frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 14 June 2024
DOI 10.3389/fgene.2024.1415249

Check for updates

OPEN ACCESS

EDITED BY
Ruzong Fan,
Georgetown University Medical Center,
United States

REVIEWED BY
Yutong Luo,
Georgetown University, United States
Shibo Wang,
University of California, Riverside, United States

*CORRESPONDENCE
Jing Gao,
gaojing@imau.edu.cn

RECEIVED 10 April 2024
ACCEPTED 29 May 2024
PUBLISHED 14 June 2024

CITATION
Li R, Gao J, Zhou G, Zuo D and Sun Y (2024), SABO-ILSTSVR: a genomic prediction method based on improved least squares twin support vector regression.
Front. Genet. 15:1415249.
doi: 10.3389/fgene.2024.1415249

SABO-ILSTSVR: a genomic prediction method based on improved least squares twin support vector regression

Rui Li[1,2], Jing Gao[1,2,3]*, Ganghui Zhou[1,2], Dongshi Zuo[1,2] and Yao Sun[1,2]

[1]College of Computer and Information Engineering, Inner Mongolia Agricultual University, Hohhot, China, [2]Inner Mongolia Autonomous Region Key Laboratory of Big Data Research and Application for Agriculture and Animal Husbandry, Hohhot, China, [3]Inner Mongolia Autonomous Region Big Data Center, Hohhot, China

In modern breeding practices, genomic prediction (GP) uses high-density single nucleotide polymorphisms (SNPs) markers to predict genomic estimated breeding values (GEBVs) for crucial phenotypes, thereby speeding up selection breeding process and shortening generation intervals. However, due to the characteristic of genotype data typically having far fewer sample numbers than SNPs markers, overfitting commonly arise during model training. To address this, the present study builds upon the Least Squares Twin Support Vector Regression (LSTSVR) model by incorporating a Lasso regularization term named ILSTSVR. Because of the complexity of parameter tuning for different datasets, subtraction average based optimizer (SABO) is further introduced to optimize ILSTSVR, and then obtain the GP model named SABO-ILSTSVR. Experiments conducted on four different crop datasets demonstrate that SABO-ILSTSVR outperforms or is equivalent in efficiency to widely-used genomic prediction methods. Source codes and data are available at: https://github.com/MLBreeding/SABO-ILSTSVR.

KEYWORDS

genomic prediction, LSTSVR, LASSO regularization, subtraction average based optimizer, high-dimensional data

1 Introduction

With the decreasing cost of high-throughput sequencing data, genomic prediction (GP) emerges as a novel breeding approach, using high-density single nucleotide polymorphisms (SNPs) to capture associations between markers and phenotypes, thereby enabling prediction of genomic estimated breeding values (GEBVs) at an early stage of breeding (Meuwissen et al., 2001). Compared with conventional breeding methods, such as phenotype and marker-assisted selection, GP greatly shortens generation intervals, reduces costs, and enhances the efficiency and accuracy of new variety selection (Heffner et al., 2010).

From the proposal of the concept of genomic prediction to the present, a multitude of models have emerged. Early models primarily focused on improving best linear unbiased prediction (BLUP), such as ridge regression-based best linear unbiased prediction (rrBLUP) (Henderson, 1975) and genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008), etc. In addition, researchers have proposed various Bayesian methods, including

BayesA and BayesB (Meuwissen et al., 2001), BayesC (Habier et al., 2011) and BayesLasso (Park and Casella, 2008), Bayesian ridge regression (BayesRR) (da Silva et al., 2021), BSLMM(Zhou et al., 2013). Moreover, bayesian methods generally exhibit higher prediction accuracy than GBLUP in the majority of cases (Rolf et al., 2015). However, the Markov Chain Monte Carlo (MCMC) steps involved in parameter estimation for Bayesian methods can significantly increase computational costs. With advancements in high-throughput sequencing technologies, the increasing dimensionality of genotype data poses new challenges for GP models. To address this problem, some researchers have begun employing regularization term to mitigate the overfitting problem, such as ridge regression (Ogutu et al., 2012), Lasso (Usai et al., 2009), elastic net (Wang et al., 2019). Meanwhile, machine learning (ML) methods such as support vector regression (SVR) (Maenhout et al., 2007; Ogutu et al., 2011), random forest (RF) (Svetnik et al., 2003), gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) and light gradient boosting machine (LightGBM) (Ke et al., 2017), have made great performance in genomic prediction methods. With the development of deep learning (DL), researchers have also combined it with genomic prediction models, such as DeepGS proposed by Ma et al. (Ma et al., 2018), based on convolutional neural networks (CNN), and DNNGP proposed by Wang et al. (Kelin et al., 2023) for application in multi-omics, which have achieved better performance compared with other classic models. However, genotype data for most species exhibit high-dimensional small-sample characteristics, leading models often to fail to learn effective features from the training data. Moreover, most GP models contain a large number of parameters and there are significant differences in genotype data among different species, leading to a tedious parameter tuning process for each species, significantly increasing breeding costs. Therefore, enhancing the prediction performance of GP models and reducing the complexity of parameters tuning is of crucial importance for shortening generation intervals and reducing breeding costs.

This study explores a machine learning model, least squares twin support vector regression (LSTSVR), to address above problem. LSTSVR, proposed by Zhong et al. (2012), is a regression model that integrates the ideas of least squares method and twin support vector machine (TSVM) (Jayadeva et al., 2007). LSTSVR contains a kernel function; when linearly inseparable data exists in the original input space, it can become linearly separable after being mapped into a higher-dimensional feature space through an appropriate kernel function (Kung, 2014). LSTSVR improves the computational efficiency during the training process of traditional SVR models by introducing the least squares paradigm to replace the ε-insensitive loss function in SVR, thereby transforming the originally nonlinear optimization problem into an easier-to-solve system of linear equations, offering more stable performance. Simultaneously, adopting a two sets of support vectors for regression enhances the model's learning capability and robustness (Huang et al., 2013). However, as a general-purpose regression model, LSTSVR still has certain limitations when dealing with various high-dimensional, small-sample genotype datasets.

To better address model overfitting and complex parameter tuning when the number of genotype samples is far less than the number of SNPs markers (Crossa et al., 2017; Tong and Nikoloski, 2020) this study has made improvements and optimizations to LSTSVR. Firstly, inspired by Lasso regularization, a penalty term was introduced to constrain model complexity on LSTSVR. Concurrently, in order to reduce the complexity stemming from model parameter tuning, the subtraction average-based optimizer (SABO) (Trojovský and Dehghani, 2023) was adopted to perform parameter optimization on the LSTSVR model. By combining the Lasso regularization-based ILSTSVR with the efficient optimization of SABO, this study successfully developed a genomic prediction model named SABO-ILSTSVR. To validate the effectiveness of SABO-ILSTSVR, comparative experiments were conducted using SABO-ILSTSVR on four different species datasets (maize, potato, wheat, and brassica napus) against commonly used genomic prediction models (LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP). The results demonstrate that SABO-ILSTSVR exhibits equivalent or superior performance compared with widely-used genomic prediction methods. Finally, in order to reduce the difficulty of using the model, this study provides an easy-to-use python-based tool for breeders to use conveniently.

2 Materials and methods

2.1 Dataset

Four different crops datasets are used in this study, including potatoes, wheat, maize, and Brassica napus. The following provides detailed descriptions of the genotype and phenotype data for each dataset.

Potato dataset (Selga et al., 2021) is derived from a total of 256 cultivated varieties across three locations in northern and southern Sweden. Over 2000 SNPs markers used for genome-wide prediction were obtained from germplasm resources at both the Centro Internacional de la Papa (CIP, Lima, Peru) and those in the United States. According to Selga et al. (2021), this number of SNPs is already sufficient for predict genomic estimated breeding values (GEBVs) without loss of information. In this dataset, the total weight of tubers serves as the phenotype data.

