IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 1, 2022, accepted April 24, 2022, date of publication April 29, 2022, date of current version May 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171341

Predicting Phenotypes From High-Dimensional Genomes Using Gradient Boosting Decision Trees

TINGXI YU[1], LI WANG[1], WUPING ZHANG[1], GUOFANG XING[2], JIWAN HAN[1], FUZHONG LI[1], AND CHUNQING CAO[1]

[1]College of Software, Shanxi Agricultural University, Taigu, Shanxi 030801, China
[2]College of Agriculture, Shanxi Agricultural University, Taigu, Shanxi 030801, China

Corresponding author: Wuping Zhang (zwping@126.com)

This work was supported in part by the Major Research Plan of Shanxi Province, China, under Grant 201703D211002-2; in part by the Agricultural Big Data Innovation Platform under Grant K481811076; and in part by the Major Research Plan through the Screening of Seenriched Millet Germplasm and Study on Se-Enriched Molecular Mechanism of Shanxi Province, China, under Grant 201803D21008-4.

ABSTRACT Genomic selection (GS) is an emerging technique for predicting unknown phenotypes using genome-wide marker coverage, allowing the use of efficient computational models to select individuals with high phenotypic values as candidate breeding populations. However, GS remains challenging inefficient crop breeding due to the limited size of training populations, the nature of genotype-environment interactions, and the complex interaction patterns between molecular markers. In this study, we use ensemble learning algorithms to construct gradient boosted decision tree (GBDT) models to achieve the prediction of phenotypic values from genotypic markers. We trained GBDT using the wheat GS dataset and compared the predictive performance with six other widely used GS models. The mean normalized discounted cumulative gain (MNDCG) method was used to evaluate the ability of each model to select individuals with high phenotypic values. The results of the study show that: (1) Bayesian models converge and reach a steady-state only when a sufficient number of iterations are set. As the number of iterations increases, the prediction accuracy of the Bayesian model increases, but the computational efficiency of the model decreases significantly. When 200,000 iterations are performed, the prediction performance of the five Bayesian models is similar and converges to a smooth state, and their prediction accuracy is 7.60% better than the GBDT model overall, and the computational efficiency of the GBDT model is 70 times that of the Bayesian model. (2) Overall, the overall prediction performance of the RRBLUP model was the best, but for some traits, the GBDT model still had a higher ability to select individuals with high phenotypic values than the RRBLUP and Bayesian models. (3) The prediction accuracy of GBDT and RRBLUP models was influenced by the subset of markers, and the higher the number of markers the higher the prediction accuracy of the models, so the reasonable selection of genetic marker data of appropriate size could improve the prediction performance of the models.

INDEX TERMS Genomic selection, gradient boosted decision tree, ensemble learning, phenotypic prediction, wheat.

I. INTRODUCTION

Genomic selection, originally proposed by Meuwissen [1] *et al.* to predict unknown phenotypes by using genome-wide markers, is an effective marker-assisted breeding paradigm. Researchers have explored the application of GS not only in animal breeding [2]–[5] but also in plant and crop breeding [6]–[11]. Many important traits in plant breeding are controlled by multiple genes, and plant or crop phenotypes for polygenic traits can be better predicted through the use of whole-genome markers, Unlike MAS, GS can predict the phenotypic trait values of individuals before planting the crop, thus contributing to the rapid selection of superior genotypes and accelerating the breeding cycle [12], [13]. Despite this, the application of GS in crop breeding is still at a nascent stage, mainly because its high-dimensional marker dataset may make it difficult to accommodate more anomalous and discrete data when performing phenotype

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

prediction, which leads to lower prediction accuracy and stability. Therefore, the application of GS in crop breeding is still challenging in terms of high predictive performance.

In recent years, genomic selection computational methods, application strategies, and breeding programs have proliferated through continuous development and research by many scholars. Various statistical models for GS breeding have been developed, including the GBLUP [13], [14], RRBLUP [15], [16] models based on the improved BLUP algorithm model, and HBLUP [17], [18], which is commonly used for animal breeding, and BayesA [1], BayesB [19], Bayesian LASSO (BL) [20], Bayesian ridge regression (BRR) [21], BayesC and BayesCπ [19], [22], which are based on the improved Bayesian algorithm. These GS models based on traditional statistical methods usually make assumptions before performing linear regression analysis. As an example, the RRBLUP model, which predicts phenotypes based on a linear function of genotypic markers only when the effects of all marker genes are assumed to arrive at a minimum and non-zero normal distribution [54]. These GS models not only face the statistical challenges associated with high-dimensional marker data but also have difficulty capturing the complex relationships within genotypes and between genotypes and genes. How to optimize the model, minimize artificially set parameters, and greatly improve computational efficiency while ensuring high accuracy is the future direction of genome-wide selection model optimization. Therefore, novel methods are urgently needed to improve the potential of GS in plant breeding.

With the continuous development of big biological data, machine learning has attracted the attention of many biologists and researchers working in the field of genetics, and it has been well applied in many disciplines [23], and researchers have successfully applied it to gene expression inference [24]–[26], functional annotation of genetic variants [27], gene and disease association prediction [28]–[30], the identification of protein folds and prediction of genome accessibility [11], [31]. These applications demonstrate the powerful ability of machine learning to learn complex relationships from biological data [32]–[34]. Therefore, machine learning has also started to be tried for GS breeding. Ma,wenlong *et al.* [36] attempted to use deep convolutional networks to capture complex interactions between markers to predict wheat phenotypes, Qin C [37] *et al.* improved GS prediction performance by constructing random forests, Wang H [35], Montesinos-López [39], *et al.* attempted to use deep learning for genomic prediction of plants and crops, Mathieu Blonde [47] *et al.* attempted to use gradient boosting regression trees and random forests for genomic prediction of crops, and Rosado R [54] *et al.* attempted to use artificial neural networks to improve the prediction accuracy of flowering traits in beans. Min-Gyoung Shin [55] *et al.* implemented random forest and gradient boosting, to evaluate the ability of significant markers in predicting phenotype values, and demonstrated the contribution of different marker combinations on trait values via prediction trees. Yan J, Xu Y, Cheng Q [56], *et al.* implemented a machine learning method, namely light gradient gradient boosting machine (LightGBM), to evaluate the ability of significant markers in predicting phenotype values, and LightGBM has been implemented as a toolbox, Crop Genomic Breeding Machine CropGBM [57], encompassing multiple novel functions and analytical modules to facilitate genomically designed breeding in crops, and demonstrate its use on diverse maize lines containing high-density markers. Overall, these studies show the potential of machine learning to capture complex interactions between markers relative to traditional GS methods.

The gradient boosting (GB) is an ensemble learning strategy that brings together multiple weak learners to build a strong model [42], [43]; therefore, its prediction accuracy is significantly better than that of a single model. The GB algorithm has been applied to animal and plant genome prediction [56], [57], and the prediction ability of GB is better than that of artificial neural networks (ANN), feed-forward neural networks (FNN), and random forest (RF) [38]. In contrast to RF, both GB and RF use ensemble learning algorithms, but GB is constructed differently from RF [41]. RF builds independent trees through a bagging ensemble strategy and averages the results of all trees as the final prediction. GB, on the other hand, builds the tree by gradient boosting iterations. In each iteration, the current tree is built based on the previous tree, and the error between the predicted and actual values of the previous tree is set as the prediction target of the current tree. Subsequently, the error value is gradually minimized by performing hundreds of iterations, and the results of all trees are summed up as the final prediction value. Compared with ANN and FNN, GB employs a unique feature extraction strategy that can accomplish feature selection and prediction simultaneously. During the tree building process, GB needs to traverse all the features to select the important nodes, and the prediction of the model is only based on the features with high importance [45]. In addition, the selected features always keep their original form and no weights are set on them. However, unlike GB, ANN and FNN models first perform feature selection by calculating weights for each marker, and then calculate recombinant new features based on the sum of weighted features to represent a set of neighboring markers within a predetermined genomic region [40]. The validity of the recombinant new traits may be diminished if the region contains too much information about markers that are not relevant to the predicted traits. Therefore, excessive feature weighting may reduce the prediction accuracy and stability of the neural network, and may also lead to model training failure due to gradient explosion.

In that study, we explore the application of integrated learning in the GS domain. The contributions of this paper are as follows:

- Integrating multiple decision trees using ensemble learning techniques to construct GBDT regression prediction models for predicting individual phenotypic values from genetic markers.

- We construct an evaluation method of GS model prediction performance (MNDCG) for evaluating the ability of the model to select individuals with high phenotypic values.
- We discuss in this paper the effect of wheat GS dataset size on the predictive performance of GBDT models and confirm that selecting appropriately sized markers is an important way to improve the predictive performance of the models. It is also confirmed that a single evaluation system does not give a good indication of how good the model is, and although the predictive performance of the RRBLUP model is still higher than that of GBDT, for some traits, the GBDT model will also have a higher ability to select individuals with high phenotypic values than RRBLUP.

This paper is structured as follows, Section 2 gives an account of the sources of experimental data, model evaluation methods, the experimental environment, and the overall architecture and algorithmic principles of the main models; Section 3 presents the experimental results; Section 4 focuses on the discussion of the experimental results, and Section 5 concludes this study.

II. MATERIALS AND METHODS

A. EXPERIMENTAL DATASETS

The GS dataset used in this study was obtained from the wheat gene bank of the International Maize and Wheat Improvement Center (http://genom ics.cimmy t.org/mexic an_irani an/traverse/irani an/stand arize dData_univa riate .RData), which The gene bank contains 2000 samples of Iranian bread wheat (Triticum aestivum) local varieties, each containing genotypes for 33,709 genetic markers and eight traits (traits: Grain length (GL), Grain width (GW), Grain hardness (GH), Thousand-kernel weight (TKW), Test weight (TW), Sodium dodecyl sulphate sedimentation (SDS), Grain protein (GP) and plant height (PHT)). Gene markers were generated by the DArT-Seq platform and obtained by genotyping sequencing methods. Genomic heritability ($h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$) is shown in Table 1, and all gene markers are coded by 0 and 1, indicating the absence or presence of alleles, respectively. Since the data used in this study are not raw data, the phenotypic data have been normalized (mean = 0, standard deviation = 1). More details about this GS dataset can be found in the literature [10].

