BMC Plant Biology

RESEARCH

Open Access

Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction

Mitchell Gill[1], Robyn Anderson[1], Haifei Hu[1], Mohammed Bennamoun[2], Jakob Petereit[1], Babu Valliyodan[3,4], Henry T. Nguyen[3], Jacqueline Batley[1], Philipp E. Bayer[1] and David Edwards[1*]

Abstract

Recent growth in crop genomic and trait data have opened opportunities for the application of novel approaches to accelerate crop improvement. Machine learning and deep learning are at the forefront of prediction-based data analysis. However, few approaches for genotype to phenotype prediction compare machine learning with deep learning and further interpret the models that support the predictions. This study uses genome wide molecular markers and traits across 1110 soybean individuals to develop accurate prediction models. For 13/14 sets of predictions, XGBoost or random forest outperformed deep learning models in prediction performance. Top ranked SNPs by F-score were identified from XGBoost, and with further investigation found overlap with significantly associated loci identified from GWAS and previous literature. Feature importance rankings were used to reduce marker input by up to 90%, and subsequent models maintained or improved their prediction performance. These findings support interpretable machine learning as an approach for genomic based prediction of traits in soybean and other crops.

Keywords: Machine learning, XGBoost, Interpretable models, Feature selection, Genomic selection, Soybean

Introduction

Soybean (*Glycine max*) has a variety of uses including human consumption, livestock and aquaculture feed, and biofuel production [1, 2]. The demand for soybean is expected to increase [3], whilst climate change is expected to decrease overall crop productivity, threatening global food security [4]. The production of large quantities of genomic data in the last 10-15 years has supported the development of genomics-based approaches for crop improvement that can address these challenges [5]. Genomic Selection (GS) has been applied to associate Single Nucleotide Polymorphisms (SNPs) with breeding values to accelerate crop improvement. GS has the potential to reduce breeding cycle length [6] and accelerate genetic gains in crops by improving breeding selection [7], supported by methods such as speed breeding [8].

Studies have shown that using non-linear prediction algorithms such as Machine Learning (ML) can improve prediction accuracy in GS [9–11]. The application of ML in crop breeding provides advantages such as the use of more complex data, along with potentially providing solutions to problems such as epistasic effects and genomic imprinting [12]. A relatively new subcategory of ML, Deep Learning (DL), has provided promising results in a range of fields and disciplines using interconnected neural networks such as Convolutional Neural Networks

*Correspondence: dave.edwards@uwa.edu.au
[1] School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia
Full list of author information is available at the end of the article

(CNN) and Deep Neural Networks (DNN). The advantage of DL is that when optimised appropriately, it can identify complex multidimensional patterns in large data sets [13]. CNNs specifically have already demonstrated this for GS through its application in selecting high value phenotypes from genomic data [6].

The underlying mechanisms of genotype to phenotype predictions remain unclear. Adding associated SNPs from Genome Wide Association Studies (GWAS) into linear prediction models has shown varying results [14, 15]. For non-linear models, tools are being developed to increase the accuracy of predictions in ML and DL using GWAS data [16]. The changes in accuracy using GWAS related inputs suggests that prediction models can, for certain traits, place importance on these SNPs for building the model. Identification of these loci provides the opportunity to guide the reduction of input data through feature selection in further models.

For genotype to phenotype predictions, it is appropriate to test a range of ML/DL algorithms as each algorithm has its own underlying assumptions and biases, and no algorithm provides the best performance for all traits [17]. Studies in this area often compare DL models to linear models and with older ML models such as random forest [18], but often forgo the inclusion of one of the most recent ML models, XGBoost. This omission is of concern due to XGBoost's accurate prediction in other disciplines [19–21], and XGBoost has shown to outperform DL in some tabular data problems [22].

In this study we build robust prediction models for seven agronomic traits in soybean and compare the suitability of prediction models for use in crop breeding. Our results suggest that XGBoost has an affinity for genotype to phenotype prediction and should be considered in future model development. The trait associated regions identified by XGBoost overlap with regions of significantly associated loci from GWAS, and XGBoost can independently identify these regions. Furthermore, reducing the input data to a targeted selection of SNPs based upon initial regions of importance to XGBoost tree building can provide equal or better results using far fewer SNPs.

Methods

SNP discovery & phenotype data

A total of 1110 diverse soybean accessions were selected from the USDA Soybean Germplasm Collection for SNP calling [23]. The sequence metadata for all 1110 soybean accessions are summarized in Table S2. Clean reads were mapped to the pangenome using BWA-MEM [24] v0.7.17 with default settings and duplicates removed by Picard tools (http://broadinstitute.github.io/picard/). Reads were realigned using GATK [25] v3.8-1-0 RealignerTargetCreator and IndelRealigner, followed by variant calling using GATK HaplotypeCaller. The resulting SNPs were filtered following the SNP filtering methodology described in Marsh et al [26], (QD < 2.0 || MQ < 40.0 || FS > 60.0 || QUAL < 60.0 || MQrankSum < −12.5 || ReadPosRankSum < −8.0) to remove low-quality SNPs. High-confidence SNPs were identified by removing SNPs with minor allele frequency (MAF) < 0.05 and missing genotype rate < 10% using VCFtools [27]. Phenotype data for flower colour, seed coat colour, pod colour, pubescence density, seed oil content, seed protein content and seed weight were downloaded from the USDA-GRIN database (https://npgsweb.ars-grin.gov/) for the accessions that had available observation data. The range of phenotype data is summarised in Table S3.