Wheat dataset (Crossa et al., 2014) originates from the Global Wheat Program at CIMMYT, comprising information on 599 wheat lines. The project carried out numerous experiments across various environmental settings, with the dataset divided into four core environments according to distinct environmental parameters. Average grain yield (GY) serves as the phenotypic trait data within this dataset. It contains 1,279 SNPs markers, which were acquired following the removal of those with minor allele frequencies below 0.05 and the estimation of missing genotypes utilizing samples from the genotype edge distribution.

Maize dataset (Crossa et al., 2014) originates from CIMMYT's maize project, comprising 242 maize lines and 46,374 SNP markers. The project encompasses multiple phenotypic data points, and we use the most significant yield-related traits as our phenotype data for this study.

Brassica napus dataset (Kole et al., 2002) is part of the MTGS package. The dataset comprises 50 lines derived from two varieties, 100 SNP markers, and phenotype information on flowering days across three distinct time periods (flower0, flower4, flower8).

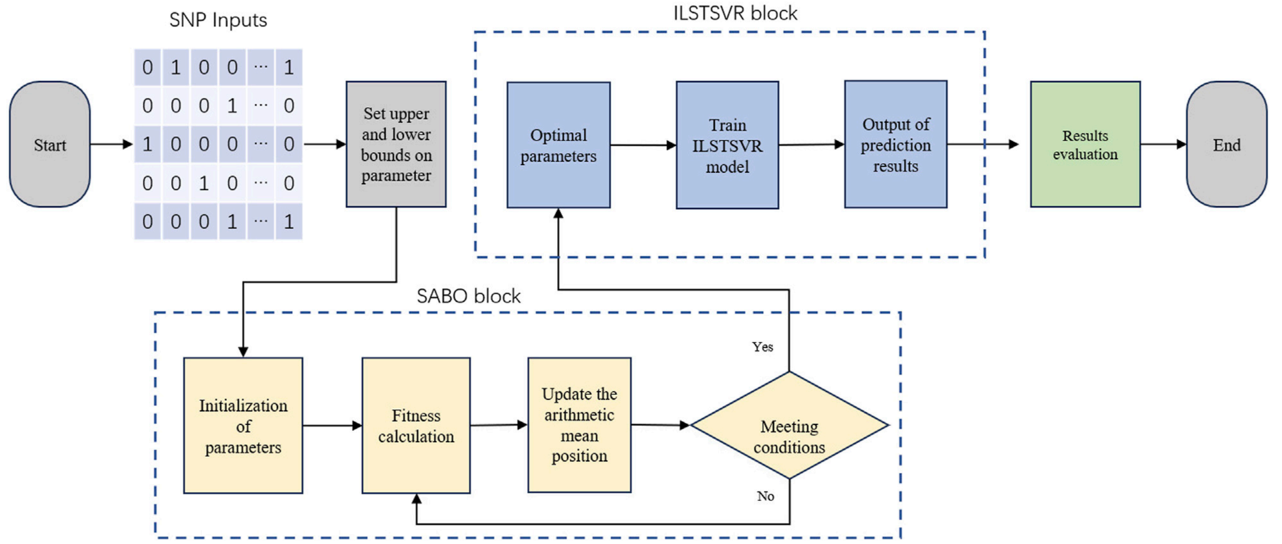

FIGURE 1
Illustration of SABO-ILSTSVR model frameworks.

2.2 SABO-ILSTSVR model

The SABO-ILSTSVR model integrates the subtraction-based average optimizer (SABO) and an improved LSTSVR method. Its overall framework is depicted in Figure 1, followed by an elaboration of the model.

2.2.1 LSTSVR

Least squares twin support vector regression (LSTSVR) is an improved twin support vector regression (TSVR) (Peng, 2010; Shao et al., 2013) model by Yang et al. (Lu Zhenxing, 2014). TSVR is an extension of the regression model derived from twin support vector machine (TSVM) (Jayadeva et al., 2007), suitable for addressing continuous value prediction problems, which accomplishes this by determining two regression functions, shown as follows,

$$f_1(X) = w_1^T X + b_1, f_2(X) = w_2^T X + b_2 \tag{1}$$

where Eq. (1) determines ε-insensitive down-bound function f_1 and ε-insensitive up-bound function f_2, w_1, w_2 are weight vectors, b_1, b_2 are bias term. These can be obtained by solving the following quadratic programming problems (QPPs),

$$\begin{aligned} &\min_{w_1,b_1} \frac{1}{2}\|Y - e\varepsilon_1 - (Xw_1 + eb_1)\|_2^2 + C_1 e^T \xi_1 \\ &s.t. \quad Y - (Xw_1 + eb_1) \geq -e\varepsilon_1 - \xi_1, \xi_1 \geq 0, \end{aligned} \tag{2}$$

$$\begin{aligned} &\min_{w_2,b_2} \frac{1}{2}\|Y + e\varepsilon_2 - (Xw_2 + eb_2)\|_2^2 + C_2 e^T \xi_2 \\ &s.t. \quad -Y + (Xw_2 + eb_2) \geq -e\varepsilon_2 - \xi_2, \xi_2 \geq 0, \end{aligned} \tag{3}$$

where C_1, C_2 are the positive penalty parameters, ε_1, ε_2 are up- and down-bound parameters, ξ_1, ξ_2 are slack vectors, e is a vector of ones with appropriate dimensions. The result of the final regression function is decided by the mean of upper and lower bound functions, as Eq. (4),

$$f(X) = \frac{1}{2}(f_1(X) + f_2(X)) = \frac{1}{2}(w_1^T + w_2^T)X + \frac{1}{2}(b_l + b_2) \tag{4}$$

In the spirit of LSTSVM, Yang et al. apply least squares method to TSVR. TSVR finds the optimal weight vector and bias terms by solving two QPPs, whereas LSTSVR transforms the original TSVR problem into two systems of linear equations for solution, which is typically faster and more stable than directly solving q QPPs, with the loss in accuracy being within an acceptable range. For LSTSVR model, the inequality constraints of (2) and (3) are replaced with equality constraints as follows,

$$\begin{aligned} &\min_{w_1,b_1} \frac{1}{2}\|Y - e\varepsilon_1 - (Xw_1 + eb_1)\|_2^2 + C_1 {\xi_1}^T \xi_1 \\ &s.t. \quad Y - (Xw_1 + eb_1) = -e\varepsilon_1 - \xi_1, \end{aligned} \tag{5}$$

$$\begin{aligned} &\min_{w_2,b_2} \frac{1}{2}\|Y + e\varepsilon_2 - (Xw_2 + eb_2)\|_2^2 + C_2 {\xi_2}^T \xi_2 \\ &s.t. \quad -Y + (Xw_2 + eb_2) = -e\varepsilon_2 - \xi_2, \end{aligned} \tag{6}$$

In formula (5) and (6), the square of L2-norm of slack variable ξ_1, ξ_2 is used, instead of L1-norm in (2) and (3), which makes constraint $\xi_1 \geq 0$, $\xi_2 \geq 0$ redundant, so the following formulas is obtained,

$$\min_{w_1,b_1} \frac{1}{2}\|Y - (Xw_1 + eb_1)\|_2^2 + \frac{C_1}{2}\|(Xw_1 + eb_1) - Y - e\varepsilon_1\|_2^2, \tag{7}$$

$$\min_{w_2,b_2} \frac{1}{2}\|Y - (Xw_2 + eb_2)\|_2^2 + \frac{C_2}{2}\|-(Xw_2 + eb_2) + Y - e\varepsilon_2\|_2^2, \tag{8}$$