B. EXPERIMENTAL MODELS

1) RRBLUP PREDICTION MODEL

RRBLUP is one of the widely used regression models in genome-wide association analysis [16], and its corresponding standard linear regression equation is shown in equation (1).

$$y = \mu + Gg + \varepsilon \tag{1}$$

$y(n+1)$ is the vector corresponding to the phenotypic values, $G(n \times m)$ is the gene matrix, μ is the mean of the phenotypic vector y, $g(m \times 1)$ is the marker effect vector associated with the G matrix and g obeys a normal distribution $g \sim N(0, I\sigma_g^2)$, and $\varepsilon(n \times 1)$ is the random effect vector.

2) BAYESIAN PREDICTION MODEL

Five Bayesian regression methods (BayesA, BayesB, BayesC, Bayesian LASSO, Bayesian ridge regression) are applied in this study. Bayesian methods include three necessary conditions, prior, likelihood, and posterior. The prior probability is a quantitative indicator of the parameters' self-generation before the data are analyzed, and generally the parameters have a prior distribution of their own. The likelihood is the conditional probability, and the posterior probability is derived by combining the prior and the likelihood using Bayesian theory. The basic theory of Bayesian is as follows:

$$f(\theta \mid y) = \frac{f(y \mid \theta)f(\theta)}{f(y)} \propto f(y \mid \theta)f(\theta) \tag{2}$$

$f(\theta)$ is denoted as the prior probability density of θ, $f(\theta \mid y)$ is denoted as the likelihood value, and $f(y \mid \theta)$ is denoted as the posterior density of θ.

The prediction accuracy of various Bayesian models depends to a large extent on the suitability of their model assumptions to the genetic construction of the predicted phenotypes.When Meuwissen *et al.* [1] first proposed GS theory, they provided two Bayesian approaches for solving the problem that the number of genetic markers is usually much higher than the number of phenotypic records, namely BayesA and BayesB. BayesA assumes that all markers have effects, and that all genetic markers obey a positive-tax distribution with a scale inverse cardinality distribution, where the degrees of freedom and scale parameters are associated with genetic structure and are able to determine genetic structure [1], both of which are given a priori. BayesA uses Markov Chain Monte Carlo (MCMC) to construct Gibbs sampling chains, and solves for the markers in the model. BayesB differs from BayesA in the different prior assumptions on SNP effects. BayesA assumes that all genetic markers have effects, while BayesB assumes that only a small fraction of marker loci have effects and most other chromosomal segments have 0 effects (the proportion of invalid marker loci is π). BayesB uses a mixed distribution as the marker effect variance a priori, so it is difficult to construct fully conditional posterior distributions for the respective marker effects and variance, so BayesB uses Metropolis-Hasting (MH) sampling for joint sampling of markers and methods [19]. BayesC [22] differs from BayesB in that π is unknown and needs to be solved in the model to obtain it. Bayesian LASSO (BL) [20] assumes that genetic markers obey the Laplace distribution, which is equivalent to a normal distribution where the variance obeys an exponential distribution. The Bayesian ridge regression (BRR) [21] model assumes that all markers have the same genetic variance, and its theoretical model is the same as that of RRBLUP, but differs in that the BRR model is based on the MCMC method sampling and thus solving for marker effects.

TABLE 1. Phenotypic traits of wheat gene bank, number of observations, number of markers and heritability of the trait.

Trait	Number of observations	Number of Markers	Heritability (h^2)
Thousand-kernel weight (TKW)	2000	33,709	0.833
Test weight (TW)	2000	33,709	0.754
Grain hardness (GH)	2000	33,709	0.839
Grain protein (GP)	2000	33,709	0.625
Grain length (GL)	2000	33,709	0.881
SDS sedimentation (SDS)	2000	33,709	0.681
Grain width (GW)	2000	33,709	0.848
Plant height (PHT)	2000	33,709	0.434

3) GBDT PREDICTION MODEL

Ensemble learning contains three main methods, boosting, bagging, and stacking [41]. Gradient boosting [42], [43] algorithm is a machine learning technique for regression, classification, and ranking tasks and is part of the boosting algorithm family, which builds a learner capable of reducing the loss along the steepest gradient at each step of the iteration to compensate for the deficiencies of the existing model, and can boost weak learners to strong learners. The training process of the algorithm is tandem, and the training of the weak learners is sequential, with each weak learner learning from the previous one and finally combining the predictions of all the learners to produce the final prediction results.

This study mainly uses decision trees as the base learner, reduces the fitting ability of a single decision tree by suppressing the complexity of the decision tree, and then integrates multiple decision trees by gradient boosting to construct the gradient boosting decision tree algorithm (GBDT), which can finally solve the overfitting problem well. The GBDT algorithm, also called MART (Multiple additive regression) [44], [45], is the iterative decision tree algorithm, which can be viewed as an additive model composed of M trees, and its corresponding formula is shown in equation (3), where x is the input sample; w is the model parameter; h is the regression tree, and α is the weight of each tree.

$$F(x, w) = \sum_{m=0}^{M} \alpha_m h_m (x, w_m) = \sum_{m=0}^{M} f_m (x, w_m) \tag{3}$$

The GBDT model predicts the genome as shown in Figure 1, with a decision tree type using CART [46].

- Initialize the first weak learner $f_0(x)$, the weak learner is defined in equation (4). The squared difference $L(Y_i, f(x_i))$ is chosen as the loss function in the GBDT model as in equation (6), where Y_i are the observed phenotype value, $f(x_i)$ is the predicted phenotype value, and the squared loss function is convex, The direct derivation yields $f_0'(x) = c = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i\right)$, the phenotypic mean of the training sample.

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(Y_i, c) \tag{4}$$

$$r_{m,i} = -\left[\frac{\partial L(Y_i, f(x_i))}{\partial f(x)}\right]_{f(x)=f_{m-1}(x)} \tag{5}$$

$$L(Y_i, f(x_i)) = (Y_i - f(x_i))^2 \tag{6}$$

- Construct $m(m = 1, 2, 3, \ldots, M)$ classification regression trees, and calculate the residual value $r_{m,i}$ corresponding to the $i(i = 1, 2, 3, \ldots, N)$ sample in the m_{th} tree. The calculation formula is shown in equation (5). The residuals are used as the true phenotype values of the training samples to train the next tree, and the mth regression tree is obtained. Its corresponding leaf node region is $R_{m,j}, j = 1, 2, \ldots, J_m$, J_m is the number of leaf nodes of the m_{th} regression tree. The process requires finding all possible best division nodes of CART [46] and calculating the squared loss of the two sets of data after splitting, SE_l is the squared loss of the left node and SE_r is the squared loss of the right node so that the division node with the smallest value of $SE_l + SE_r$ is the best, followed by calculating the best fit value $c_{m,j}$ of the leaf node regions $j = 1, 2, \ldots, J_m$ as in equation (7), update the learner $f_m(x)$ as in equation (8), and after m iterations get M decision trees and integrate them to get the strong learner as in equation (9).

$$c_{m,j} = \arg\min \sum_{x \in R} L(Y_i, f_{m-1}(x_i) + c) \tag{7}$$

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} c_{m,j} I\left(x \in R_{m,j}\right) \tag{8}$$

$$F_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J_m} c_{m,j} I\left(x \in R_{m,j}\right) \tag{9}$$

- The constructed training set is fed into the GBDT predictor for testing.
- Finally, the sum of the prediction scores of these decision trees and the product of the learning rate is used as the final prediction result.

C. MODELS PERFORMANCE EVALUATION

The whole experimental procedure uses the Hold-Out [51] validation method to extract 70% of the original data for model training and 30% of the data for model

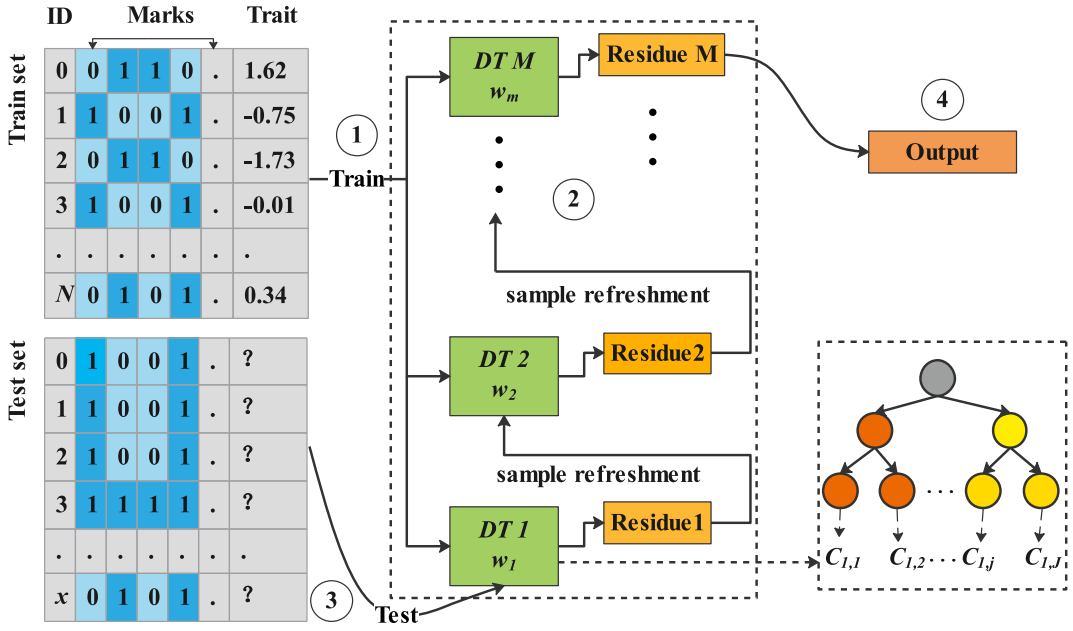


FIGURE 1. General architecture of gradient boosting decision tree algorithm.