Genome wide association studies

The R package rMVP v0.99.15 [28] was used to conduct GWAS, with the FarmCPU statistical technique. FarmCPU allows for the population structure to be controlled by using the first three principal components (PCs) from an automatic principal component analysis (PCA) based on the marker data [29], whilst the significance threshold was defined as 0.05/marker size. GWAS was run with rMVP using the following settings, nPC.FarmCPU = 3, priority = "memory", vc.method = "BRENT", maxLoop = 10, method.bin = "EMMA", threshold = 0.05.

Data pre-processing for machine learning model building

Vcftools [27] vcf-to-tab was used to remove vcf preamble. A python script was used to reformat this file into a structured csv. Soybean lines with over 1% missing data were excluded.

SNPs were reduced by 95% by extracting 1 in 20 sequential SNPs, as our initial total of approximately 5 million SNPs was not compatible with GPU memory requirements. For each trait, 20% of samples were randomly excluded to create a holdout set by using the packages pandas and numpy within a python script before model building. This holdout data was later used for model validation.

For models with reduced feature input, SNPs for genomic regions based on XGBoost feature importance were extracted using a custom python script. Holdout sets were produced in the same manner as for the complete dataset.

Model building and feature importance

A virtual sandbox environment was built on GPU servers using singularity with a tensorflow docker image (version 'tensorflow:20.03-tf2-py3'). Jupyter notebooks were connected to the server, and are available at https://github.com/mitchgill16/Soybean_Trait_Prediction. The python package Scikit learn v0.21 [30] was used to perform a variety of ML relevant tasks.

For multiclass classification problems, XGBClassifier was initiated with objective = 'multiclass' and num_classes set to the number of trait classes. For other regression and classification problems, XGRegressor and XGBClassifier were initiated with default settings, all of which were loaded in with the XGBoost package v1.1.1. XGBoost model parameters were optimised using the BayesSearchCV object from scikit optimisation package v0.8.1. The setting space was follows: learning_rate(0.01 - 1.0), min_child_weight(0 - 10), max_depth(0 - 50), max_delta_step(0 - 20), subsample(0.01 - 1.0), colsample_bytree(0.01 - 1.0), colsample_bylevel(0.01 - 1.0), reg_lambda(1e^9 - 1000), reg_alpha(1e^9 - 1.0), gamma(1e-9 – 0.5), min_child_weight(0 - 5), n_estimators(50 – 200) and scale_pos_weight(1e^6 - 500). The optimum XGBoost parameters were then used to fit the XGBoost models using stratified k-fold cross validation for classification tasks, and standard k-fold cross validation (k = 10) from the scikit learn model selection package. The best performing model from cross validation was saved using the inbuilt pickle python package (protocol version 5) and used for prediction on the holdout dataset.

The functions 'get_booster' and 'get_scores' was used to generate dictionaries of importance scores for each feature used by the XGBoost model. The top 20 scores were retrieved and ranked. Original SNP names were retrieved from a stored list of headers, and allele values were retrieved by a custom inverse one hot encoding function.

Random forest objects for classification (RandomForestClassifier), and continuous (RandomForestRegressor) traits were loaded from the scikit learn ensemble python package. Both the random forest classifier object and random forest regressor object were initialised with n = estimators = 100, max_features = "sqrt" and the random state set to a random integer between 0 and 5000. To ensure an optimised and fitted model could be selected for use on the holdout dataset, the random forest objects were fitted using stratified k-fold cross validation for classification tasks, and standard k-fold cross validation (k = 10) from the scikit learn model selection package. The best performing model from all folds during cross validation was selected and subsequently saved using the inbuilt pickle python package and used for prediction on the holdout dataset.

The Keras 2.4.3 interface (https://github.com/keras-team/keras) for the Tensorflow v2.1.0 [31] python library was used to build sequential DL models. The CNN architecture was initially adapted from the successful models in the GMStool [16] as a baseline for further adjustment, whilst the DNN architecture involved adapting elements from both the GMStool paper and a successful DNN architecture for prediction of yield [32]. The final hyperparameter and architecture choices were a mixture of trial and error, grid searching and adaption from the aforementioned papers.

The CNN models used three 1D convolution layers using Rectified Linear Unit (ReLU) activation, with a 20% dropout between layer one and two, and a 10% dropout between layer two and three. The convolution layers had 12, 10 and 8 filters and a kernel size of 14, 10 and 8 respectively. The convolution layers were followed by a 1D max pooling layer of size 2 and batch normalisation before being flattened and fed into three dense layers. These dense layers had 48, 32 and 16 nodes, each with ReLU activation and were followed by a batch normalisation layer.