(7) and (8) are two unconstrained QPPs, hence the solutions for w and b can be directly obtained by setting the derivatives to zero, as Eq. (9),

$$-\mathbf{H}^{\mathrm{T}}((\mathbf{Y} - \varepsilon_1 \mathbf{e}) - \mathbf{Hu}) + \mathbf{c}_1 \mathbf{H}^{\mathrm{T}}(\mathbf{Hu} - (\mathbf{Y} - \varepsilon_1 \mathbf{e})) = \mathbf{0}, \tag{9}$$

where $u = \begin{bmatrix} w \\ b \end{bmatrix}$, H = [X e], then, we have Eq. (10) and Eq. (11),

$$\mathbf{u_1} = \begin{bmatrix} \mathbf{w_1} \\ \mathbf{b_1} \end{bmatrix} = (\mathbf{1}+\mathbf{C_1})^2 \left(\mathbf{H^T H}\right)^{-1} \mathbf{H^T Y} + (\mathbf{1}+\mathbf{C_1})\mathbf{C_1} \left(\mathbf{H^T H}\right)^{-1} \mathbf{H^T} \varepsilon_1 \mathbf{e}, \tag{10}$$

$$\mathbf{u_2} = \begin{bmatrix} \mathbf{w_2} \\ \mathbf{b_2} \end{bmatrix} = (\mathbf{1}+\mathbf{C_2})^2 \left(\mathbf{H^T H}\right)^{-1} \mathbf{H^T Y} - (\mathbf{1}+\mathbf{C_2})\mathbf{C_2} \left(\mathbf{H^T H}\right)^{-1} \mathbf{H^T} \varepsilon_2 \mathbf{e}, \tag{11}$$

thus, the final regression function is $f(X) = \frac{1}{2} X(w_1^T + w_2^T) + \frac{1}{2}(b_l + b_2)$.

2.2.2 Improved LSTSVR (ILSTSVR)

When applying LSTSVR to handle high-dimensional genotype datasets with small samples, the model is prone to a high risk of overfitting. This is due to the fact that the model may overly fit noise and feature details in the training set, leading to decrease generalization performance on new samples and thereby affecting the effectiveness and reliability of the predictive results. Therefore, this study adds a Lasso regularization term for the weight parameter w in the LSTSVR framework. For linear problems, the function of ILSTSVR is as follows,

$$\begin{gathered} \min_{w_1, b_1} \frac{1}{2}\|Y - \varepsilon_1 e - (Xw_1 + b_1 e)\|_2^2 + \frac{C_1}{2}{\xi_1}^T \xi_1 + \frac{C_3}{2}\left(\|w_1\|_1 + b_1^2\right), \\ s.t. \quad Y - (Xw_1 + b_1 e) = -\varepsilon_1 e - \xi_1 \end{gathered} \tag{12}$$

$$\begin{gathered} \min_{w_2, b_2} \frac{1}{2}\|Y + \varepsilon_2 e - (Xw_2 + b_2 e)\|_2^2 + \frac{C_2}{2}{\xi_2}^T \xi_2 + \frac{C_4}{2}\left(\|w_2\|_1 + b_2^2\right), \\ s.t. \quad -Y + (Xw_2 + b_2 e) = -\varepsilon_2 e - \xi_2 \end{gathered} \tag{13}$$

where $C_1, C_2, C_3, C_4, \varepsilon_1, \varepsilon_2$ are positive penalty parameters, ξ_1, ξ_2 are slack variables, e is a vector of ones of appropriate dimensions. For the non-differentiability with L1 regularization and the convenience of calculations, we assume $\|w\|_1 = \|\alpha^* w\|_2^2$, then (12) and (13) can be converted into as follow,

$$\begin{gathered} \min_{w_1, b_1} \frac{1}{2}\|Y - \varepsilon_1 e - (Xw_1 + b_1 e)\|_2^2 + \frac{C_1}{2}\|(Xw_1 + b_1 e) - Y - \varepsilon_1 e\|_2^2 \\ + \frac{C_3}{2}\left(\|\alpha_1 * w_1\|_2^2 + b_1^2\right), \end{gathered} \tag{14}$$

$$\begin{gathered} \min_{w_2, b_2} \frac{1}{2}\|Y + \varepsilon_2 e - (Xw_2 + b_2 e)\|_2^2 + \frac{C_2}{2}\|-(Xw_2 + b_2 e) + Y - \varepsilon_2 e\|_2^2 \\ + \frac{C_4}{2}\left(\|\alpha_2 * w_2\|_2^2 + b_2^2\right), \end{gathered} \tag{15}$$

where α_1, α_2 are vectors of appropriate dimensions, * represents element-wise multiplication. Expand (14) yields,

$$\begin{aligned} L = \min \frac{1}{2}&(Y - \varepsilon_1 e - (Xw_1 + b_1 e))^T (Y - \varepsilon_1 e - (Xw_1 + b_1 e)) \\ &+ \frac{C_1}{2}((Xw_1 + b_1 e) - Y - \varepsilon_1 e)^T ((Xw_1 + b_1 e) - Y - \varepsilon_1 e) \\ &+ \frac{C_3}{2}(diag(\alpha_1)w_1)^T (diag(\alpha_1)w_1) + \frac{C_3}{2}b_1^2, \end{aligned} \tag{16}$$

let the derivatives of (16) with respect to w_1 and b_1 respectively be zero, we obtain Eq. (17) and Eq. (18),

$$\begin{aligned} \frac{\partial L}{\partial w_1} = &- X^T (Y - Xw_1 - b_1 e - \varepsilon_1 e) + C_1 X^T (Xw_1 + b_1 e - \varepsilon_1 e - Y) \\ &+ C_3 diag(\alpha_1)^T diag(\alpha_1) w_1 = \mathbf{0}, \end{aligned} \tag{17}$$

$$\begin{aligned} \frac{\partial L}{\partial b_1} = &- e^T (Y - Xw_1 - b_1 e - \varepsilon_1 e) + C_1 e^T (Xw_1 + b_1 e - \varepsilon_1 e - Y) \\ &+ C_3 b_1 = \mathbf{0}, \end{aligned} \tag{18}$$

composed in matrix form as Eq. (19),

$$\begin{aligned} &\begin{bmatrix} (1 + C_1)X^T X + C_3 D_1 & (1 + C_1)X^T e \\ (1 + C_1)e^T X & (1 + C_1)e^T e + C_3 \end{bmatrix} \begin{bmatrix} w_1 \\ b_1 \end{bmatrix} \\ &= \begin{bmatrix} (1 + C_1)X^T & (C_1 - 1)X^T e \\ (1 + C_1)e^T & (C_1 - 1)e^T e \end{bmatrix} \begin{bmatrix} Y \\ \varepsilon_1 \end{bmatrix} \end{aligned} \tag{19}$$

where $D_1 = diag(\alpha_1)^T diag(\alpha_1)$. Similarly, the following result Eq. (20) can be obtained through (15),

$$\begin{aligned} &\begin{bmatrix} (1 + C_2)X^T X + C_4 D_2 & (1 + C_2)X^T e \\ (1 + C_2)e^T X & (1 + C_2)e^T e + C_4 \end{bmatrix} \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} \\ &= \begin{bmatrix} (1 + C_2)X^T & (1 - C_2)X^T e \\ (1 + C_2)e^T & (1 - C_2)e^T e \end{bmatrix} \begin{bmatrix} Y \\ \varepsilon_2 \end{bmatrix} \end{aligned} \tag{20}$$

where $D_2 = diag(\alpha_2)^T diag(\alpha_2)$. Because of the assumption $\|w\|_1 = \|\alpha * w\|_2^2$, we set an initial α^0, and calculate the final w and b through the iterative formula that updates alternately, as Eq. (21)

$$\alpha_i^{k+1} = \sqrt{\frac{1}{|w_i^k|}} \tag{21}$$

To make the process clear, the computational process is summarized as Table 1.

