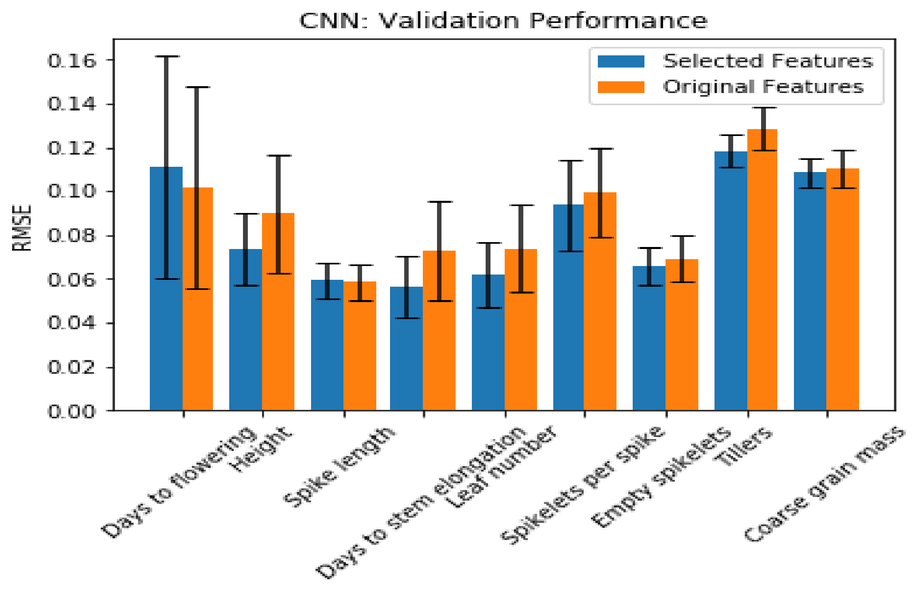


Figure 3: Performances of CNN model.

The performance of CNN model (250 epochs & 64 batch size) is improved to 0.081 and 0.073 RMSEs, when random forest was used as feature selection method (**Figure 4**).

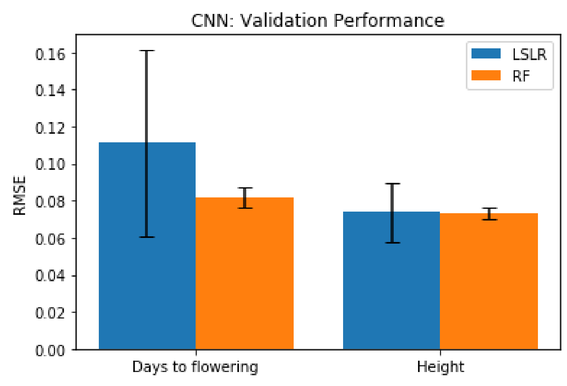


Figure 4: Performances of CNN model. LSLR = least square linear regression ($p \leq 0.01$), RF = random forest (Threshold = 0.002)

Conclusion

Considering the promising performance of our initial model, our current focus is to develop a robust DL method to predict wheat yield components from multi- 'omics data spanning to multiple field experiments. It will support breeding decisions and crop management on-farm.

FOR FURTHER INFORMATION

Md Zakir Hossain
Agriculture & Food, Black Mountain, Canberra
zakir.hossain@csiro.au
csiro.au/ Agriculture & Food

REFERENCE

[1] Sandhu KS, Lozada DN, Zhang Z, Pumphrey MO and Carter AH (2021) Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. Front. Plant Sci. 11:613325. doi: 10.3389/fpls.2020.613325

Australian National University

CSIRO