A stacking ensemble learning framework for genomic prediction

Mang Liang
Chinese Academy of Agricultural Sciences Institute of Animal Science
Tianpeng Chang
Chinese Academy of Agricultural Sciences Institute of Animal Science
Bingxing An
Chinese Academy of Agricultural Sciences Institute of Animal Science
Xinghai Duan
Chinese Academy of Agricultural Sciences Institute of Animal Science
Lili Du
Chinese Academy of Agricultural Sciences Institute of Animal Science
Xiaoqiao Wang
Chinese Academy of Agricultural Sciences Institute of Animal Science
Jian Miao
Chinese Academy of Agricultural Sciences Institute of Animal Science
Lingyang Xu
Chinese Academy of Agricultural Sciences Institute of Animal Science
Xue Gao
Chinese Academy of Agricultural Sciences Institute of Animal Science
Lupei Zhang
Chinese Academy of Agricultural Sciences Institute of Animal Science
Junya Li
Chinese Academy of Agricultural Sciences Institute of Animal Science
Huijiang Gao (gaohuijiang@caas.cn)
Chinese Academy of Agricultural Sciences Institute of Animal Science

Research

Keywords: Ensemble learning, Stacking, Genomic prediction, Machine learning, Prediction accuracy

Posted Date: August 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52592/v1

A stacking ensemble learning framework for genomic prediction

Mang Liang[1†], Tianpeng Chang[1], Bingxing An[1], Xinghai Duan[1], Lili Du[1], Xiaoqiao Wang[1], Jian Miao[1], Lingyang Xu[1], Xue Gao[1], Lupei Zhang[1], Junya Li[1], Huijiang Gao[1*]

[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, 100193, China.

Mang Liang: liangmang87@163.com

Tianpeng Chang: changtianpeng@126.com

Bingxing An: ABX2HF@126.com

Xinghai Duan: xhduan0411@163.com

Lili Du: dulili1996@126.com

Xiaoqiao Wang: longmao7@live.com

Miao Jian: miaojian6363@163.com

Lingyang Xu: xulingyang@163.com

Xue Gao: gaoxue76@126.com

Lupei Zhang: zhanglupei@caas.cn

Junya Li: lijunya@caas.cn

† These authors contributed equally to this work.

***Corresponding authors:** Dr. Huijiang Gao: gaohuijiang@caas.cn

Running title: A stacking ensemble learning framework for genomic prediction

Address: Yuanmingyuan West Road 2#, Haidian District, Institute of Animal Science, Beijing, 100193, China.

Tel: 86-10-62812769

Fax: 86-10-62817806

Abstract

Background: Machine learning (ML) is perhaps the most useful for the interpretation of large genomic datasets. However, the performance of a single machine learning method in genomic selection (GS) was unsatisfactory in existing research. To improve the genomic predictions, we constructed a stacking ensemble learning framework (SELF) integrated three machine learning methods to predict genomic estimated breeding values (GEBVs).

Results: We evaluated the prediction ability of SELF by three real datasets and compared the prediction accuracy of SELF, base learners, GBLUP and BayesB. For each trait, SELF performed better than base learners, which included support vector regression (SVR), kernel ridge regression (KRR) and elastic net (ENET). The prediction accuracy of SELF had an average 7.70% improvement compared with GBLUP in three datasets. Except for the milk fat percentage (MFP) traits of the German Holstein dairy cattle dataset, SELF more robust than BayesB in the remaining traits.

Conclusions: In this study, we utilized a stacking ensemble learning framework (SELF) to genomic prediction and it performed much better than GBLUP and BayesB in three real datasets with different genetic architecture. Therefore, we believed SEFL had the potential to be promoted to estimate GEBVs in other animals and plants.

Keywords: Ensemble learning, Stacking, Genomic prediction, Machine learning, Prediction accuracy

Background

Genomic selection (GS), introduced by Meuwissen [1], using whole-genome markers' information to predict the genomic estimated breeding values (GEBVs). GS was first applied to daily cattle, which facilitates the rapid selection of superior genotypes and accelerates genetic gain by shortening the breeding cycles [2-4]. After more than ten years of development, GS had been wildly used in livestock and plant breeding programs with high prediction accuracy[4, 5]. Moreover, GS had even applied to improve the prediction of complex disease phenotypes using genotype data[6, 7]. A critical concern in genomic prediction is the prediction accuracy calculated by the Pearson correlation between the predicted estimated breeding values (EBV) and the true breeding values. A novel approach with higher accuracy is what breeders have usually been pursuing, while the exploration of more robust genomic prediction methods has never stopped.