Then we obtain Eq. (22),

$$f(X) = \frac{1}{2} X\left(w_1^T + w_2^T\right) + \frac{1}{2}(b_l + b_2) \tag{22}$$

For nonlinear problems, kernel functions typically provide a good solution, and here we have chosen RBF as the kernel function for our model. The formula of RBF as Eq. (23),

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{23}$$

where σ is kernel parameter, the value has a significant impact on prediction performance, easily leading to overfitting or underfitting.

We map the training data through $K(X, X^T)$ into a high-dimensional reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950), obtaining matrix H. Thus, we obtain the function for the nonlinear problem as Eq. (24) and Eq. (25),

$$\begin{gathered} \min_{w_1, b_1} \frac{1}{2}\|Y - \varepsilon_1 e - (Hw_1 + b_1 e)\|_2^2 + \frac{C_1}{2}{\xi_1}^T \xi_1 + \frac{C_3}{2}\left(\|w_1\|_1 + b_1^2\right) \\ s.t. \quad Y - (Xw_1 + b_1 e) = -\varepsilon_1 e - \xi_1, \end{gathered} \tag{24}$$

$$\begin{gathered} \min_{w_2, b_2} \frac{1}{2}\|Y + \varepsilon_2 e - (Hw_2 + b_2 e)\|_2^2 + \frac{C_2}{2}{\xi_2}^T \xi_2 + \frac{C_4}{2}\left(\|w_2\|_1 + b_2^2\right) \\ s.t. \quad -Y + (Xw_2 + b_2 e) = -\varepsilon_2 e - \xi_2, \end{gathered} \tag{25}$$

TABLE 1 Iterative algorithm to solve the L1 regularization problem.

```
Algorithm 1: An iterative algorithm to solve the L1 regularization problem in (14) and (15)
Input: $\varepsilon_1 > 0, \varepsilon_2 > 0, C_1 > 0, C_2 > 0, C_3 > 0, C_4 > 0, X \in R^{n\times m}, Y \in R^{n\times 1}$
Output: $w_1, b_1$ and $w_2, b_2$
1: Random initialization $\alpha_1^0, \alpha_2^0$, set the number of iterations t = 0
2: repeat
    Calculate and update $w_1, b_1$ using (19)
    Calculate and update $w_2, b_2$ using (20)
    Accordingto the formula $\alpha_i{}^{k+1} = \sqrt{\frac{1}{|w_i{}^k|}}$,item-by-item update of $\alpha_1, \alpha_2$ $t = t + 1$
until $\alpha_1, \alpha_2$ convergence
```

Similarly, we can obtain Eq. (26) and Eq. (27),

$$\begin{bmatrix} (1+C_1)H^TH + C_3D_1 & (1+C_1)H^Te \\ (1+C_1)e^TH & (1+C_1)e^Te + C_3 \end{bmatrix} \begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} (1+C_1)H^T & (C_1-1)H^Te \\ (1+C_1)e^T & (C_1-1)e^Te \end{bmatrix} \begin{bmatrix} Y \\ \varepsilon_1 \end{bmatrix}, \tag{26}$$

$$\begin{bmatrix} (1+C_2)H^TH + C_4D_2 & (1+C_2)H^Te \\ (1+C_2)e^TH & (1+C_2)e^Te + C_4 \end{bmatrix} \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = \begin{bmatrix} (1+C_2)H^T & (1-C_2)H^Te \\ (1+C_2)e^T & (1-C_2)e^Te \end{bmatrix} \begin{bmatrix} Y \\ \varepsilon_2 \end{bmatrix}, \tag{27}$$

the final nonlinear model as Eq. (28),

$$f(H) = \frac{1}{2}H\left(w_1^T + w_2^T\right) + \frac{1}{2}(b_l + b_2) \tag{28}$$

2.2.3 Optimizing ILSTSVR using SABO

For the non-linear ILSTSVR model, the choice of kernel parameter σ for RBF has a significant impact on prediction performance. Moreover, in this study, we set $C_1 = C_2$ and $C_3 = C_4$ for the ILSTSVR parameters C_1, C_2, C_3, C_4 as given in (12) and (13). Parameter optimization plays an important role in machine learning, as selecting appropriate parameters can significantly enhance a model's predictive capability and accuracy. Different combinations of parameters may lead to vastly different performances of the model on both training and testing data. Furthermore, parameters affect the model complexity and learning capacity. By adjusting them, we can better strike a balance between overfitting and underfitting (Young et al., 2015). Although grid search can find the global optimal solution, it will result in enormous computational resource consumption and neglect the correlations among parameters (Vincent and Jidesh, 2023). Instead, we use the recently proposed SABO for parameter optimization, which updates the positions of population members in the search space using subtraction averages of individuals, characterized by strong optimization capability and fast convergence rates (Moustafa et al., 2023).

The basic inspiration for the design of the SABO is mathematical concepts such as averages, the differences in the positions of the search agents, and the sign of difference of the two values of the objective function. The idea of using the arithmetic mean location of all the search agents (i.e., the population members of the tth iteration), instead of just using, e.g., the location of the best or worst search agent to update the position of all the search agents (i.e., the construction of all the population members of the $(t + 1)$th iteration), is not new, but the SABO's concept of the computation of the arithmetic mean is wholly unique, as it is based on a special operation " $-_v$", called the v-subtraction of the search agents B from the search agent A, which is defined as Eq. (29):

$$A -_v B = sign(F(A) - F(B))(A - \vec{v} * B), \tag{29}$$

where $\vec{v}$ is a vector of the dimension m, the operation "*" represents the Hadamard product of the two vectors, F(A) and F(B) are the values of the objective function of the search agents A and B, respectively, and sign is the signum function (Trojovský and Dehghani, 2023).

In the proposed SABO, the displacement of any search agent X_i in the search space is calculated by the arithmetic mean of the v-subtraction of each search agent X_j, j = 1,2, . . . , N, from the search agent X_i. Thus, the new position for each search agent is calculated using (30).

$$X_i^{new} = X_i + \vec{r}_i * \frac{1}{N}\sum_{j=1}^{N}(X_i -_v X_j), i = 1, 2, \ldots, N, \tag{30}$$

where X_i^{new} is the new proposed position for the ith search agent X_i, N is the total number of the search agents, and $\vec{r}_i$ is a vector of the dimension m. Then, if this proposed new position leads to an improvement in the value of the objective function, it is acceptable as the new position of the corresponding agent, according to (31)

$$X_i = \begin{cases} X_i^{new}, F_i^{new} < F_i; \\ X_i, else, \end{cases} \tag{31}$$

where F_i and F_i^{new} are the fitness function values of the search agents X_i and X_i^{new}, respectively.

Similar to other optimization algorithms, the primary positions of the search agents in the search space are randomly initialized using (32).

$$x_{i,d} = lb_d + r_{i,d} \cdot (ub_d - lb_d), i = 1, \ldots, N, d = 1, \ldots, m, \tag{32}$$

where $x_{i,d}$ is the dth dimension of X_i, N is the number of search agents, m is the number of decision variables, $r_{i,d}$ is a random number in the interval [0, 1], and lb_d and ub_d are the lower and upper bounds of the dth decision variables, respectively (Trojovský and Dehghani, 2023).

Here, we define the fitness function for SABO as MSE function shown in (34), $C_1 \in [C_{1_{min}}, C_{1_{max}}]$, $C_3 \in [C_{3_{min}}, C_{3_{max}}]$ and $\sigma \in [\sigma_{min}, \sigma_{max}]$. The parameter corresponding to the smallest fitness function value obtained through iteration is the optimal combination of parameters for ILSTSVR. Finally, train the model according to the obtained parameters combination.