The DNN models for this study were a fully connected feed forward multilayer perceptron network consisting of five dense layers with ReLU activation functions, a dropout layer of 3, 2 and 1% after the first three layers respectively, and a batch normalisation layer after the final dense layer. The layers consisted of 200, 100, 64, 32 and 16 nodes each. For both CNN's and DNN's, the output layer had one node with linear activation for continuous traits, 1 node with sigmoid activation for binary categorical traits and x nodes with softmax activation for multiclass traits, where x = the number of possible classes. The optimiser used was Adamax with a learning rate of 0.003. The batch size was 1/50th of the total amount of samples to ensure memory could be managed when running through the neural network.

Results

XGBoost, random forest, CNN and DNN models were trained and evaluated on SNP input data uniformly distributed across the genome to predict each of the following traits: flower colour, pod colour, pubescence density, seed coat colour, seed oil content, seed protein content and seed weight (Table 1, Fig. 1). Categorical traits were evaluated with classification accuracy, whereas continuous traits used Root Mean Squared Error (RMSE) as a percentage of trait mean to evaluate models and allow a comparison of prediction error across traits and models. XGBoost models outperformed other models for all categorical traits, however for the continuous traits of seed oil and seed protein percentage, random forest was the best performer, and for seed weight the CNN was most accurate. In comparison to the DL architectures (CNN & DNN), XGBoost was on average 10.32% more accurate across classification predictions. XGBoost's performance in comparison to DL prediction error for traits requiring regression analysis was negligible, with an average of 0.16% reduction in error for regression traits when

Table 1 Evaluation and comparison of prediction models on whole genome SNP input data

		Learning Algorithm					
Evaluation Metric	**Trait**	**XGB†**	**RF†**	**CNN†**	**DNN†**	**XGB-DL Diff‡**	**RF-DL Diff‡**
Accuracy	Flower Colour	**96.79%**	95.51%	87.82%	81.41%	12.18%	10.90%
	Pod Colour	**84.52%**	76.13%	70.32%	70.32%	14.20%	5.81%
	Pubescence Density	**91.56%**	81.17%	83.77%	80.52%	9.42%	-0.97%
	Seed Coat Colour	**89.68%**	85.16%	84.52%	83.87%	5.49%	0.97%
RMSE % of Mean	Seed Oil Percentage	11.14%	**10.67%**	13.44%	10.74%	-0.95%	-1.42%
	Seed Protein Percentage	6.41%	**6.32%**	6.33%	6.91%	-0.21%	-0.30%
	Seed Weight	19.03%	21.63%	**17.77%**	18.91%	0.69%	3.29%

† XGB = XGBoost, RF = Random Forest, CNN = Convolutional Neural Network, DNN = Deep Neural Network

‡ XGB-DL Diff = Difference in performance between XGBoost & deep learning architectures, RF-DL Diff = Difference in performance between random forest and deep learning architectures