performance validation. The Pearson correlation coefficient (PCC) was used to evaluate the prediction performance of different models, and the mean normalized discounted cumulative gain (MNDCG) method [40], [47] was used to evaluate the ability of different models to predict individuals with high phenotypic values. PCC and MNDCG are defined as follows.

$$PCC_{(Y,\hat{Y})} = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2}} \tag{10}$$

$$MNDCG_{(K,\hat{Y},Y)} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^{K} y(i,\hat{Y})d(i)}{\sum_{i=1}^{K} y(i,Y)d(i)} \tag{11}$$

where N in equation (10) denotes the number of test samples, and Y_i and $\hat{Y}_i$ denote the phenotypic observations and phenotypic predictions at i, respectively, where i takes a range of values $1 \leq i \leq N$, $\bar{Y}$ denotes the mathematical expectation of phenotypic observations, and $\overline{\hat{Y}}$ denotes the mathematical expectation of phenotypic predicted values. The $d(i) = 1/\left(\log_2 i + 1\right)$ in Eq. (11) denotes the monotonically decreasing discount function at i, $y(i, Y)$ denotes the i_{th} phenotypic observation, and Y is in descending order, $y(1, Y) \geqslant y(2, Y) \geqslant \cdots y(N, Y)$, $y(i, \hat{Y})$ is the i_{th} phenotypic predicted value corresponding to Y in the two-dimensional score matrix $(\hat{Y}, Y)$, where $\hat{Y}$ is in descending order, and the higher value of MNDCG indicates that the GS prediction model performs better in selecting the first K individuals with higher phenotypic values ($\alpha = K/N$; $1 \leqslant K \leqslant N$; $1\% \leqslant \alpha \leqslant 100\%$, N is the number of test samples), and $MNDCG = y(i, \hat{Y})/y(i, Y)$ when K = 1.

D. EXPERIMENTAL ENVIRONMENT

Experimental hardware environment configuration Inter(R) Core(TM) i7-10700kCPU@3.80Hz processor, NVIDIA Quadro P400 graphics card, 8GB running memory, 1T hard disk capacity. The experimental software environment is JetBrains PyCharm Community Edition 2019.2.4 × 64 (https://www.jetbrains.com/pycharm/) and RStudio (https://rstudio.com/). The GBDT model is based on the sklearn framework (https://sklearn.apachecn.org/), and the RRBLUP model is based on the "rrBLUP" package (https://cran.r-project.org/web/packages/rrBLUP/index.html), the BayesA, BayesB, BayesC, BL, and BRR models are based on the R language package "BGLR" (https://cran.r-project.org/web/packages/ BGLR/index.htm) for experiments.

III. RESULTS

A. COMPARISON OF THE PREDICTION PERFORMANCE OF GBDT AND THREE MACHINE LEARNING MODELS

We constructed four GS prediction models based on the sklearn framework, including gradient boosting decision tree (GBDT), random forest (RF), artificial neural network (ANN), and k-nearest neighbor (KNN) algorithm, and evaluated the predictions of these four models using the wheat dataset. The experimental results showed (Figure 2) that GBDT had the highest prediction correlation for eight traits (TKW, TW, GL, GW, GH, GP, SDS, and PHT) with PCC values of 0.615, 0.564, 0.689, 0.706, 0.542, 0.527, 0.402, and 0.352, respectively; KNN had the lowest prediction correlation with PCC values of Compared with the ANN, RF and KNN models, the average prediction accuracy of GBDT for the eight traits was improved by 7.82%, 3.12%, and 24.78%, respectively.

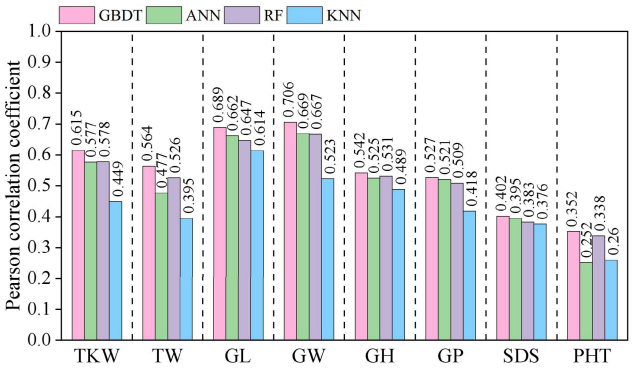


FIGURE 2. Pearson correlation coefficients of GBDT, ANN, RF and KNN models for eight tested traits.

B. COMPARISON OF THE PREDICTION PERFORMANCE OF GBDT AND FIVE BAYESIAN MODELS

Genomic predictions were made for eight test traits of wheat using BayesA, BayesB, BayesC, BRR, BL, and GBDT models, and the predictive performance of the models was evaluated using Pearson correlation coefficients of phenotypic predictions and phenotypic observations. We verified the prediction accuracy of Bayesian models for different numbers of iterations, and their prediction performance leveled off when 200,000 iterations were performed. The results in Table 2 show that for the trait GW, the GBDT model has the highest prediction accuracy with PCC values of 0.706; for several other traits, BayesA, BayesB, BayesC, BRR, and BL have similar prediction results with average prediction accuracy of 0.591, 0.592, 0.591, 0.590 and 0.589, respectively, relative to the overall improvement of prediction accuracy of GBDT is 7.60%. In addition, Bayesian models are far less computationally efficient than the GBDT model when 200,000 iterations are performed. For PHT trait with lower heritability, each model showed lower prediction accuracy than several other traits, indicating that the prediction performance of the models was somewhat influenced by the magnitude of heritability.

The MNDCG was also used to evaluate the ability of the model to predict individuals with high phenotypic values (Figure 3). At different α levels, the MNDCG values of GBDT were higher than the five Bayesian models for traits TKW ($53\% \leq \alpha \leq 100\%$), GW ($1\% \leq \alpha \leq 100\%$) and GP ($1\% \leq \alpha \leq 18\%$); for traits TW ($1\% \leq \alpha \leq 27\%$), GL ($1\% \leq \alpha \leq 100\%$), GH ($1\% \leq \alpha \leq 100\%$), GP ($1\% \leq \alpha \leq 100\%$), SDS ($1\% \leq \alpha \leq 100\%$) and PHT ($5\% \leq \alpha \leq 100\%$), the MNDCG values of GBDT were overall higher than those of BRR; for traits GL and GH, the MNDCG values of BayesA, BayesB, BayesC, and BL were not significantly different when $1\% \leq \alpha \leq 100\%$ and were all higher than those of GBDT.

C. COMPARISON OF THE PREDICTION ACCURACY OF GBDT AND RRBLUP MODELS

The prediction accuracy of the GBDT and RRBLUP models are discussed separately in this subsection. The results in Table 3 show that the correlation between phenotypic predictions and observations is higher for the RRBLUP model than for the GBDT. Studies have shown that the RRBLUP model still exhibits good predictive power in genomic prediction [37], [54]. Based on this, further comparison of the predictive ability of the RRBLUP and GBDT models for individuals with high phenotypic values (Figure 4) showed that for traits GH (Figure 4.e), SDS (Figure 4.g), GL (Figure 4.c) and PHT (Figure 4.h), the MNDCG values of RRBLUP were higher than those of GBDT overall when $1\% \leq \alpha \leq 100\%$, where the GBDT model had the lowest MNDCG values of 0.093 and -0.016 for traits GH ($\alpha = 97\%$) and SDS ($\alpha = 100\%$), respectively; for traits GW ($1\% \leq \alpha \leq 100\%$) (Figure 4.d), TKW ($60\% \leq \alpha \leq 100\%$) (Figure 4.a), TW ($8\% \leq \alpha \leq 17\%$) (Figure 4.b), and GP ($1\% \leq \alpha \leq 73\%$) (Figure 4.f), the MNDCG values of GBDT were higher than those of RRBLUP.

In terms of model prediction accuracy, the overall prediction accuracy of the GBDT model for the eight traits was lower than that of RRBLUP, and in terms of the ability of the models to select individuals, both models showed their respective selection advantages for different traits. For some individuals with higher phenotypic values, GBDT also showed good selection ability. In addition, GBDT differs from the RRBLUP model in that the GBDT model does not make a priori assumptions about the markers, and its prediction process uses a different feature extraction strategy by building a regression tree through a gradient boosting algorithm and traversing all features to select important feature nodes. The prediction process of GBDT is based only on features with high importance and can accomplish both feature selection and phenotype prediction, so this unique computational approach also greatly improves its computational efficiency in the phenotype prediction process.

D. EFFECT OF THE NUMBER OF MARKERS ON MODEL PREDICTION PERFORMANCE

In the course of this study, to investigate the effect of the selection of different marker numbers on the prediction accuracy of GBDT and RRBLUP models, a subset of markers with different dimensions (Markers = 1000, 2000, 5000, 8000, 12000, 15000, 18000, 20000) were selected to construct the prediction models, and the Hold-Out [51] method was used to validate for each of the different marker subsets.

The PCC of both GBDT and RRBLUP models were affected to some extent when the number of markers changed. For traits GL, GH, and PHT, the PCC of GBDT and RRBLUP showed a significant increasing trend as the number of markers increased, however, for the other five traits, the PCC showed a more moderate increasing trend as the number of markers increased. For individual models, when the number of markers varied, the PCC was highest for TKW (2K), TW (8K), GL (15K), GW (8K), GH (15K), GP (12K), SDS (15K), and PHT (20K) with 0.640, 0.576, 0.699, 0.717, 0.558 0.537, 0.418 and 0.332; for traits TKW (15K), TW (20K), GL (20K),

TABLE 2. Pearson correlation coefficients of six GS models for eight tested traits. (Notes: TKW: Thousand-kernel weight; TW: Test weight; GL: Grain length; GW: Grain width; GH: Grain hardness; SDS: Sodium dodecyl sulphate sedimentation; GP: Grain protein; PHT: Plant height).