2.2.4 Performance evaluation

To evaluate the prediction performance of GEBVs by the model, while avoiding the problem that Pearson correlation coefficient fails to measure the distance between true and predicted values, we adopt both Pearson correlation coefficient and MSE as evaluation metrics for the relationship between predicted and true values. Furthermore,

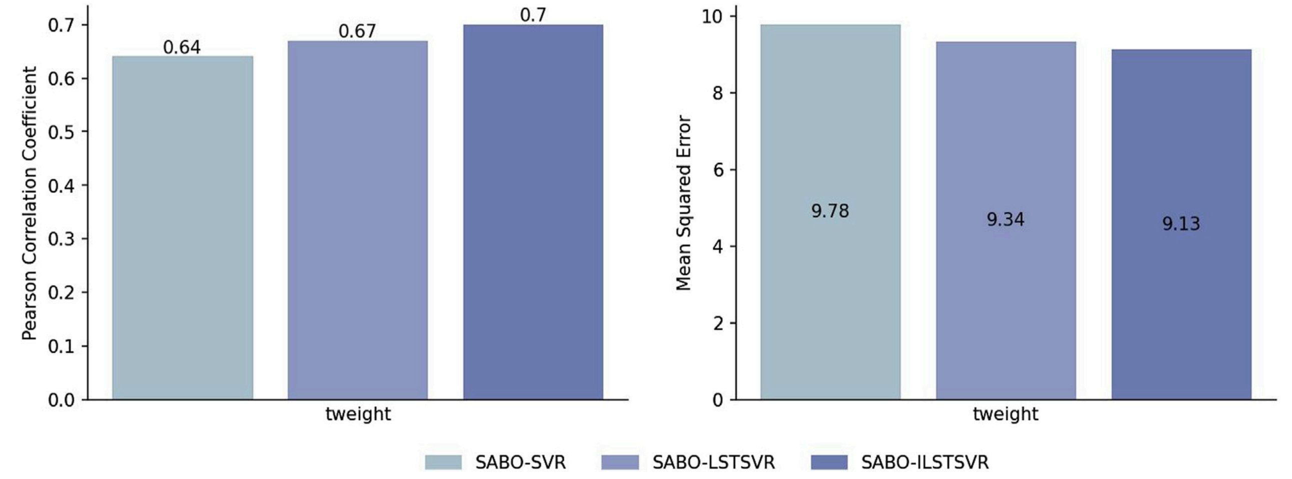

FIGURE 2
Performance of Pearson correlation coefficient and MSE prediction for SABO-SVR, SABO-LSTSVR, and SABO-ILSTSVR on the potato dataset.

we use ten-fold cross-validation to assess the model's performance. The original dataset is divided into ten equally-sized (or nearly equal) folds; for each fold, it serves as the validation set, while the remaining nine folds constitute the training set. Training a model using the training set, and its performance is evaluated using the validation set. After completing all ten iterations, the average of the performance measures obtained from each validation set is taken, thereby yielding an overall assessment of the model's performance.

The Pearson correlation coefficient (PCC) is used to measure the strength and direction of the linear relationship between two continuous variables, and is defined as follows,

$$\rho_{y,y'} = \frac{cov(y, y')}{\sigma_y \sigma_{y'}} \tag{33}$$

where y and y' represent the true values and predicted values respectively, $\operatorname{cov}(y, y')$ denotes the covariance of vectors y and y', σ_y and $\sigma_{y'}$ are the standard deviations of vector y and y' respectively.

Mean squared error (MSE) measures the degree of difference between predicted values and actual values, and is defined as follows,

$$MSE(y, y') = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2 \tag{34}$$

3 Result

3.1 Comparison of SABO-ILSTSVR with the base model

To validate the effectiveness of ILSTSVR, this section compares SABO-ILSTSVR with some methods prior to its improvement on four datasets (potato, wheat, maize and brassica napus). The results on the potato dataset (Figure 2) show that the SABO-ILSTSVR exhibits a 4% increase in Pearson correlation coefficient and a 2% decrease in MSE compared with SABO-LSTSVR. Furthermore, when contrasted with SABO-SVR, the SABO-ILSTSVR demonstrates a 9% improvement in the Pearson correlation coefficient and a 6% decrease in MSE.

The results on other datasets are shown in Table 2. On the wheat dataset, SABO-ILSTSVR improves the Pearson correlation coefficient in four environments by an average of 2%, 2%, 8%, and 2% and reduces the MSE in four environments by an average of 1%, 3%, 1% and 1%, respectively, compared with SABO-SVR and SABO-LSTSVR. On the maize dataset, SABO-ILSTSVR improves the Pearson correlation coefficient by an average of 6% and reduces the MSE by 1%, respectively, compared with SABO-SVR and SABO-LSTSVR. On the brassica napus dataset, SABO-ILSTSVR improves the Pearson correlation coefficient in three traits by an average of 11%, 6%, and 24% and reduces the MSE in three traits by an average of 17%, 1% and 12%, respectively, compared with SABO-SVR and SABO-LSTSVR.

3.2 Comparison of SABO-ILSTSVR with other methods

Considering the complexity of the genetic architecture, this study employs real data to evaluate the prediction performance of models, including datasets from public available sources for maize, wheat, potato, and Brassica napus. And we performed standardization on the phenotype data of all datasets. Due to the small sample sizes in the adopted datasets, random sampling errors may be relatively substantial, leading to decreased model prediction accuracy, reduced statistical power of tests, and difficulty in obtaining stable and reliable statistical inferences (Bengio and Grandvalet, 2004). Consequently, a ten-fold cross-validation is applied to each dataset in this study, with the average of the results over ten iterations used to represent the ultimate prediction performance of the models.

This section compares the prediction performance of LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP, and SABO-ILSTSVR across four datasets (potato, wheat,

TABLE 2 Prediction performance in ten-fold cross-validation for each trait in wheat, maize and brassica napus datasets.

	SABO-SVR		SABO-LSTSVR		SABO-ILSTSVR	
Trait	PCC	MSE	PCC	MSE	PCC	MSE
env1	0.58	0.66	0.58	0.67	0.59	0.65
env2	0.5	0.76	0.49	0.75	0.51	0.73
env3	0.43	0.82	0.42	0.82	0.45	0.81
env4	0.52	0.71	0.5	0.76	0.52	0.73
yield	0.37	0.73	0.38	0.73	0.4	0.72
flower0	0.64	0.045	0.61	0.047	0.7	0.038
flower4	0.68	0.045	0.66	0.04	0.71	0.042
flower8	0.36	0.017	0.38	0.02	0.46	0.016