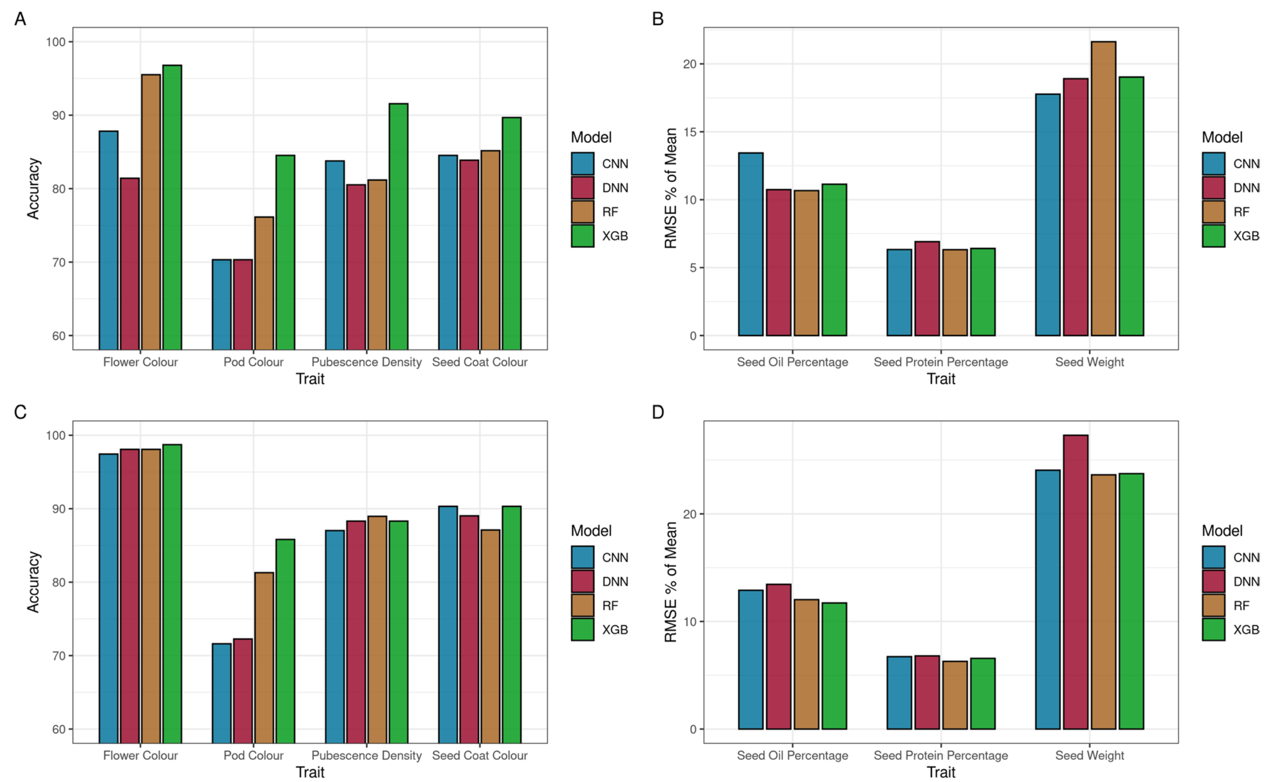


Fig. 1 Model Prediction Performance Across Soybean Traits. **A** Accuracy for flower colour, pod colour, pubescence density and seed coat colour for models trained on SNP input data uniformly distributed across the soybean genome. **B** Root mean square error as a percentage of mean trait value for seed oil as a percentage of total seed weight, seed protein as a percentage of total seed weight and total seed weight. Models were trained on SNP input data uniformly distributed across the soybean genome. **C** Accuracy for flower colour, pod colour, pubescence density and seed coat colour for models trained on reduced SNP input data set. **D** Root mean square error as a percentage of mean trait value for seed oil as a percentage of total seed weight, seed protein as a percentage of total seed weight and total seed weight. Models were trained on a reduced SNP input data set

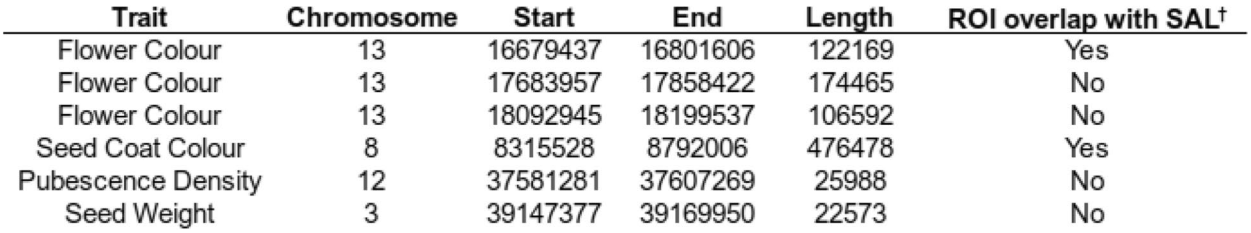
compared to DL prediction error. The best performing XGBoost model on each continuous trait was outperformed by the trained DNN for seed oil prediction, the trained CNN for seed protein prediction and both the CNN and DNN models for seed weight prediction. Like XGBoost, random forest on average performed better than DL models for classification traits, with a 4.17% increase in accuracy when compared to DL accuracy across classification prediction. Random forest models also had the lowest error for both seed oil and seed protein prediction.

To compare the results and assess the potential to reduce input data we interpreted the XGBoost models and found SNPs that were important for prediction. For each trait, the XGBoost input SNPs were ranked by their F-score, measured in gain, which is a measure of the relative contribution of each SNP. For flower colour, seed coat colour, pubescence density and seed weight, the 20 highest ranking SNPs by XGBoost F-score included subsets of at least 3 SNPs within close proximity. Each SNP in a subset was separated by a maximum of 100kbp from their nearest neighbouring SNP within a given subset. The genetic regions that these subsets of SNPs spanned were defined as regions of importance (ROI), of which 6 in total were identified across 4 traits, and are summarised in Table 2. To investigate the ROIs further we performed a GWAS with all available SNPs to determine whether any of the ROI overlapped with loci identified from GWAS. GWAS identified one major loci for each of flower colour, seed coat colour and pod colour (Table 3) (Fig. S1). A flower colour ROI and seed coat colour ROI overlapped with significant GWAS loci for their respective trait, whereas the seed weight ROI did not overlap with any significant GWAS loci. There were no significant regions identified for pubescence density using GWAS, however a region identified from our XGBoost models overlaps with a previously identified locus on chromosome 12 for pubescence density [33]. In summary, the majority of the significant GWAS loci identified in this study had an overlapping ROI whilst three out of six ROI identified had an overlapping significant GWAS loci from this study and a previous study.