GS Models	Eight test traits								
	TKW	GL	TW	GH	GP	SDS	PHT	GW	Mean
GBDT	0.615	0.689	0.564	0.542	0.527	0.402	0.352	0.706	0.550
BayesC	0.633	0.723	0.610	0.684	0.561	0.441	0.376	0.703	0.591
BayesB	0.634	0.722	0.610	0.683	0.562	0.441	0.380	0.704	0.592
BRR	0.632	0.722	0.611	0.683	0.577	0.434	0.377	0.702	0.590
BayesA	0.634	0.720	0.619	0.683	0.560	0.439	0.371	0.703	0.591
BL	0.622	0.721	0.612	0.670	0.561	0.442	0.376	0.705	0.589

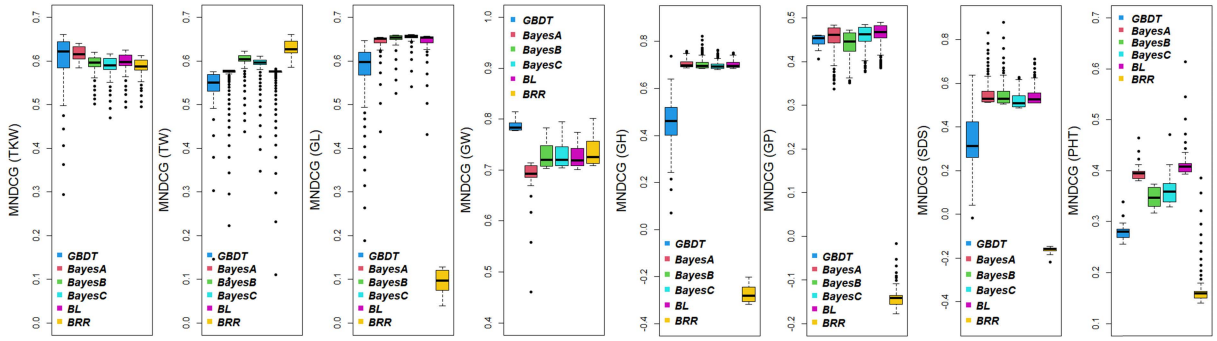


FIGURE 3. Box plot of the MNDCG value for GBDT and five Bayesian models with top-ranked α increasing from 1% to 100%.

TABLE 3. Pearson correlation coefficients of CBDT and RRBLUP for eight tested traits. (Notes: TKW: Thousand-kernel weight; TW: Test weight; GL: Grain length; GW: Grain width; GH: Grain hardness; SDS: Sodium dodecyl sulphate sedimentation; GP: Grain protein; PHT: Plant height).

GS Models	Eight test traits								
	TKW	GL	TW	GH	GP	SDS	PHT	GW	Mean
GBDT	0.615	0.689	0.564	0.542	0.527	0.402	0.352	0.706	0.550
RRBLUP	0.633	0.756	0.629	0.677	0.544	0.494	0.377	0.730	0.605

GW (15K), GH (20K), GP (8K), SDS (20K) and PHT (15K), RRBLUP had the highest PCC with 0.633, 0.626, 0.785, 0.731 0.651, 0.552, 0.487 and 0.384; for traits TKW, GW, and GP, the PCC values and trends of the GBDT and RRBLUP models were similar. However, the predictive performance of RRBLUP for the eight tested traits was slightly higher than that of the GBDT model (Figure 5).

The MNDCG results showed (Figure 6) that the GBDT model predicted the highest mean MNDCG values of 0.492, 0.662, 0.826, 0.389, 0.670, 0.482, 0.541, and 0.283 for traits GP (2K), GL (5K), GW (5K), SDS (5K), TKW (8K), GH (8K), TW (18K) and PHT (20K), respectively when different numbers of markers were used. The RRBLUP model predicted the highest mean MNDCG for traits GP (2K), GL (5K), GW (12K), SDS (20K), TKW (12K), GH (18K), TW (5K) and PHT (15K), with 0.480, 0.668, 0.631, 0.680, 0.751, 0.420, 0.736, and 0.641, respectively. For traits GH, TW, SDS, and PHT, although the MNDCG values of the GBDT model were lower than those of RRBLUP, there was a significant improvement in the MNDCG values of the GBDT model as the number of markers increased (Figure 6).

At different α, when the number of markers was 1K, except for traits TW and PHT, for traits TKW ($1\% \leq \alpha \leq 100\%$), GL ($93\% \leq \alpha \leq 100\%$), GW ($1\% \leq \alpha \leq 100\%$), GH ($97\% \leq \alpha \leq 100\%$), GP ($1\% \leq \alpha \leq 59\%$), and SDS ($1\% \leq \alpha \leq 2\%$), the MNDCG values of GBDT were higher than RRBLUP by 3.91%, 0.30%, 8.76%, 8.67%, 6.66%, and 6.50%, respectively (Figure 7.a).

When the number of markers was 2K, the MNDCG values of GBDT increased over RRBLUP for traits TKW ($3\% \leq \alpha \leq 100\%$), GL ($22\% \leq \alpha \leq 100\%$), GW ($1\% \leq \alpha \leq 100\%$), GH ($98\% \leq \alpha \leq 100\%$), and GP ($1\% \leq \alpha \leq 47\%$), in addition to traits TW, SDS, and PHT, respectively 4.18%, 2.62%, 5.88%, 10.48%, and 5.32%, respectively (Figure 7.b).

When the number of markers was 5K, except for traits TW and PHT, for traits TKW ($50\% \leq \alpha \leq 100\%$), GL

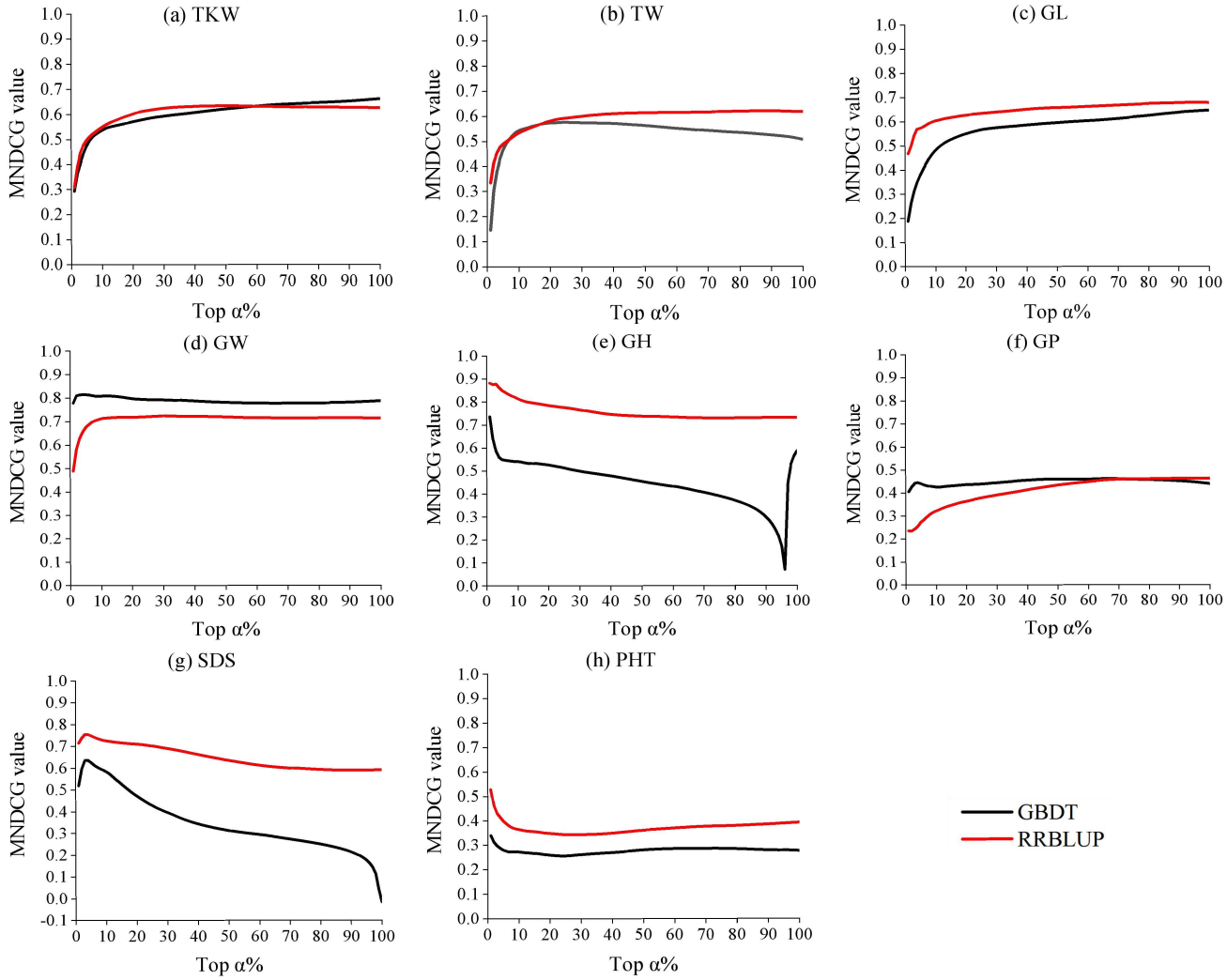


FIGURE 4. MNDCG value curves for GBDT and RRBLUP with top-ranked α increasing from 1 to 100%.

(1% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (97% $\leq \alpha \leq$ 100%), GP (1% $\leq \alpha \leq$ 56%), and SDS (1% $\leq \alpha \leq$ 6%), the MNDCG values of GBDT over RRBLUP were increased by 2.41%, 3.12%, 7.61%, 8.58%, 4.56%, and 8.24%, respectively (Figure 7.c).

When the number of markers was increased to 8K, except for traits TW, SDS and PHT, for traits TKW (48% $\leq \alpha \leq$ 100%), GL (31% $\leq \alpha \leq$ 56%, 84% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (98% $\leq \alpha \leq$ 100%), and GP (1% $\leq \alpha \leq$ 51%), the MNDCG values of GBDT over RRBLUP by 2.09%, 0.56%, 7.87%, 9.23%, and 2.00%, respectively (Figure 7.d).