In recent years, there is an increasing interest in applying machine learning (ML) to genomic prediction. ML is a programming computer to optimize a performance criterion using training data[8], which make predictions or decisions without being explicitly programmed to do so. The excellent predictive ability for complex problems lead to ML employed in those industries required high accuracy. e.g., email filtering, face recognition, natural language processing, stock market forecasting. ML had also been used in GS, and it might has the best performance in the interpretation of large-scale genomic data[6]. González et al. suggested that ML was a valuable alternative to well-known parametric methods for genomic selection [9]. Montesinos-López et al. found that the predictions of the multi-trait deep learning model were very competitive with the Bayesian multi-trait and multi-environment model [10]. Jubair and Domaratzki et al. predicted genomic breeding values of Iranian wheat landraces using ensemble learning and concluded that the ensemble learning

was better than single machine learning [11]. There was a clear trend that more and more breeders were trying to use machine learning methods to estimate GEBVs in genomic prediction.

Nowadays, the machine learning methods applied in animal and plant breeding mainly include support vector regression (SVR), random forest (RF), kernel ridge regression (KRR), multi-layer prediction (MLP) and convolutional neural network (CNN). Those machine learning methods possessed the ability to predict GEBVs by building a complex nonlinear model considered the interaction effects and epistatic effects[12]. Nevertheless, the prediction accuracy of those single machine learning methods did not improve much than the traditional genomic prediction methods (genomic bets linear unbiased regression (GBLUP), ridge regression BLUP (rrBLUP), BayesB et al.). Ogutu et al. compared the prediction accuracy of random forest (RF), boosting and support vector machine (SVM) with rrBLUP in simulated dataset, in which rrBLUP outperformed the three machine learning methods [13]. Montesinos-López et al. compared the prediction performance of multi-layer prediction, support vector machine with the Bayesian threshold genomic best linear unbiased prediction (TGBLUP) and believed that the reliability of two machine learning methods was comparable to TGBLUP, in some case, outperformed TGBLUP [14]. Even though the achievement of ML in GS had not been fantastic, the breeders still had the confidence in exploration of ML because of its outstanding performance in other majors.

To further improve the prediction accuracy of ML in GS, one available solution is to integrate several machine learning methods simultaneously in genomic prediction. Ensemble learning is a ML paradigm that multiple learners trained to solve the same problem and usually much more robust than that of a single learner [15, 16]. Stacking, Boosting and Bagging was the most commonly utilized integration strategy of ensemble learning and the stacking is gradually showing its powerful

prediction capabilities for complex problems. In other majors, stacking had been applied to date prediction, protein interaction structure prediction, credit scoring and cancer detection et al [17-20]. However, the application of stacking in GS had rarely been reported.

To this end, the objective of this study was to improve genomic predictions by using a stacking ensemble learning framework (SELF). In the experiment, support vector regression (SVR), kernel ridge regression (KRR) and elastic net (ENET) were selected as the base learner and the ordinary Least Squares (OLS) linear regression was chosen as the meta learner to construct the SELF model. Afterward, we evaluated the SELF model using two animal datasets (Chinese Simmental Beef Cattle dataset and German Holstein dairy cattle dataset) and a plant dataset (Loblolly pine dataset). In order to assess the performance of SELF, we compared the prediction accuracy calculated by the Pearson correlation of SELF with the base learners, GBLUP and BayesB. Finally, the 20-folds cross-validation was employed to mitigate the impact of the accidental error.

Materials and methods

Dataset

Chinese Simmental Beef Cattle dataset

The Chinese Simmental Beef Cattle population included 1,217 individuals, all of them were born between 2008 and 2014 in Ulgai, Xilingolia of China, and were slaughtered at the average age of 16-18 months. In the progress of slaughter, the carcass trait was assessed according to the Institutional Meat Purchase Specifications for Fresh Beef Guidelines. In this study, we selected three important economic traits for latter analysis: live weight (LW), carcass weight (CW) and eye muscle

area (EMA). Before training the model, we used fix effects (year of birth, sex, fattening duration, initial body weights) to correct the phenotype data as the following formula:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \ y^*$$

where $\boldsymbol{y}$ is the vector of observed phenotypic values, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{X}$ is the design matrix of fix effects, y^* is the random residual. In subsequent analysis, the residual y^* was used as the phenotype data for each prediction model. The describe of statistics used for each trait to estimate variance components were presented in Table 1.