We selected a subset of targeted SNPs for each trait based on the XGBoost interpretation (Figs. S2-S5, Table S1). Input data was reduced, retaining between 27 and 4% depending on the trait being predicted. Subsequent models for predicting discrete traits showed an increase in accuracy across all classification predictions, ranging from 2.42 to 7.70% (Table 4). However, models for continuous traits showed mixed results (Table 4). Seed protein prediction showed a negligible increase in average error when compared to the original models. Seed oil prediction had a minor increase in average error across models, whilst seed weight prediction had a larger increase in average error of 5.35% compared to the original models.

When using the reduced SNP input data to evaluate the ML and DL architectures (Table 5), XGBoost performed well for classification traits as it had the highest accuracy for flower colour, pod colour and seed coat colour, while for regression traits, XGBoost had the lowest error for seed oil prediction. In comparison to the DL models, XGBoost was on average 4.03% more accurate for classification traits when using reduced SNP input data. For regression traits, XGBoost showed a 1.20% lower prediction error when compared to the average DL prediction error, and outperformed both the CNN and DNN model for each trait.

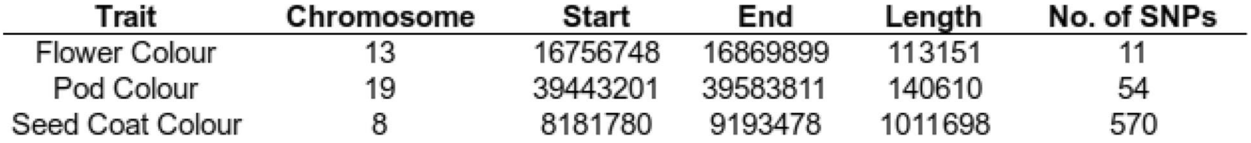
Table 2 Regions of Importance (ROI) from XGBoost



Trait	Chromosome	Start	End	Length	ROI overlap with SAL[†]
Flower Colour	13	16679437	16801606	122169	Yes
Flower Colour	13	17683957	17858422	174465	No
Flower Colour	13	18092945	18199537	106592	No
Seed Coat Colour	8	8315528	8792006	476478	Yes
Pubescence Density	12	37581281	37607269	25988	No
Seed Weight	3	39147377	39169950	22573	No

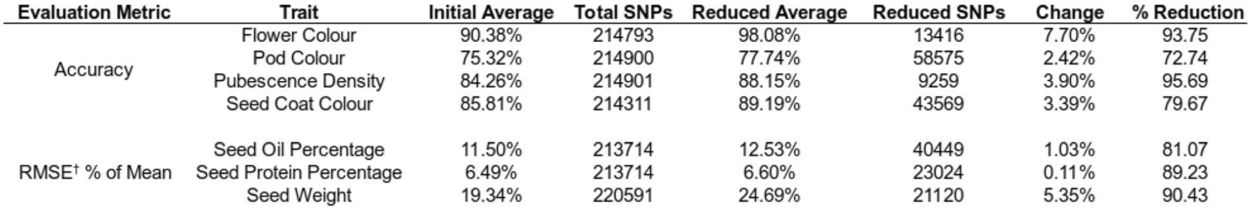
[†] SAL = Significantly Associated Loci identified from GWAS

Table 3 Significantly Associated Loci (SAL) from GWAS



Trait	Chromosome	Start	End	Length	No. of SNPs
Flower Colour	13	16756748	16869899	113151	11
Pod Colour	19	39443201	39583811	140610	54
Seed Coat Colour	8	8181780	9193478	1011698	570

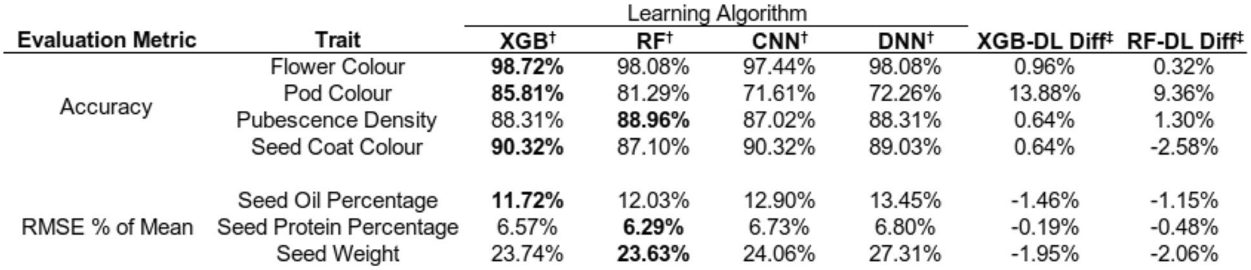
Table 4 Model performance from whole genome SNP input data compared to Reduced Input SNP Data



Evaluation Metric	Trait	Initial Average	Total SNPs	Reduced Average	Reduced SNPs	Change	% Reduction
Accuracy	Flower Colour	90.38%	214793	98.08%	13416	7.70%	93.75
	Pod Colour	75.32%	214900	77.74%	58575	2.42%	72.74
	Pubescence Density	84.26%	214901	88.15%	9259	3.90%	95.69
	Seed Coat Colour	85.81%	214311	89.19%	43569	3.39%	79.67
RMSE[†] % of Mean	Seed Oil Percentage	11.50%	213714	12.53%	40449	1.03%	81.07
	Seed Protein Percentage	6.49%	213714	6.60%	23024	0.11%	89.23
	Seed Weight	19.34%	220591	24.69%	21120	5.35%	90.43