When the number of markers was 12K, the MNDCG values of GBDT increased over RRBLUP for traits TKW (82% $\leq \alpha \leq$ 100%), GL (17% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (98% $\leq \alpha \leq$ 100%), and GP (1% $\leq \alpha \leq$ 60%), in addition to traits TW, SDS, and PHT, respectively 0.64%, 0.78%, 5.75%, 9.20%, and 3.47%, respectively (Figure 7.e).

When the number of markers was 15K, except for traits TW, SDS and PHT, the MNDCG values for traits TKW (71% $\leq \alpha \leq$ 100%), GL (78% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (98% $\leq \alpha \leq$ 100%), GP (1% $\leq \alpha \leq$ 98%), and GBDT increased over RRBLUP by 0.72%, 0.72%, 10.12%, 7.44%, and 6.52%, respectively (Figure 7.f).

When the number of markers was 18K, except for traits SDS and PHT, for traits TKW (67% $\leq \alpha \leq$ 100%), GL (12% $\leq \alpha \leq$ 39%, 85% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (98% $\leq \alpha \leq$ 100%), GP (1% $\leq \alpha \leq$ 92%), and TW (5% $\leq \alpha \leq$ 15%), the GBDT MNDCG values increased by 1.10%, 0.48%, 11.48%, 5.75%, 5.40%, and 0.51%, respectively, compared to RRBLUP (Figure 7.g).

When the number of markers was 20K, except for traits TW, SDS and PHT, the MNDCG values for traits TKW (51% $\leq \alpha \leq$ 100%), GL (98% $\leq \alpha \leq$ 100%), GW (1% $\leq \alpha \leq$ 100%), GH (98% $\leq \alpha \leq$ 100%), GP (1% $\leq \alpha \leq$ 86%),

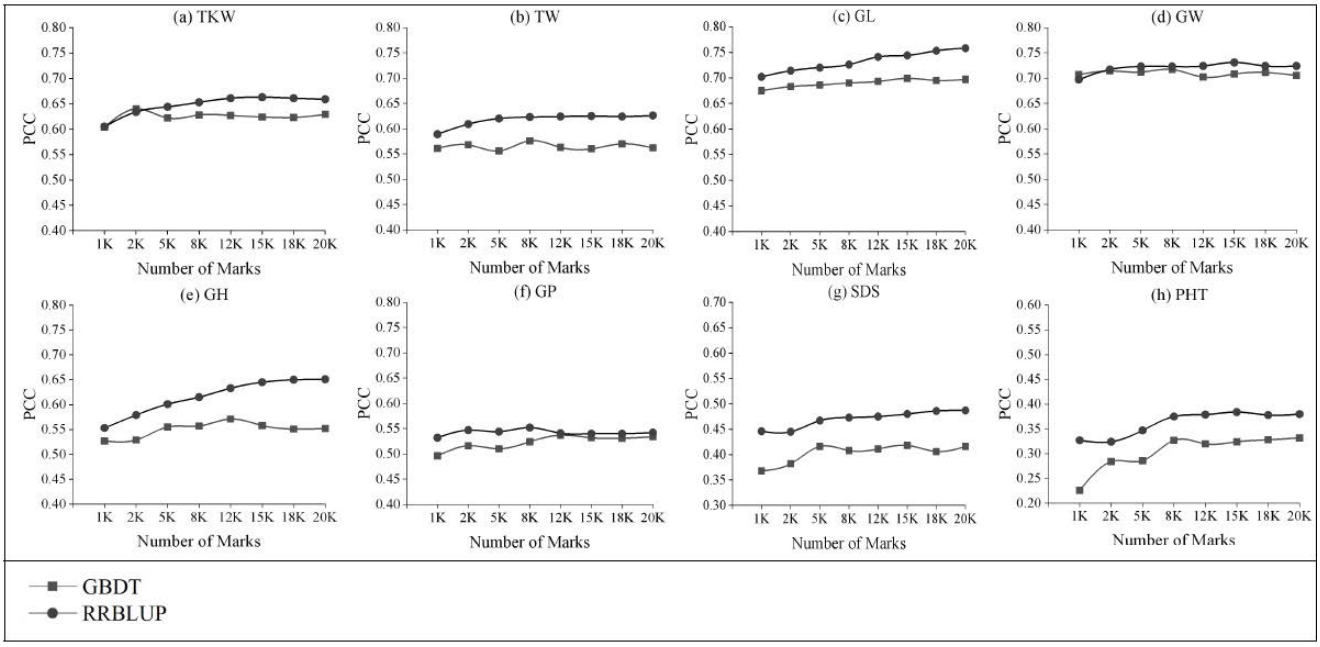


FIGURE 5. PCC curves for GBDT and RRBLUP models when subsets of different markers were used.

and GBDT were improved over RRBLUP by 2.62%, 0.11%, 10.87%, 15.35%, and 4.61%, respectively (Figure 7.h).

IV. DISCUSSION

Genome-wide prediction is an emerging technique for predicting unknown phenotypes using genome-wide marker coverage, and with the continuous updating and increasing maturity of sequencing technologies and the increasingly low cost of genotyping, genome-wide prediction is gradually being promoted in plant and animal breeding [51-52], and the selection of genomic models play a crucial role and directly affects the prediction of phenotypes. In this study, we mainly used an integrated learning approach (GBDT) for genomic prediction in wheat and compared the prediction performance with five typical Bayesian models (BL, BRR, BayesA, BayesB, BayesC) and RRBLUP model, in addition, the prediction performance of each model for the top K ranking of individual phenotypic values was analyzed in the study, and the effect of a different number of markers on the prediction performance of GBDT and RRBLUP models was explored.

Overall, the prediction results of GBDT and the five Bayesian models were relatively similar, but their applicability differed for the different traits tested, with the GBDT model having the highest prediction performance for traits TKW and GW, the BL model for traits PHT and GL, and the BayesA model for traits SDS, GP, GH, and TW. The results demonstrated that it is difficult to have optimal methods in the genomic prediction that are adapted to all traits [50], [52], and even when Bayesian theoretical models are used in different population experiments, the prediction performance varies because of the differences between traits [19]. And even before this study, Bayesian theoretical models were used by Mathieu Blondel *et al.* to predict maize, rice, and barley and compared with another ensemble learning method (Gradient boosted regression trees, GBRT) [47]. For the barley dataset, the average result of GBRT prediction was 0.554, the average result of BayesC was 0.593, and the average result of BL was 0.581; for the maize dataset, the average result of GBRT prediction was 0.419, the average result of BayesC was 0.393, and the average result of BL was 0.383; for the rice dataset, the average result of GBRT prediction of 0.713, BayesC of 0.688, and BL of 0.714 for the rice dataset, and these previous findings again support the conclusions we reached.

However, Bayesian models tend to have more parameters with estimation, which brings more computational effort while improving prediction accuracy. Half of the parameter solution process first assumes the distribution type of the variables to be sought in the model, i.e., assumes the prior distribution of the parameters, determines the joint distribution of the samples and parameters, and derives the posterior distribution of the parameters according to Bayes' theorem, constructs a Markov Chain Monte Carlo (MCMC), samples based on Gibbs or Metropolis-Hasting (MH) sampling method, set a sufficient number of iterations until convergence, and reach a smooth state. In the study, the Bayesian model is set to the default 15,000 iterations, and to determine whether the Bayesian model is fully converged, we set a series of gradient iterations of 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, 100,000, and 200,000, and the number of burn in under each iteration is set to the total number of iterations. number of iterations is set to 50% of the total number of

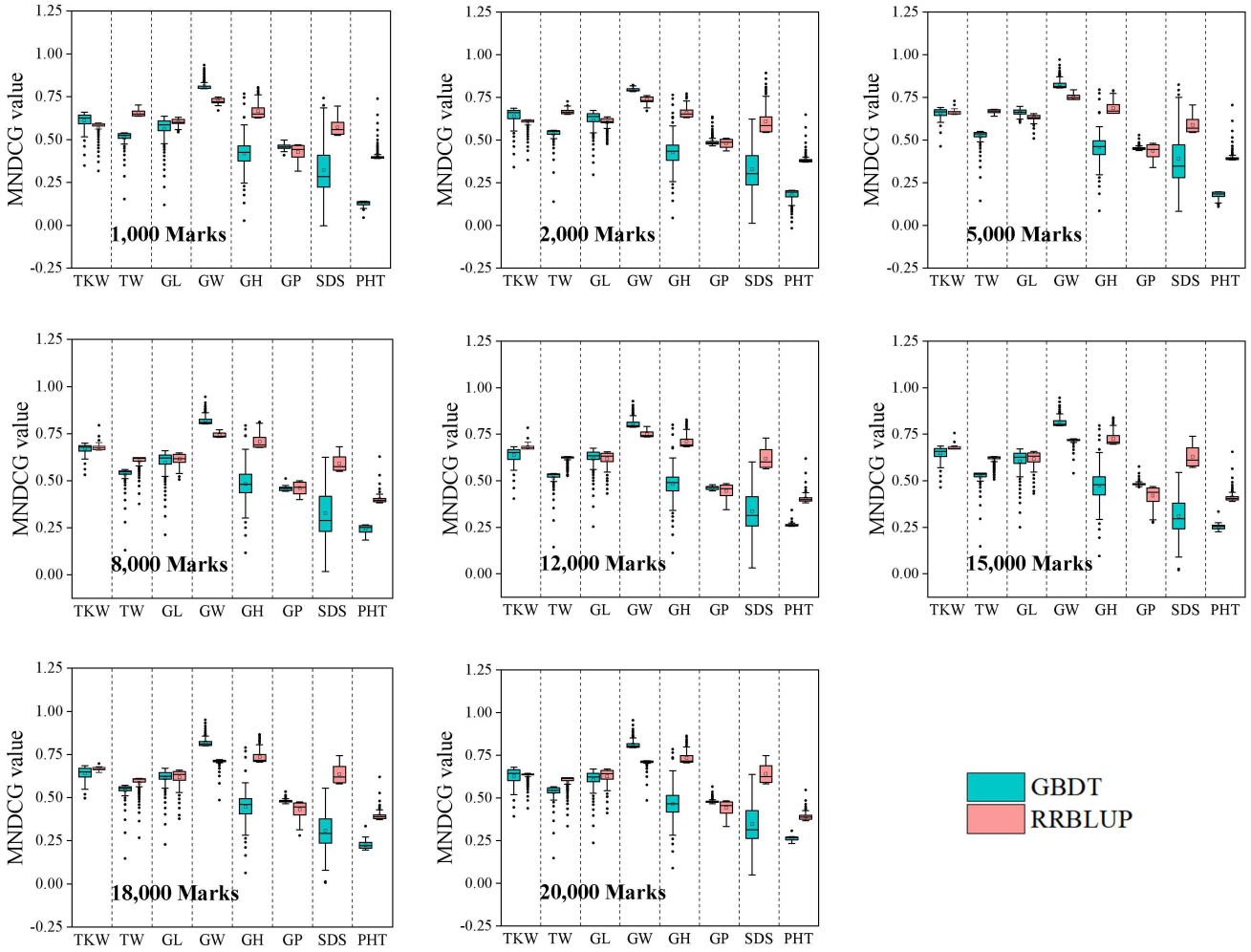


FIGURE 6. Box plot of the MNDCG value for GBDT and RRBLUP when subsets of markers were used with top-ranked α increasing from 1% to 10%.