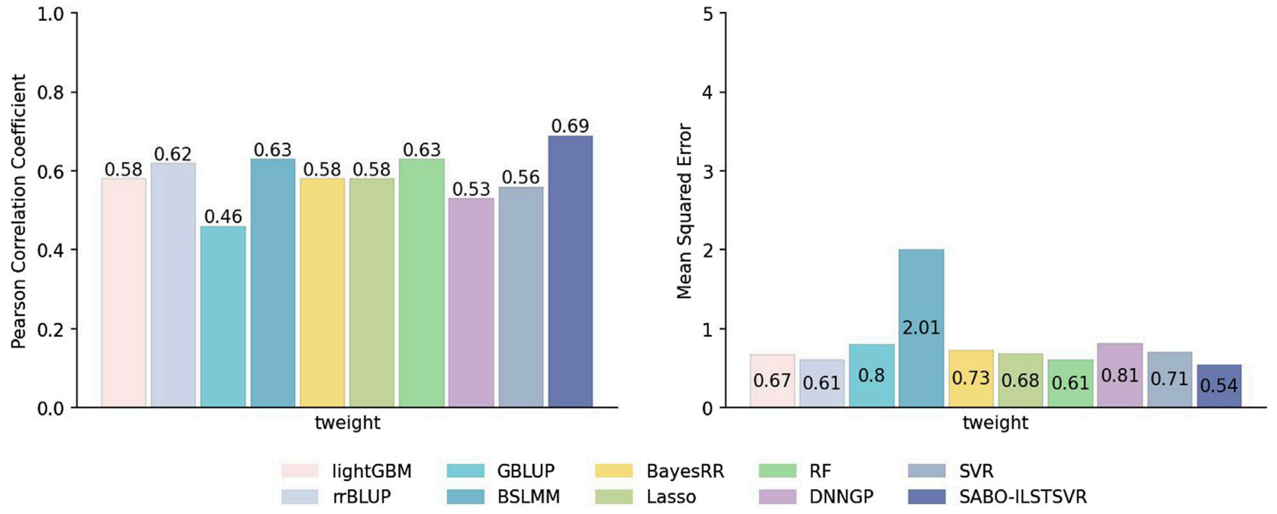

FIGURE 3
Performance of prediction for various models on the potato dataset in terms of Pearson correlation coefficients and MSE.

maize and brassica napus). The LightGBM model originates from a LightGBM Python package developed by Microsoft. The BayesRR, Lasso, RF, and SVR models are included in the Scikit-learn python library. The GBLUP, BSLMM and rrBLUP models utilize the sommer R package (Covarrubias-Pazaran, 2016), hibayes R package (Yin et al., 2022) and the rrBLUP R package (Endelman, 2011), respectively. The DNNGP model is mentioned in the paper by (Kelin et al., 2023).

3.2.1 Potato dataset

This paper first compares the prediction performance of SABO-ILSTSVR with LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP on the potato dataset. Detailed information about SNPs and phenotypes in the potato dataset has been described in the Materials and Methods section. As shown in Figure 3, SABO-ILSTSVR outperforms comparative models across key performance metrics. Specifically, SABO-ILSTSVR improves the Pearson correlation coefficient by 18%, 11%, 50%, 9%, 18%, 18%, 9%, 23% and 30% and reduces the MSE by an average of 19%, 11%, 32%, 73%, 26%, 20%, 11%, 23% and 33%, respectively, compared with LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP.

3.2.2 Wheat dataset

Similarly, prediction performance comparisons were conducted for SABO-ILSTSVR, LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP on the wheat dataset. As shown in Figure 4, the DNNGP model exhibits the highest Pearson correlation coefficients in predicting yield under environments env1 and env2 of the wheat dataset, whereas its performance in env3 and env4 is inferior to that of other models. In contrast, our proposed SABO-ILSTSVR model demonstrates higher Pearson correlation coefficients and lower mean squared errors for yield data across all four environments compared with the other models. Specifically, SABO-ILSTSVR achieves the best performance in env3 and env4 and is second only to DNNGP in env1 and env2. The reason for the differing performance may be due to their prediction performance varying across different agroclimatic regions.

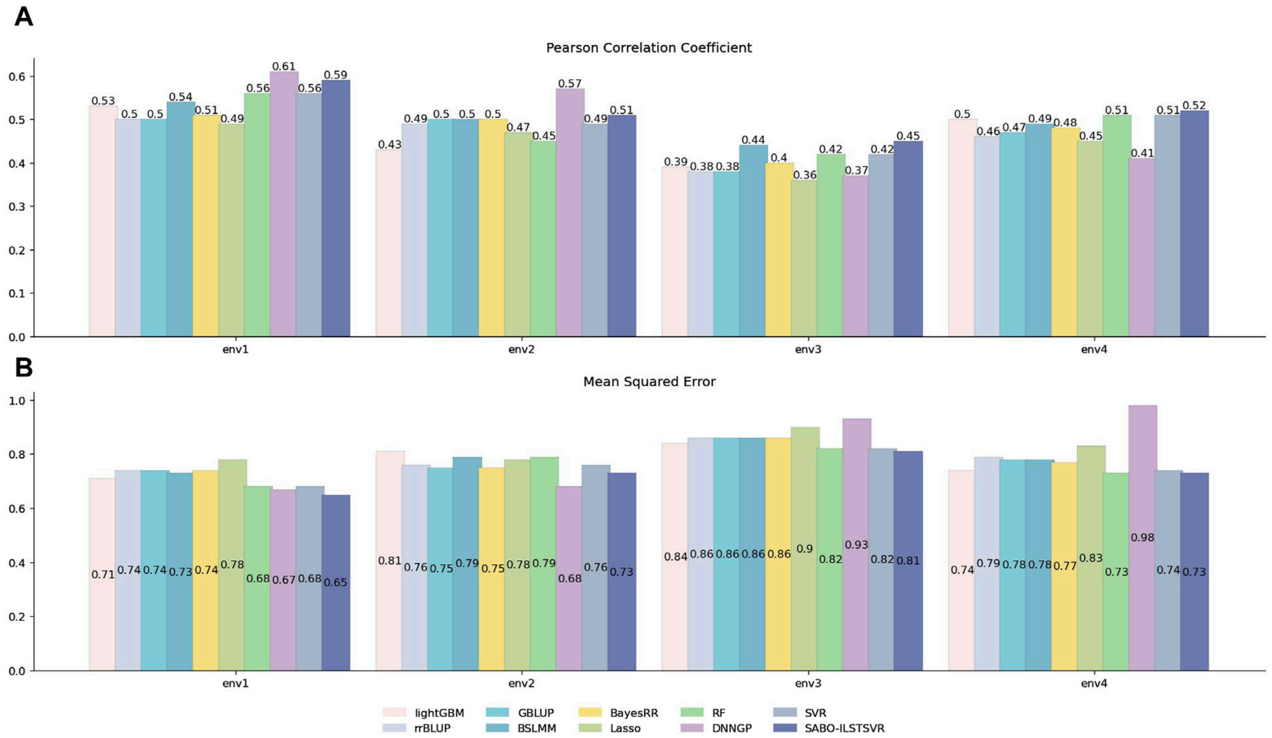

FIGURE 4
Performance of prediction for various models on the wheat dataset. Pearson correlation coefficient **(A)** and MSE **(B)** metrics of four phenotypes (env1, env2, env3 and env4), as evaluated through ten-fold cross-validation.

3.2.3 Maize dataset

The detailed information on SNPs and phenotypes in the maize dataset has been described in the Materials and Methods section. Unlike the potato and wheat datasets, this section performs ten replicates ten-fold cross-validation separately for SABO-ILSTSVR, LightGBM, rrBLUP, GBLUP, BayesRR, Lasso, RF, SVR, DNNGP on the maize dataset. Comparison was not conducted with BSLMM due to its unstable results. The Pearson correlation coefficients of the ten results are represented (A) in Figure 5, while the mean squared errors (MSEs) of the ten results are averaged and depicted as (B). As shown in Figure 5, Lasso exhibits the lowest prediction performance, whereas DNNGP, despite having the highest Pearson correlation coefficient prediction performance, displays significantly large variations across the ten runs, resulting in elongated bars in the boxplot. In contrast, SABO-ILSTSVR has a stable Pearson correlation coefficient prediction performance and exhibits the lowest MSE.

3.2.4 Brassica napus dataset

Similarly, on the brassica napus dataset, ten replicates ten-fold cross-validation was performed for SABO-ILSTSVR, LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP, respectively. The Pearson correlation coefficients of the ten results are represented by (A) in Figure 6, while the mean squared errors (MSEs) after averaging are depicted by (B). As shown in (A) of Figure 6, SABO-ILSTSVR exhibits the highest Pearson correlation coefficient and lowest MSE prediction performance on flower0 and flower8, whereas BSLMM achieves the highest Pearson correlation coefficient prediction performance on flower4. The prediction performance of the SABO-ILSTSVR model is slightly lower than that of the GBLUP and BSLMM model but higher than that of the other comparative models on flower4.