[†] RMSE = Root Mean Square Error

Table 5 Evaluation and comparison of prediction models with reduced input data



		Learning Algorithm					
Evaluation Metric	**Trait**	**XGB†**	**RF†**	**CNN†**	**DNN†**	**XGB-DL Diff‡**	**RF-DL Diff‡**
Accuracy	Flower Colour	**98.72%**	98.08%	97.44%	98.08%	0.96%	0.32%
	Pod Colour	**85.81%**	81.29%	71.61%	72.26%	13.88%	9.36%
	Pubescence Density	88.31%	**88.96%**	87.02%	88.31%	0.64%	1.30%
	Seed Coat Colour	**90.32%**	87.10%	90.32%	89.03%	0.64%	-2.58%
RMSE % of Mean	Seed Oil Percentage	**11.72%**	12.03%	12.90%	13.45%	-1.46%	-1.15%
	Seed Protein Percentage	6.57%	**6.29%**	6.73%	6.80%	-0.19%	-0.48%
	Seed Weight	23.74%	**23.63%**	24.06%	27.31%	-1.95%	-2.06%

† XGB = XGBoost, RF = Random Forest, CNN = Convolutional Neural Network, DNN = Deep Neural Network

‡ XGB-DL Diff = Difference in performance between XGBoost & deep learning architectures, RF-DL Diff = Difference in performance between random forest and deep learning architectures

Like XGBoost, random forest on average performed better than DL models for classification traits using reduced SNP input data, with a 2.10% increase in accuracy when compared to DL accuracy across classification traits. Using reduced SNP input data, random forest was the best performing model for seed oil and seed weight prediction. Random forest had a reduction in error of 1.23% when compared to DL error and outperforms both the CNN and DNN model for each of three regression traits.

Across all traits with the complete SNP data, XGBoost was the best performing model for 4/7 traits, random forest for 2/7 traits, and a CNN for 1/7 traits. With reduced SNP data, XGBoost was the best performing model for 4/7 traits, and random forest for 3/7 traits. For both datasets XGBoost was the best predictor of flower colour, pod colour and seed colour whilst random forest was the best predictor of seed protein content. Overall, XGBoost was the best performer 8 of 14 times, and remained robust after feature reduction.

Discussion

Machine learning performance

Here we demonstrate the performance of XGBoost and random forest for genotype to phenotype predictions in comparison to widely used DL architectures. From the 14 sets of models, DL outperformed the ML models on only one occasion. The models in this study were all non-linear, which has the advantage of being able to include non-additive variances [34, 35]. Our results suggest that XGBoost and random forest models are better able to account for non-additive effects in genotype to phenotype prediction than DL architectures. The study emphasises the need to test a variety of models and provides an evaluation of the XGBoost algorithm in this research space.

Building and training effective DL models requires abundant high-quality training data [36], and for this study there was access to over 5.5 million SNPs for between 700 and 1000 individuals per trait. Khaki and Wang [32] produced successful yield prediction models in maize using over 2000 individuals, and Jeong et al. [16], trained CNNs on soybean with 1928 individuals. However it is to be noted that Jeong et al. [16], also trained successful CNN's for rice genomic prediction on 413 individuals which suggests that sample size might not be the sole reason that a deep learning model underperforms. Whilst model optimisation and training were done according to best practices, it is possible that ML algorithms were better able to train on this smaller dataset than DL algorithms. Recent research supports this notion by concluding that random forest was able to better predict from small, tabulated datasets than DL, whilst DL was able to better predict from larger tabulated datasets [37]. Our work remains consistent with this finding and suggests that XGBoost has the ability to predict effectively from relatively small, tabulated datasets. Another consideration is recent evidence to suggest that DL models do not estimate complex marker effects, but rather use the genetic relatedness between markers to make predictions. This may partially explain the underperformance of DL in crop genomic prediction problems [38].

The importance of marker density for prediction varies between datasets, with some evidence demonstrating increasing SNP density can increase accuracy and is positively correlated with heritability [39], whereas other evidence suggests that marker density is less important, and that heritability of a trait is more important [40]. A study in cattle found that highly ranked subsets of 400-3000 SNPs provided better results than evenly spaced SNPs across the genome [41], similar to the results of our study. The ability to reduce the number of SNPs whilst retaining a high prediction performance increases the feasibility of genotyping for genomic selection with tools such as SNP arrays.