iterations. Figure 8 shows the trend of the prediction accuracy of the five Bayesian models with increasing number of iterations. GBDT and RRBLUP in the figure do not involve MCMC iterations, and their prediction accuracy is used as the reference standard. It can be seen from Figure 8 that the prediction accuracy of the five Bayesian models gradually improves as the total number of MCMC iterations increases, but the MCMC needs to re-estimate all marker effect values for each iteration and the process is continuous and non-parallel, which consumes a large amount of computation time, as shown in Figure 9. The high accuracy of Bayesian methods is based on the result of successful convergence, and when the iterative process fails to converge successfully, it leads to low prediction accuracy. Therefore, it is a challenge for Bayesian methods to set the number of iterations when the genetic structure of traits is unknown, which to some extent limits its application in plant and animal breeding practices with strong time-sensitive needs.

In this study, the average predictive performance of the GBDT model for the eight traits was lower than that of the RRBLUP model, and the experimental results suggest that RRBLUP is still a valid genomic prediction model [37]. However, in past studies, most researchers have singularly used Pearson correlation coefficient as a method to evaluate the goodness of genomic prediction models; in fact, the correlation coefficient is easily influenced by extreme individual phenotypic values, and its magnitude can only indicate the fit between phenotypic predictions and observations and the feasibility of the model and does not fully represent the goodness of the prediction model [53]. Therefore, Mathieu Blonde [47], Wenlong Ma [36], and others used the MNDCG method to measure the ability of GS models to select individuals with high breeding values before K. In this study, although the RRBLUP model predicted higher correlations than the GBDT model, for some traits, the GBDT model still had a higher ability to select individuals with high phenotypic values than RRBLUP.

However, in genomic prediction, many factors affect the predictive performance of the models. We grouped these factors into three categories: the first category is the influence

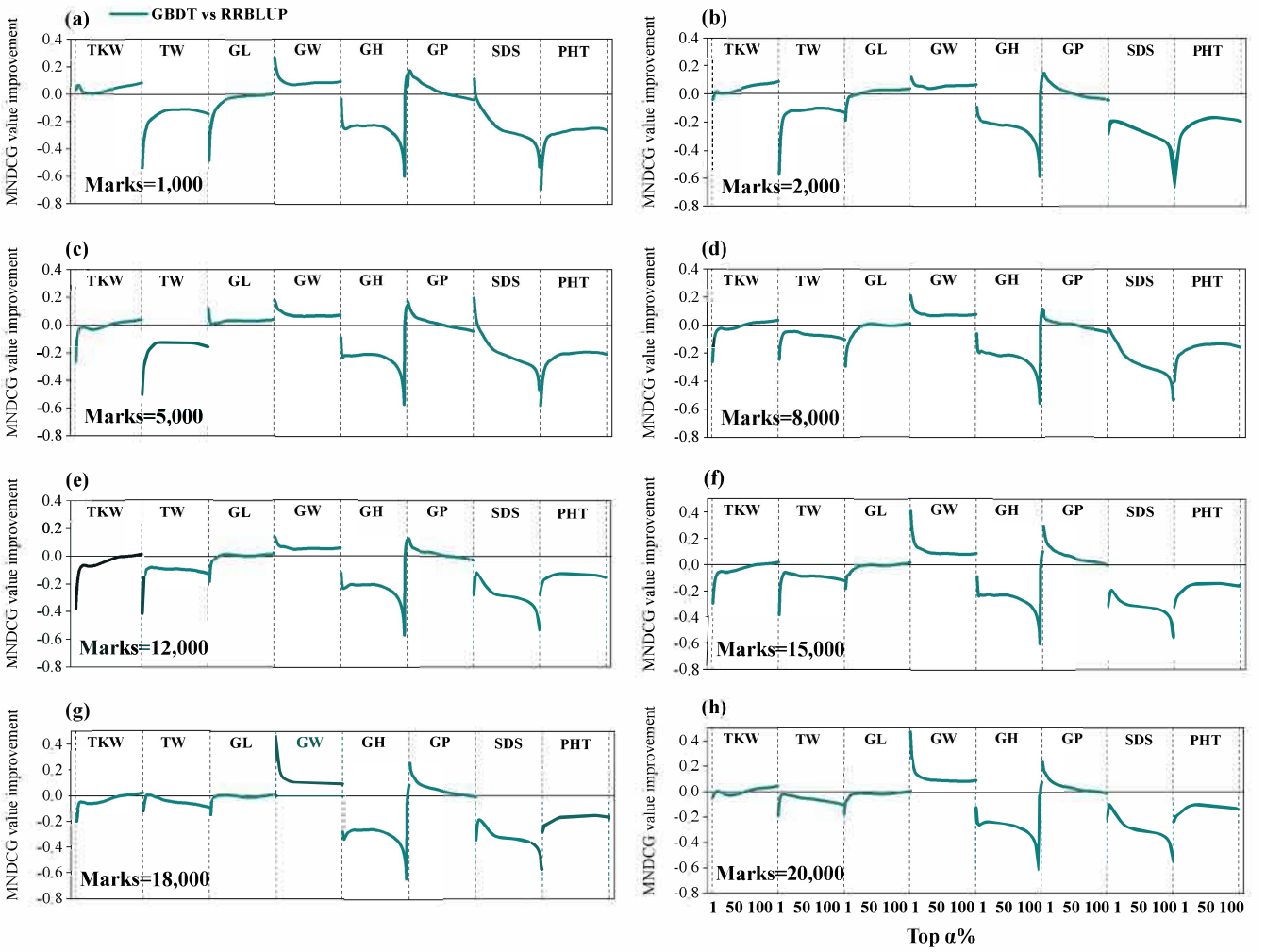


FIGURE 7. Improvement of GBDT over RRBLUP for eight traits when subsets of markers were used with top-ranked α increasing from 1% to 100%.

brought by the data itself, containing the training population size, heritability and genetic structure of traits, genetic marker types and sampling strategies, minimum allele frequencies, and the number of markers and QTL. The second category is the limitations of the models themselves, including algorithm design, model a priori assumptions, parameter selection, and type of model. The third category is the choice of validation method, which contains K-fold cross-validation, leave-one-out cross-validation, and leave-out cross-validation. This study focuses on validating the effect of the number of markers on the GBDT and RRBLUP models. The experimental results also show that the prediction performance of the models is affected by the marker size, but not regularly and that the ability of the models to select individuals with high phenotypic values varies depending on the marker subset size [36], [42], so the selection of the appropriate size of marker data has a significant impact on the prediction performance of the models.

Hold-Out validation was used for the experiments in this study, and the GBDT model was optimized by continuously optimizing the model parameters during the training process, and finally, the number of weak learners was adjusted to 12000, the learning rate was 0.02, the maximum depth of the decision tree was 5, the maximum number of leaf nodes was 20, and the minimum sample weight sum of leaf nodes was 0.025, so the GBDT model may still have some limitations, the experimental accuracy may not be the best, and the parameters of the model need to be adjusted for different experimental groups.

In the next study, we will try to further explore the research in the following aspects: (1) perform GS linear regression prediction experiments with more complex algorithms and richer marker datasets using GPU acceleration techniques; (2) downscale genetic marker data using data downscaling methods in machine learning to reduce computing time and improve model efficiency; (3) treat genomic selection as a classification problem and try to use more complex (3) treat genomic selection as a classification problem and try to perform nonlinear binary or multiclassification GS experiments using more sophisticated machine learning techniques and