4 Discussion

In this study, we integrate Least Squares Twin Support Vector Regression (LSTSVR) with Lasso regularization, constructing a GP model named ILSTSVR. We use the SABO optimization algorithm to effectively optimize the parameters of the model. To address the number of genotype samples is far less than the number of SNPs markers, we introduce a Lasso regularization term into LSTSVR. By the unique feature selection property of Lasso, it can effectively shrink the coefficients of non-key features to zero, achieving parameter sparsity and effectively preventing overfitting. Meanwhile, to cope with the potential nonlinear relationships in genotype data, we adopt the radial basis function (RBF) kernel, mapping the raw data into a high-dimensional space to attain linear separability.

Considering the differences in genotype data among various species may lead to distinct optimal parameters, we used the SABO optimization algorithm to automatically tune the parameters of the ILSTSVR model. To validate the effectiveness of this model, we conducted evaluations on multiple datasets spanning potato, maize, wheat, and Brassica napus, and compared its prediction performance against a series of widely-used models such as LightGBM, rrBLUP, GBLUP, BSLMM, BayesRR, Lasso, RF, SVR, DNNGP.

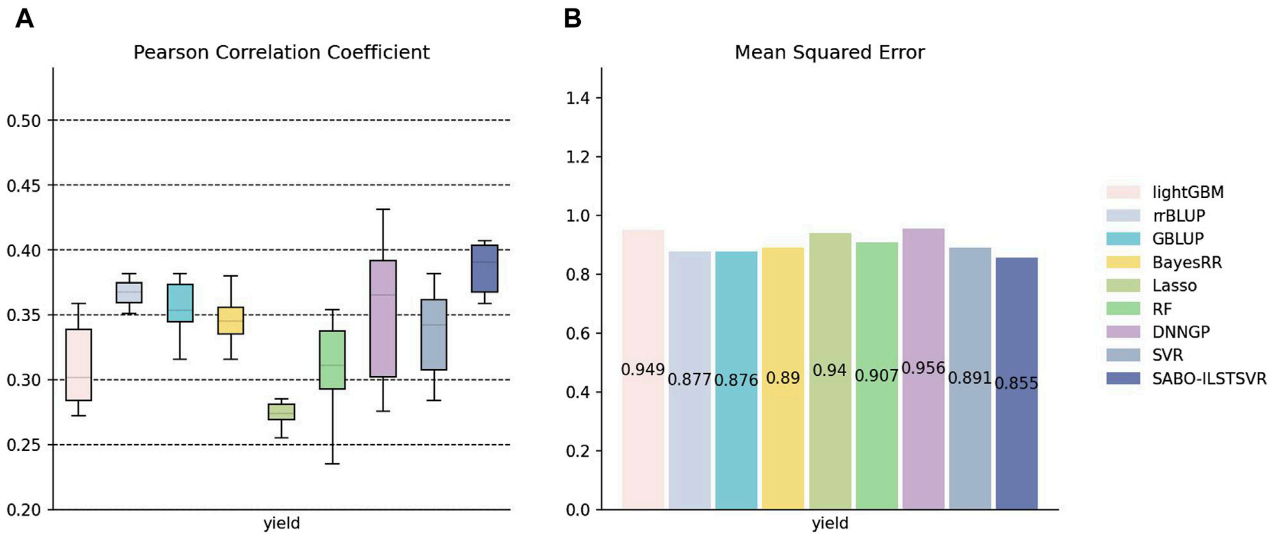

FIGURE 5
Performance of prediction for various models on the maize dataset. **(A)** Pearson correlation coefficients of maize yield traits, represented by box plots, after ten replicates ten-fold cross-validation. **(B)** Average MSE of maize yield traits after ten replicates ten-fold cross-validation.

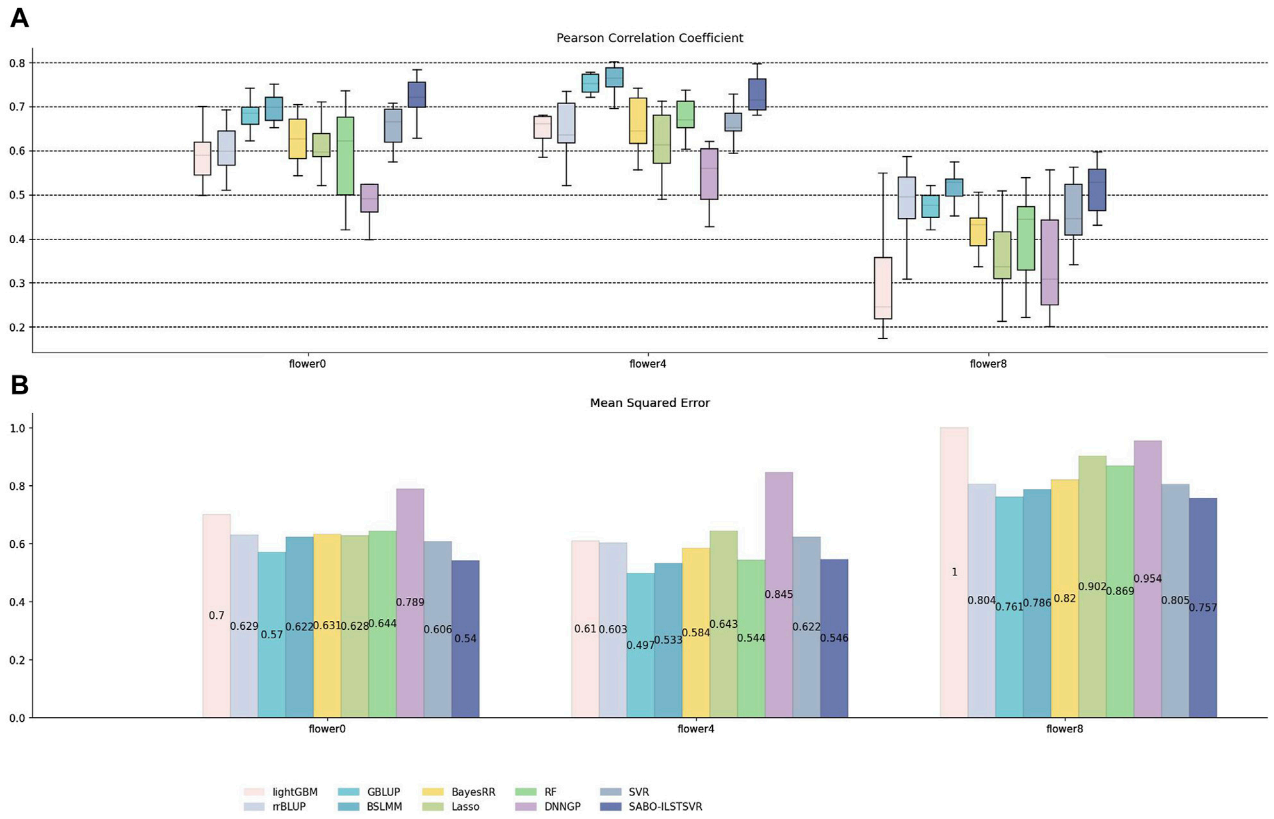

FIGURE 6
Performance of prediction for various models on the brassica napus dataset. **(A)** Pearson correlation coefficients of three phenotypes (flower0, flower4 and flower8), represented by box plots, after ten replicates ten-fold cross-validation. **(B)** Average MSE of three phenotypes after ten replicates ten-fold cross-validation.