The whole of the Chinese Simmental Beef Cattle population was genotyped by Illumina BovineHD BeadChip (770K). The quality control criteria of genotype data were as follows: minor allele frequency (MAF) >0.05, call rate (CR) >0.95 and *P*-value > 10^{-5} from Hardy-Weinberg equilibrium (HWE).

German Holstein dairy cattle dataset

The dataset of German Holstein dairy cattle consisted of 5024 bulls with genotypes and phenotypes [21]. The genotype data were generated with the Illumina Bovine SNP50 Beadchip (42,551 SNPs). All of the SNPs meet the following standards: HWE *P*-value > 10^{-4}, CR > 0.95 and MAF > 0.01[22]. Because the dataset that we used was not the original data, all the phenotype data had been standardized (mean=0, standard deviation=1). More details about the original dataset can refer to Zhang et al. [21]. The phenotypes involves three traits, milk yield (MY), milk fat percentage (MFP) and somatic cell score (SCS). The describe of statistics for German Holstein dairy cattle dataset was demonstrated in Table 1[21]. The three traits may represent three genetic architectures of complex traits compose, (1) one major gene and a large number of small effect loci (MFP), (2) few moderate

effect loci and many small effect loci (MY) and (3) many loci with small effects (SCS), respectively[21, 22].

Loblolly pine dataset

The loblolly pine dataset comprised 951 individuals from 61 families with each individual systemically recorded 17 traits [23]. The original dataset All the individuals were genotyped with an Illumina Iminium assay (7216 SNPs) [21]. After quality control, the genotypes contained 4853 polymorphic SNPs, which also used in Resende et al. and Zhang et al. [21, 23]. The phenotypes that we used was a subset of the original phenotype data. Only one trait in growth traits (total stem height, HT), development traits (crown width along the planting beds, CWAL) and wood quality traits (tree stiffness, TS) were chosen to implement prediction models, respectively. The describe of statistics for the loblolly pine dataset was shown in Table 1.

Methods

Stacking

Stacking is a form of meta-learning that yielded impressive results by designing novel deep learning architectures[24]. The core idea of stacking is using the base learners to generate metadata for the inputs and then utilize another learner, generally called the meta-learner, to process metadata. Due to the base learners usually called level 0 learner, the meta learner called level 1 learner and the meta learner stacked on the based learner, hence the name stacking[24]. In genomic prediction, the SELF performed in two steps: firstly, training a series of single machine learning methods to generate metadata using markers' information; secondly, training a meta

learner to predict GEBVs using metadata. The data flow of SELF for genomic prediction was shown in Figure 1.

The base learner we employed to construct SELF involves SVR, KRR and elastic net (ENET). SVR and KRR constructed a nonlinear model to predict GEBVs and ENET estimated the GEBVs by building a linear regression. SVR and ENET had been applied to GS in previous researches [9, 25, 26]. Generally, the meta learner should be a relatively simple ML algorithm in order to avoid overfitting and possess the ability to handle correlated inputs with no assumptions about the independence of features because of the inputs of meta-learner will be highly correlated[24]. Considered the above requirements, the ordinary least squares (OLS) linear regression was chosen as the meta-learner in the SELF. During the SELF-model training, we did not take the genotypes as the inputs directly but replaced it with the genomic relationship matrix derived by genotypes[12]. Although this might reduce the prediction accuracy of a single machine learning method, it would significantly reduce the time and the memory required for computation. In theory, the calculation time of SELF will be five times a single machine learning method, because of the 5-fold cross-validation was used to generate metadata. If we used the same steps of previous researches to apply the genotypes as the inputs, the computation time of SELF would be unacceptable. Finally, SELF was run in Python (V3.7) with the help of *sklearn* (V0.22) package.The genomic relationship matrix **G** was calculated as[27]:

$$\mathrm{G} = \frac{MM'}{\sum_{l=1}^{m} 2p_j q_j}$$

where **M** is a n×m matrix (n is the number of individuals, m is the number of markers) and elements of column j in **M** are $0 - 2p_j$, $1 - 2p_j$ and $2 - 2p_j$ for genotypes A_1A_1, A_1A_2 and A_2A_2; q_j is allele frequency A_1 at locus j, p_j is allele frequency A_2 at locus jth.