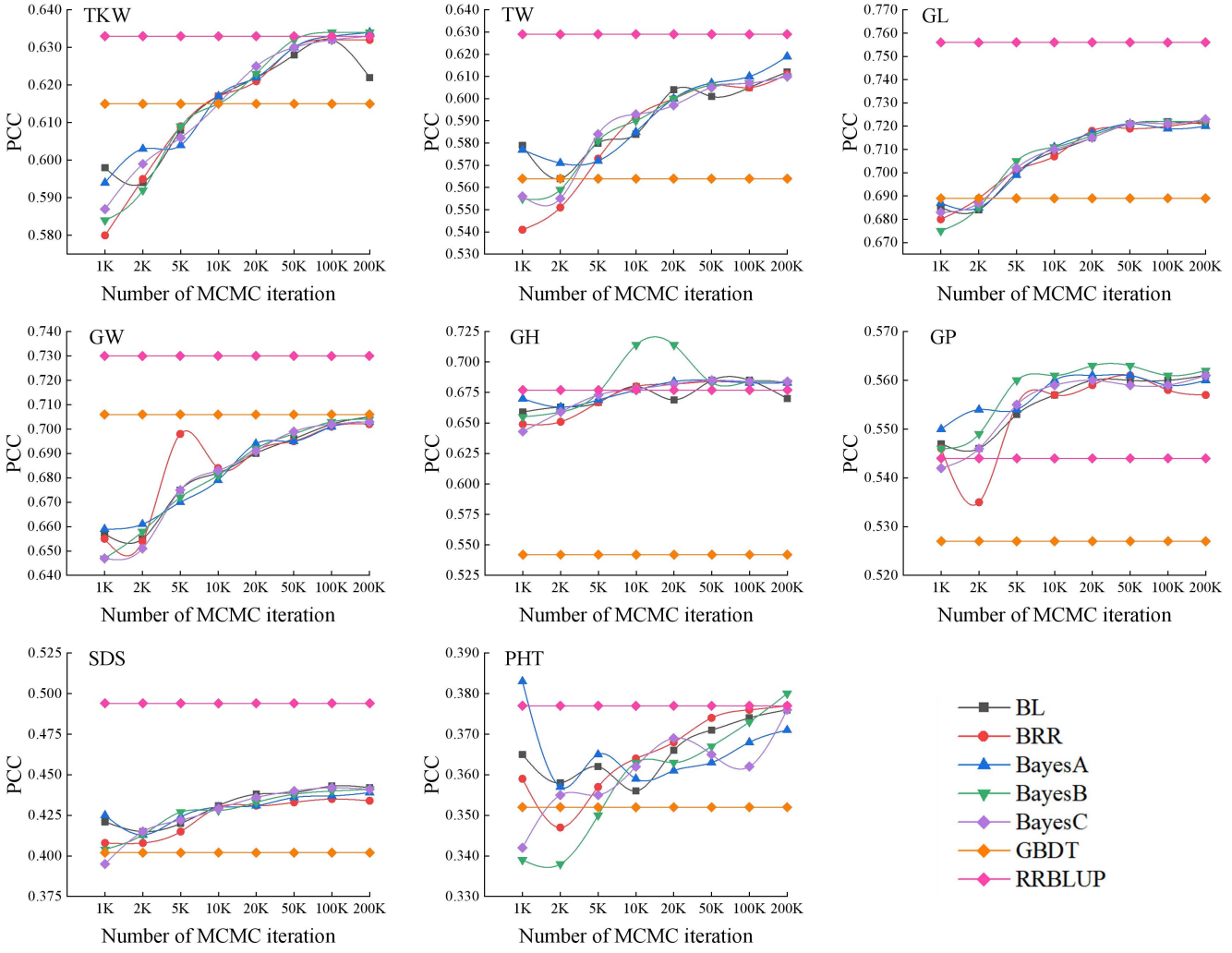


FIGURE 8. Trend of prediction accuracy of Bayesian model with different number of MCMC iterations.

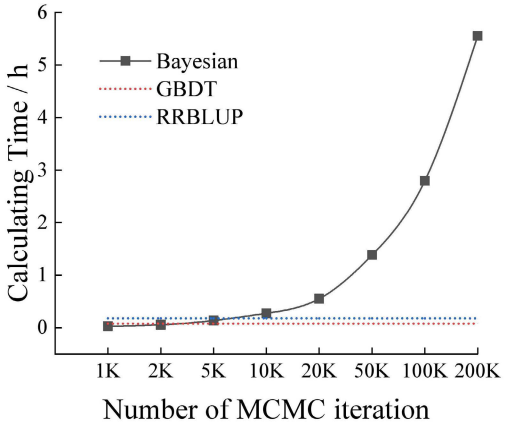


FIGURE 9. Trend of computational efficiency of Bayesian model with different number of MCMC iterations.

artificial intelligence techniques; (4) combine genome-wide association analysis (GWAS) methods to filter marker data and select genotypic marker data with high genetic effects as target data, which can avoid the impact of invalid genetic markers on model prediction performance; (5) try to use more GS data from different populations for model validation.

V. CONCLUSION

In this study, genomic prediction analysis was performed for eight traits of wheat TKW, TW, GL, GW, GH, GP, SDS, and PHT using seven methods: GBDT, RRBLUP, BayesA, BayesB, BayesC, BL, and BRR. It was found that (1) except for the GW trait, the prediction accuracy of GBDT was lower than the overall prediction performance of the five Bayesian models, but the overall computational efficiency of GBDT was higher than that of the Bayesian models. (2) The high accuracy of Bayesian methods is based on the result of successful convergence, and when the iterative process fails to converge successfully, it leads to low prediction accuracy. Therefore, how to set the number of iterations is a challenge for Bayesian methods when the genetic structure of traits is unknown, which to a certain extent limits its application in breeding practice due to the high requirement of timeliness in

plant and animal breeding. (3) RRBLUP is still a promising GS model, and its overall prediction effect is better than several other models. In addition, the prediction accuracy of the model is affected by the size of the number of markers to a certain extent. (4) For the PHT traits with low heritability, all seven models showed low prediction performance. Therefore, how to optimize the models and improve the computational efficiency of GS models while ensuring higher accuracy and robustness is the direction of future genome-wide prediction model optimization.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their helpful comments and the International Maize and Wheat Improvement Center (CIMMYT) for providing data support for this study.

REFERENCES

[1] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, Apr. 2001.

[2] T. Meuwissen, B. Hayes, and M. Goddard, "Genomic selection: A paradigm shift in animal breeding," *Animal Frontiers*, vol. 6, no. 1, pp. 6–14, 2016.

[3] J. J. Hayward, M. G. Castelhano, K. C. Oliveira, E. Corey, C. Balkman, T. L. Baxter, M. L. Casal, S. A. Center, M. Fang, S. J. Garrison, S. E. Kalla, P. Korniliev, M. I. Kotlikoff, N. S. Moise, L. M. Shannon, K. W. Simpson, N. B. Sutter, R. J. Todhunter, and A. R. Boyko, "Complex disease and phenotype mapping in the domestic dog," *Nature Commun.*, vol. 7, no. 1, pp. 1–11, Apr. 2016.

[4] H. D. Daetwyler *et al.*, "Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle," *Nature Genet.*, vol. 46, no. 8, pp. 858–865, Aug. 2014.

[5] H. Song, J. Zhang, Q. Zhang, and X. Ding, "Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig," *Frontiers Genet.*, vol. 9, p. 730, Jan. 2019.

[6] J. Spindel, H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña, G. Atlin, J.-L. Jannink, and S. R. McCouch, "Genomic selection and association mapping in Rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of Rice genomic selection in elite, tropical Rice breeding lines," *PLOS Genet.*, vol. 11, no. 2, Feb. 2015, Art. no. e1004982.

[7] C. Guzman, R. J. Peña, R. Singh, E. Autrique, S. Dreisigacker, J. Crossa, J. Rutkoski, J. Poland, and S. Battenfield, "Wheat quality improvement at CIMMYT and the use of genomic selection on it," *Appl. Transl. Genomics*, vol. 11, pp. 3–8, Dec. 2016.

[8] J. J. Marulanda, X. Mi, A. E. Melchinger, J.-L. Xu, T. Würschum, and C. F. H. Longin, "Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, Rice and triticale," *Theor. Appl. Genet.*, vol. 129, no. 10, pp. 1901–1913, Oct. 2016.

[9] J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. De Los Campos, and J. Burgueño, "Genomic selection in plant breeding: Methods, models, and perspectives," *Trends Plant Sci.*, vol. 22, no. 11, pp. 961–975, 2017.

[10] J. Crossa, D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, C. Saint-Pierre, P. Vikram, C. Sansaloni, C. Petroli, D. Akdemir, C. Sneller, M. Reynolds, M. Tattaris, T. Payne, C. Guzman, R. J. Peña, P. Wenzl, and S. Singh, "Genomic prediction of gene bank wheat landraces," *G3, Genes, Genomes, Genet.*, vol. 6, no. 7, pp. 1819–1834, Jul. 2016.

[11] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, pp. 831–838, Jul. 2015.

[12] L. L. Bhering, V. S. Junqueira, L. A. Peixoto, C. D. Cruz, and B. G. Laviola, "Comparison of methods used to identify superior individuals in genomic selection in plant breeding," *Genet. Mol. Res.*, vol. 14, no. 3, pp. 10888–10896, 2015.

[13] P. M. VanRaden, "Efficient methods to compute genomic predictions," *J. Dairy Sci.*, vol. 91, no. 11, pp. 4414–4423, Nov. 2008.

[14] H. Gao, O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. Su, "Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population," *Genet. Selection Evol.*, vol. 44, no. 1, pp. 1–8, Dec. 2012.

[15] J. B. Endelman, "Ridge regression and other kernels for genomic selection with R package rrBLUP," *Plant Genome*, vol. 4, no. 3, pp. 250–255, Nov. 2011.

[16] J. C. Whittaker, R. Thompson, and M. C. Denham, "Marker-assisted selection using ridge regression," *Ann. Hum. Genet.*, vol. 63, no. 4, p. 366, 1999.

[17] I. Aguilar, I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor, "Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score," *J. Dairy Sci.*, vol. 93, no. 2, pp. 743–752, Feb. 2010.

[18] O. F. Christensen and M. S. Lund, "Genomic prediction when some animals are not genotyped," *Genet. Selection Evol.*, vol. 42, no. 1, pp. 1–8, Dec. 2010.

[19] D. Habier, R. L. Fernando, K. Kizilkaya, and D. J. Garrick, "Extension of the Bayesian alphabet for genomic selection," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–12, Dec. 2011.

[20] N. Yi and S. Xu, "Bayesian LASSO for quantitative trait loci mapping," *Genetics*, vol. 179, no. 2, pp. 1045–1055, Jun. 2008.

[21] G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher, "Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model," *PLOS Genet.*, vol. 11, no. 4, Apr. 2015, Art. no. e1004969.

[22] M. P. L. Calus, A. C. Bouwman, C. Schrooten, and R. F. Veerkamp, "Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection," *Genet. Selection Evol.*, vol. 48, no. 1, pp. 1–19, Dec. 2016.

[23] L. Koumakis, "Deep learning models in genomics; are we there yet?" *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1466–1473, Jan. 2020.

[24] V. H. Gazestani and N. E. Lewis, "From genotype to phenotype: Augmenting deep learning with networks and systems biology," *Current Opinion Syst. Biol.*, vol. 15, pp. 68–73, Jun. 2019.