The results showed that the SABO-ILSTSVR model demonstrated outstanding prediction performance on the potato dataset, outperforming other benchmark models. In the wheat and brassica napus datasets containing multiple phenotypic traits, our model consistently exhibited higher prediction accuracy for most traits compared with other models. The box plot analysis of the maize and brassica napus dataset further revealed the robustness of the SABO-ILSTSVR model's predictions.

In our further exploration of the future, confronted with the high-dimensional challenges of genomic data, we will delve deeper into how to efficiently perform feature extraction (Burges, 2009). With the continuous decline in sequencing costs, large-scale genomic sequencing of samples is poised to become a reality, the increased sample size may offer more favorable application for DL models (Khan et al., 2020; Yang et al., 2024). We will investigate innovative DL models within the field of breeding.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

RL: Software, Writing–review and editing, Writing–original draft, Methodology. JG: Writing–review and editing. GZ: Writing–review and editing, Data curation. DZ: Writing–review and editing. YS: Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project was funded by the Major Science and Technology Projects of the Inner Mongolia Autonomous Region (2019ZD016, 2021ZD0005) and the 2023 Graduate Research Innovation Project of the Inner Mongolia Autonomous Region (S20231117Z).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* 68, 337–404. doi:10.21236/ada296533

Bengio, Y., and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* 5, 1089–1105.

Burges, C. J. C. (2009). Dimension reduction: a guided tour. *Found. Trends® Mach. Learn.* 2, 275–364. doi:10.1561/9781601983794

Chen, T., and Guestrin, C. X. G. B. (2016) *A scalable tree boosting system*. New York, NY, USA: Association for Computing Machinery, 785–794.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLOS ONE* 11, e0156744. doi:10.1371/journal.pone.0156744

Crossa, J., PéREZ, P., Hickey, J., BurgueñO, J., Ornella, L., CeróN-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi:10.1038/hdy.2013.16

Crossa, J., PéREZ-RodríGUEZ, P., Cuevas, J., Montesinos-LóPEZ, O. A., JarquíN, D., Campos, G. D. L., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends plant Sci.* 22 (11), 961–975. doi:10.1016/j.tplants.2017.08.011

Da Silva, F. A., Viana, A. P., Correa, C. C. G., Santos, E. A., De Oliveira, J. A. V. S., Andrade, J. D. G., et al. (2021). Bayesian ridge regression shows the best fit for SSR markers in Psidium guajava among Bayesian models. *Sci. Rep.* 11, 13639. doi:10.1038/s41598-021-93120-z

Endelman, J. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi:10.3835/plantgenome2011.08.0024

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinforma.* 12, 186. doi:10.1186/1471-2105-12-186

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi:10.2135/cropsci2009.11.0662

Henderson, A. C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31 (2), 423–447. doi:10.2307/2529430

Huang, H.-J., Ding, S.-F., and Shi, Z.-Z. (2013). Primal least squares twin support vector regression. *J. Zhejiang Univ. Sci. C* 14, 722–732. doi:10.1631/jzus.ciip1301

Jayadeva, R., Khemchandani, R., and Chandra, S. (2007). Twin support vector machines for pattern classification. *IEEE Trans. Pattern Analysis Mach. Intell.* 29, 905–910. doi:10.1109/tpami.2007.1068

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017) *LightGBM: a highly efficient gradient boosting decision tree*. Red Hook, NY, USA: Curran Associates Inc., 3149–3157.

Kelin, W., Muhammad Ali, A., Rasheed, A., Crossa, J., Hearne, S., and Li, H. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* 16, 279–293. doi:10.1016/j.molp.2022.11.004

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 5455–5516. doi:10.1007/s10462-020-09825-6

Kole, C., Thormann, C. E., Karlsson, B. H., Palta, J. P., Gaffney, P., Yandell, B., et al. (2002). Comparative mapping of loci controlling winter survival and related traits in oilseed Brassica rapa and B. napus. *Mol. Breed.* 9, 201–210. doi:10.1023/a:1019759512347

Kung, S. Y. (2014) *Kernel methods and machine learning*. Cambridge: Cambridge University Press.

Lu Zhenxing, Y. Z. G. A. O. X. (2014). Least square twin support vector regression. *Comput. Eng. Appl.* 50, 140–144.

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theor. Appl. Genet.* 115, 1003–1013. doi:10.1007/s00122-007-0627-9

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318. doi:10.1007/s00425-018-2976-9

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Moustafa, G., Tolba, M. A., El-Rifaie, A. M., Ginidi, A., Shaheen, A. M., and Abid, S. (2023). A subtraction-average-based optimizer for solving engineering problems with applications on TCSC allocation in power systems. *Biomimetics* 8, 332. doi:10.3390/biomimetics8040332

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* 5, S11. doi:10.1186/1753-6561-5-S3-S11

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc.* 6, S10. doi:10.1186/1753-6561-6-S2-S10

Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi:10.1198/016214508000000337

Peng, X. (2010). TSVR: an efficient twin support vector machine for regression. *Neural Netw.* 23, 365–372. doi:10.1016/j.neunet.2009.07.002

Rolf, M. M., Garrick, D. J., Fountain, T., Ramey, H. R., Weaber, R. L., Decker, J. E., et al. (2015). Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. *Genet. Sel. Evol.* 47, 23. doi:10.1186/s12711-015-0106-8

Selga, C., Koc, A., Chawade, A., and Ortiz, R. (2021). A bioinformatics pipeline to identify a subset of SNPs for genomics-assisted potato breeding. *Plants* 10, 30. doi:10.3390/plants10010030

Shao, Y.-H., Zhang, C.-H., Yang, Z.-M., Jing, L., and Deng, N.-Y. (2013). An ε-twin support vector machine for regression. *Neural Comput. Appl.* 23, 175–185. doi:10.1007/s00521-012-0924-3

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi:10.1021/ci034160g

Tong, H., and Nikoloski, Z. (2020). Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *J. plant physiology* 257, 153354. doi:10.1016/j.jplph.2020.153354

Trojovský, P., and Dehghani, M. (2023). Subtraction-average-based optimizer: a new swarm-inspired metaheuristic algorithm for solving optimization problems. *Biomimetics* 8, 149. doi:10.3390/biomimetics8020149

Usai, M. G., Goddard, M. E., and Hayes, B. J. (2009). LASSO with cross-validation for genomic selection. *Genet. Res.* 91, 427–436. doi:10.1017/S0016672309990334

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

Vincent, A. M., and Jidesh, P. (2023). An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms. *Sci. Rep.* 13, 4737. doi:10.1038/s41598-023-32027-3

Wang, X., Miao, J., Chang, T., Xia, J., An, B., Li, Y., et al. (2019). Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. *PLOS ONE* 14, e0210442. doi:10.1371/journal.pone.0210442

Yang, T., Sun, X., Yang, H., Liu, Y., Zhao, H., Dong, Z., et al. (2024). Integrated thermal error modeling and compensation of machine tool feed system using subtraction-average-based optimizer-based CNN-GRU neural network. *Int. J. Adv. Manuf. Technol.* 131, 6075–6089. doi:10.1007/s00170-024-13369-2

Yin, L., Zhang, H., Li, X., Zhao, S., and Liu, X. (2022). Hibayes: an R package to fit individual-level, summary-level and single-step bayesian regression models for genomic prediction and genome-wide association studies. *bioRxiv*. doi:10.1101/2022.02.12.480230

Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S.-H., and Patton, R. M. (2015) *Optimizing deep learning hyper-parameters through an evolutionary algorithm*. New York, NY, USA: Association for Computing Machinery.

Zhong, P., Xu, Y., and Zhao, Y. (2012). Training twin support vector regression via linear programming. *Neural Comput. Appl.* 21, 399–407. doi:10.1007/s00521-011-0525-6

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLOS Genet.* 9, e1003264. doi:10.1371/journal.pgen.1003264