Explainability and interpretability for genotype to phenotype prediction

XGBoost's inbuilt methods were used to interpret the model and guide the selection of dense areas of SNPs for further model building. Using model interpretability to guide feature selection was effective, however a potential limitation of the study is that for some traits in the initial genome-wide SNP input data models, XGBoost was not the best performing model. As an alternative, model-agnostic local explanation methods such as SHAP [42] could be used on the best performing models to rank feature importance. Other methods that could be used to reduce input include QTL-based genomic assisted prediction [43], using GWAS associated SNPs [44] or using a selection of significant and non-significant SNPs from GWAS to train genomic prediction models [16].

One concern with explaining models that were built without the intention of interpretability is that it can lead to issues of validation [45], and explanations that are misleading [46]. Interpretable models such as XGBoost enable the identification of features in the underlying architecture, and support retraining to improve model classification. In addition, interpretable models allow the identification of genetic markers, and hence a genomic location for traits, providing biological context [32]. Further model building through reduction of the input space has demonstrated the ability to improve the performance across ML and DL models for trait prediction [17, 32, 47].

Comparison and similarity of XGBoost feature importance and significantly associated regions identified through GWAS

The loci identified from GWAS for flower colour on chromosome 13, seed coat colour on chromosome 8 and pod colour on chromosome 19 are supported by previous research [48, 49]. In addition, the associated loci for flower and seed coat colour overlapped with the regions of importance identified by XGBoost. Despite being different methods, there is evidence that interpretable ML can use correlation between features and outcomes to extract associations [50], which is similar to how GWAS tests for marginal association between a target trait and SNP [51]. Our XGBoost models add evidence to this idea, as they learn which inputs to use in decisions by lowering the cumulative residual error [52], and identified loci associated with flower and seed coat colour without manual labelling to inform the model that there was a significant locus present.

This study demonstrated that XGBoost and random forest models have the ability to outperform DL architectures for genotype to phenotype prediction problems. XGBoost models and GWAS identified overlapping genomic regions for two traits. For other traits, XGBoost identified genomic regions hosting multiple SNPs that may help define new associated regions. Finally, this study used the feature importance results as a guide to reduce the number of SNPs required. These results demonstrate the feasibility of feature reduction for genotype to phenotype predictions and showcase the importance of an appropriate representation of input.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-022-03559-z.

Additional file 1: Supplementary Figure 1. *P*-value of each SNPs association for a) flower colour b) seed coat colour c) pod colour in the soybean VCF. SNPs coloured red have been determined as significantly associated for the given trait as they have a *p*-value less than the -log10(8) significance threshold for this GWAS. **Supplementary Figure 2.** Graphs ranking the top 20 most input SNPs by gain as identified by XGBoost models for trait predictions for traits with regions of importance identified from XGBoost. Blue bars are region of importance, whereas other colours represent collections of important SNPs on the same chromosome. Black bars represent left over SNPs with no relation to other SNPs in the ranking. SNP rankings for genome wide SNP input for A) flower colour B) seed coat colour C) pubescence density D) seed weight. **Supplementary Figure 3.** Top 20 ranked SNPs for XGBoost Seed Oil Prediction. **Supplementary Figure 4.** Top 20 ranked SNPs for XGBoost Pod Colour Prediction. **Supplementary Figure 5.** Top 20 ranked SNPs for XGBoost Seed Protein Prediction. **Supplementary Table 1.** Targeted Regions of SNPs for Reduced Input Models. **Supplementary Table 2.** List of soybean germplasm in the pangenome with the sequence coverage. (ND, not defined). **Supplementary Table 3.** Trait Data Types.

Acknowledgements

This work was supported by resources provided by the Pawsey Supercomputing Centre during the 2020/2021 Pawsey Summer Internship Programme with funding from the Australian Government and the Government of Western Australia. This work is funded by the Australia Research Council (Projects DP1601004497, LP160100030 and DP200100762, DE210100398) and the Grains Research and Development Corporation (Projects 9177539 and 9177591). Dr. Philipp Bayer acknowledges the support of the Forrest Research Foundation.

Authors' contributions

M.G. conducted Genome Wide Association Studies, ML model building and ML interpretation. M. G analysed model performance with guidance from R. A and P.B. H.H generated the tabulated genomic and phenotype data. M.G, R.A, P.B & D.E conceptualised the project. M.G wrote the manuscript. All authors contributed to the editing of the manuscript. The author(s) read and approved the final manuscript.

Availability of data and materials

The sequence metadata for all 1,110 soybean accessions are summarized in Table S2. Of these lines, 118 were previously published in PRJNA257011 [48] and 104 were previously published in PRJNA289660 [53]. The rest of the sequenced data is publicly available from the SRA project PRJNA639876 [54].

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1] School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia. [2] Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA, Australia. [3] Division of Plant Sciences and National Center for Soybean Biotechnology, University of Missouri, Columbia, MO 65211, USA. [4] Department of Agriculture and Environmental Sciences, Lincoln University, Jefferson City, MO 65101, USA.