[25] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: New computational modelling techniques for genomics," *Nature Rev. Genet.*, vol. 20, no. 7, pp. 389–403, Jul. 2019.

[26] T. Yue and H. Wang, "Deep learning for genomics: A concise overview," 2018, *arXiv:1802.00810*.

[27] M. E. Ekpenyong, P. I. Etebong, and T. C. Jackson, "Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy," *Heliyon*, vol. 5, no. 7, 2019, Art. no. e02080.

[28] J. You, R. D. McLeod, and P. Hu, "Predicting drug-target interaction network using deep learning model," *Comput. Biol. Chem.*, vol. 80, pp. 90–101, Jun. 2019.

[29] G. S. Araújo, M. R. B. Souza, J. R. M. Oliveira, and I. G. Costa, "Random forest and gene networks for association of SNPs to Alzheimer's disease," in in *Proc. Brazilian Symp. Bioinf.*, Cham, Switzerland: Springer, 2013, pp. 104–115.

[30] X. Chen, C.-C. Zhu, and J. Yin, "Ensemble of decision tree reveals potential miRNA-disease associations," *PLOS Comput. Biol.*, vol. 15, no. 7, Jul. 2019, Art. no. e1007209.

[31] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1301–1310, Jan. 2020.

[32] J. Crawford and C. S. Greene, "Incorporating biological structure into machine learning models in biomedicine," *Current Opinion Biotechnol.*, vol. 63, pp. 126–134, Jun. 2020.

[33] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.

[34] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019.

[35] H. Wang, E. Cimen, N. Singh, and E. Buckler, "Deep learning for plant genomics and crop improvement," *Current Opinion Plant Biol.*, vol. 54, pp. 34–41, Apr. 2020.

[36] W. Ma, Z. Qiu, J. Song, J. Li, Q. Cheng, J. Zhai, and C. Ma, "A deep convolutional neural network approach for predicting phenotypes from genotypes," *Planta*, vol. 248, no. 5, pp. 1307–1318, Nov. 2018.

[37] Z. Qiu, Q. Cheng, J. Song, Y. Tang, and C. Ma, "Application of machine learning-based classification to genomic selection and performance improvement," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2016, pp. 412–421.

[38] J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proc.*, vol. 5, no. S3, pp. 1–5, Dec. 2011.

[39] O. A. Montesinos-López, A. Montesinos-López, R. Tuberosa, M. Maccaferri, G. Sciara, K. Ammar, and J. Crossa, "Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods," *Frontiers Plant Sci.*, vol. 10, p. 1311, Nov. 2019.

[40] O. A. Montesinos-López, A. Montesinos-López, P. Pérez-Rodríguez, J. A. Barrón-López, J. W. R. Martini, S. B. Fajardo-Flores, L. S. Gaytan-Lugo, P. C. Santana-Mancilla, and J. Crossa, "A review of deep learning applications for genomic selection," *BMC Genomics*, vol. 22, no. 1, pp. 1–23, Dec. 2021.

[41] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[42] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme gradient boosting machine for miRNA-disease association prediction," *Cell Death Disease*, vol. 9, no. 1, pp. 1–16, Jan. 2018.

[43] V. K. Ayyadevara, "Gradient boosting machine," in *Pro Machine Learning Algorithms*. Berkeley, CA, USA: Apress, 2018, pp. 117–134.

[44] J. H. Friedman and J. J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statist. Med.*, vol. 22, no. 9, pp. 1365–1381, May 2003.

[45] Z. Liu, M. Zhang, F. Liu and B. Zhang, "Multidimensional feature fusion and ensemble learning-based fault diagnosis for the braking system of heavy-haul train," in *IEEE Trans. Ind. Inform.*, vol. 17, no. 1, pp. 41–51, Jan. 2021, doi: 10.1109/TII.2020.2979467.

[46] B. Li, "Classification and regression trees (CART)," *Biometrics*, vol. 40, no. 3, pp. 358–361, 1984.

[47] M. Blondel, A. Onogi, H. Iwata, and N. Ueda, "A ranking approach to genomic selection," *PLoS ONE*, vol. 10, no. 6, Jun. 2015, Art. no. e0128570.

[48] *Prediction Accuracy of Various Genomic Selection Models on Yield and Quality Traits in Chinese Winter Wheat*, Chinese Academy of Agricultural Sciences, Beijing, China, 2020.

[49] L. Yin, "Development of a machine learning based method to improve the genomic prediction accuracy and computation efficiency for complex traits," Huazhong Agricult. Univ., Wuhan, China, Tech. Rep., 2020, doi: 10.27158/d.cnki.ghznu.2020.000851.

[50] N. R. Wray, C. Wijmenga, P. F. Sullivan, J. Yang, and P. M. Visscher, "Common disease is more complex than implied by the core gene omnigenic model," *Cell*, vol. 173, no. 7, pp. 1573–1580, Jun. 2018.

[51] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *Proc. IEEE 6th Int. Conf. Adv. Comput. (IACC)*, Feb. 2016, pp. 1–6.

[52] X. Peng, L. Yin, Q. Mei, H. Wang, X. Liu, M. Zhu, X. Li, L. Fu, and S. Zhao, "A study of genome selection based on the porcine major economic traits," *Acta Veterinaria Zootechnica Sinica*, vol. 50, no. 2, pp. 439–445, 2019.

[53] W. Zhang, W. E. I. Qun, and W. U. Tianaoet, "Prediction models of reference crop evapotranspiration based on gradient boosting decision tree (GBDT) algorithm in Jiangsu province," *Jiangsu J. Agricult. Sci.*, vol. 36, no. 5, pp. 103–114, 2020.

[54] R. D. S. Rosado, C. D. Cruz, L. D. Barili, J. E. de Souza Carneiro, P. C. S. Carneiro, V. Q. Carneiro, J. T. da Silva, and M. Nascimento, "Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars," *Agriculture*, vol. 10, no. 12, p. 638, Dec. 2020.

[55] M. Shin, M. Ithnin, W. T. Vu, K. Kamaruddin, T. N. Chin, Z. Yaakub, P. L. Chang, K. Sritharan, S. Nuzhdin, and R. Singh, "Association mapping analysis of oil palm interspecific hybrid populations and predicting phenotypic values via machine learning algorithms," *Plant Breeding*, vol. 140, no. 6, pp. 1150–1165, Dec. 2021.

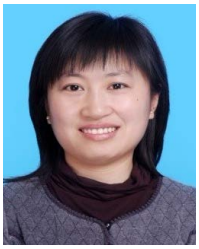
[56] J. Yan, Y. Xu, Q. Cheng, S. Jiang, Q. Wang, Y. Xiao, C. Ma, J. Yan, and X. Wang, "LightGBM: Accelerated genomically designed crop breeding through ensemble learning," *Genome Biol.*, vol. 22, no. 1, pp. 1–24, Dec. 2021.

[57] Y. Xu, J. D. Laurie, and X. Wang, "CropGBM: An ultra-efficient machine learning toolbox for genomic selection-assisted breeding in crops," in *Accelerated Breeding of Cereal Crops*. New York, NY, USA: Humana, 2022, pp. 133–150.

TINGXI YU was born in October 1996. He is currently pursuing the master's degree with the School of Software, Shanxi Agricultural University. His research interests include machine learning, genome selection, data mining, and smart agriculture.

LI WANG was born in April 1986. She is currently pursuing the master's degree with a focus on machine learning and plant phenotype-related research.

WUPING ZHANG was born in April 1973. He is currently pursuing the Ph.D. degree. He is also the Head of the Department of Artificial Intelligence and Smart Agriculture, School of Software, Shanxi Agricultural University. He is also a Professor and a Master Tutor. His research interests include plant phenomics, smart agriculture, soil-plant system model and its application, and 3S technology and its application. He is a China Computer Association and a member of Agricultural Modeling and Simulation Committee. He has presided over (participated) the completion of eight national, provincial, and ministerial projects, four software copyrights, three national invention patents, published more than 40 representative papers, and won one-second prize of the Shanxi Science and Technology Progress Award, the editor-in-chief one textbook compiled by the ten-three-five countries.



GUOFANG XING received the Doctorate degree in agronomy from China Agricultural University. She is currently an Associate Professor and a Master Tutor of the Agricultural College, Shanxi Agricultural University, the Deputy Director of the Genetics and Breeding Department, the Director of the Genetics Teaching and Research Section, the Director of the Shanxi Coarse Cereals Association, and the Deputy of the Association of Overseas Chinese Residents of Shanxi Agricultural University Secretary-General, mainly engaged in the teaching work of "Genetics," "Crop Breeding," "Epigenetics," "Transgenic Technology," "Biological Evolution," engaged in research work on the innovation and utilization of crop germplasm resources.

JIWAN HAN was born in February 1976. He received the Ph.D. degree from the University of Hertfordshire, U.K. He was a Senior Researcher and a Doctoral Supervisor at Abel University, U.K. In 2019, he was introduced to the Software College, Shanxi Agricultural University and worked as the Deputy Director of the Plant Phenotypic Research Center, Shanxi Agricultural University. He host or participate in seven projects, such as "A China-U.K. joint phenomics consortium to dissect the basis of crop stress resistance in the face of dimate change." He was responsible for the collection and pre-processing of all data of the British National Plant Phenotyping Center, and independently presided over a number of plant phenotypic research projects of the center. He has published 23 representative papers, mainly engaged in system research and development in the field of plant phenotyping.

FUZHONG LI has been working with the Student Affairs Department, Shanxi Agricultural University, the College of Animal Science and Technology, the College of Economics and Trade, and other units, since July 1992. Since 2011, he has been working as the Dean, a Professor, and a Doctoral Supervisor of the School of Software, and take over as Information Science and Engineering, in 2020, the Dean of the College. His research interests include smart agriculture, crop phenotype, spectrum detection analysis, and the application of blockchain technology in agriculture.

CHUNQING CAO was born in April 1996. He is currently pursuing the master's degree with the School of Software, Shanxi Agricultural University. His research interest includes machine learning.