Received: 30 December 2021 Accepted: 21 March 2022
Published online: 08 April 2022

References

1. Suciu V, Rusu T, Rezi R, Urdă C. Agrotechnic, economic and environmental advantages of the soybean crop. ProEnvironment/ProMediu. 2019;12:112-5.
2. Rodionova MV, et al. Biofuel production: challenges and opportunities. Int J Hydrog Energy. 2017;42:8450–61.
3. Ray DK, Mueller ND, West PC, Foley JA. Yield trends are insufficient to double global crop production by 2050. PLoS One. 2013;8:e66428.
4. Anderson R, Bayer PE, Edwards D. Climate change and the need for agricultural adaptation. Curr Opin Plant Biol. 2020;56:197–202.
5. Abberton M, et al. Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol J. 2016;14:1095–8.
6. Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta. 2018;248:1307–18.
7. Voss-Fels KP, Cooper M, Hayes BJ. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 2019;132:669–86.
8. Watson A, et al. Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants. 2018;4:23–9.
9. Crossa J, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.
10. Cuevas J, et al. Genomic prediction of genotype $\times$ environment interaction kernel regression models. Plant Genome. 2016;9:1–20.
11. Pérez-Rodríguez P, et al. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3. 2012;2:1595–605.
12. Varona L, Legarra A, Toro MA, Vitezica ZG. Non-additive effects in genomic selection. Front Genet. 2018;9:78.
13. Zou J, et al. A primer on deep learning in genomics. Nat Genet. 2019;51:12–8.
14. Rice B, Lipka AE. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and Sorghum. Plant Genome. 2019;12:180052.
15. Spindel JE, et al. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity. 2016;116:395–408.
16. Jeong S, Kim J-Y, Kim N. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci Rep. 2020;10:19653.
17. Azodi CB, et al. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3. 2019;9:3691–702.
18. Montesinos-López OA, et al. A review of deep learning applications for genomic selection. BMC Genomics. 2021;22:19.
19. Nguyen H, Bui X-N, Bui H-B, Cuong DT. Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study. Acta Geophysica. 2019;67:477–90.
20. Inoue T, et al. XGBoost, a machine learning method, predicts neurological recovery in patients with cervical spinal cord injury. Neurotrauma Rep. 2020;1:8–16.
21. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme gradient boosting as a method for quantitative structure–activity relationships. J Chem Inf Model. 2016;56:2353–60.
22. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere. 2019;10:373.
23. Song Q, et al. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, genomes, genetics. 2015;5:1999–2006.
24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv. 2013;1303:3997.
25. McKenna A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
26. Marsh JI, et al. Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus cqProt-003. Theor Appl Genet. 2022;1–13.
27. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
28. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics Proteomics Bioinformatics. 2021.
29. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.
30. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Machine Learn Res. 2011;12:2825–30.
31. Abadi M, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv. 2016;1603:04467.
32. Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci. 2019;10:621.
33. Chang H-X, Hartman GL. Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Front Plant Sci. 2017;8:670.
34. González-Camacho JM, et al. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome. 2018;11:170104.
35. Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. Crop Sci. 2012;52:146–60.
36. Taylor L, Nitschke G. Improving deep learning using generic data augmentation. arXiv preprint arXiv. 2017;1708:06020.
37. Xu H, et al. When are Deep Networks really better than Random Forests at small sample sizes? arXiv preprint arXiv. 2021;2108:13637.
38. Ubbens J, Parkin I, Eynck C, Stavness I, Sharpe A. Deep Neural Networks for Genomic Prediction Do Not Estimate Marker Effects. 2021. https://doi.org/10.1101/2021.05.20.445038.
39. Liu X, et al. Factors affecting genomic selection revealed by empirical evidence in maize. Crop J. 2018;6:341–52.
40. Zhang A, Wang H, Beyene Y, Semagn K, Liu Y, Cao S, et al. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. Front Plant Sci. 2017;8:1916.
41. Li B, et al. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet. 2018;9.
42. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
43. Zheng W, et al. Quantitative trait loci-based genomics-assisted prediction for the degree of apple fruit cover color. Plant Genome. 2020;13:e20047.
44. An Y, et al. Genome-wide association studies and whole-genome prediction reveal the genetic architecture of KRN in maize. BMC Plant Biol. 2020;20:490.
45. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Machine Intell. 2019;1:206–15.
46. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020. p. 180–6. https://doi.org/10.1145/3375627.3375830.
47. Hoffstetter A, Cabrera A, Huang M, Sneller C. Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3 Genes|Genomes|Genetics. 2016;6:2919–28.
48. Fang C, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. 2017;18:1–14.
49. Qi X, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun. 2014;5:4340.
50. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. Trends Genet. 2020;36:442–55.
51. Yang S, Wen J, Eckert ST, Wang Y, Liu DJ, Wu R, et al. Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. Bioinformatics. 2020;36:3811–7.
52. Chen T, Guestrin CX. A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–94.
53. Valliyodan B, et al. Landscape of genomic diversity and trait discovery in soybean. Sci Rep. 2016;6:1–10.
54. Bayer PE, Valliyodan B, Hu H, Marsh JI, Yuan Y, Vuong TD, et al. Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. Plant Genome TSI. 2021:1–12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.