Support vector regression

Support vector machine (SVM), developed by Vapnik[28], is grounded in statistical learning theory. Support vector regression (SVR) is an application of SVM for regression. SVR utilizes a linear or nonlinear kernel function to map the original space to a higher dimensional feature space[29, 30]. Then building the linear prediction model on feature space. The SVR problem formalized as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} L_\varepsilon\left(f(x_i) - y_i\right)$$

where C is the regularization constant, L_ε is the ε-insensitive loss

$$L_\varepsilon(z) = \begin{cases} 0, & if \;\; z < \varepsilon \\ |z| - \varepsilon, & otherwise \end{cases}$$

where $\mathrm{z} = f(x_i) - y_i$.Through a series of optimization processes, the SVR can be written as:

$$f(x) = \sum_{i=1}^{m} (\widehat{\alpha_i} - \alpha_i) k(x, x_i) + b$$

where $\mathrm{k}(x_i,\ x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function. In this study, the Gaussian kernel was used to map original data and the most suitable parameters of C and ε for each trait were determined by grid search. The function *SVR* in *sklearn* package (V 0.22) was used to implement SVR methods.

Kernel ridge regression

The difference between KRR and ridge regression is that KRR exploits the kernel trick to define a higher dimensional feature space and then build the ridge regression model in feature space [31-33]. For KRR, the final prediction function can be written as the following:

$$f(x) = k^{'}(K + \lambda I)^{-1} y$$

where K is the so-called Gram matrix with entries $K_{ij} = \langle \phi(x_i) \cdot \phi(x_j \rangle$, k is a vector with entries $k_i = \langle \phi(x) \cdot \phi(x_i \rangle = k(x, x_i)$ with $i = 1,2,3, \ldots,$ n, n is the number of training samples; ***I*** is the identity matrix, λ is the ridge parameter. In this study, the kernel we used to transform input data was selected by the grid search method.

Elastic net

Elastic net is a linear regression model trained with both ℓ_1 and ℓ_2-norm regularization of the coefficients. This combination leads to the ENET almost has the advantages of Lasso and ridge regression simultaneously. Thus, ENET can learn a sparse model where few of the weights are non-zero and maintain the regularization properties[34]. The progress of training the ENET model can be seen as an optimization process for:

$$\min_{\omega} \frac{1}{2n_{train}} \|X\omega - y\|_2^2 + \alpha\rho\|\omega\|_1 + \frac{\alpha(1-\rho)}{2}\|\omega\|_2^2$$

for this study, ***X*** is a matrix of the training section of **G** matrix, **ω** is the vector of weights, α and ρ are the parameters that determined by grid search.

Genomic best linear unbiased prediction

The basic GBLUP method can be built by the following equation[27, 35]:

$$y = 1\mu + Zg + \ e$$

where ***y*** is the vector of the correct phenotype, μ is the overall mean, ***1*** is a vector of ones, **Z** is a design matrix that allocates records to breeding values, ***g*** is a vector of genomic breeding values, ***e*** is a vector of residuals. Random residuals were assumed that $e \sim N(0,\ I\sigma_e^2)$ where σ_e^2 is the residual variance, ***I*** is an identity matrix. g assumed that $g \sim N(0, \mathbf{G}\sigma_g^2)$ where σ_g^2 is the

additive genetic variance, $\boldsymbol{G}$ is the marker-based genomic relationship matrix. To implement GBLUP, we used the *mixed.solve* function of *rrBLUP* package in the R V3.5.

BayesB

BayesB assumed a priori that many markers have no effects, and some have an effect coming from gamma or exponential distribution [36]. The formula of BayesB can be written as the following:

$$\mathrm{y} = \sum_{j=1}^{p} m_j \alpha_j + e$$

where **y** is a vector of phenotypes; m_j is the *j*th maker; α_j is the effect of the *j*th maker and $\alpha_j \sim N(0, \sigma^2_{\alpha_j})$. The variance of α_j is assigned an informative prior to show the presence (with the probability of $1-\pi$) and absence (with the probability of π) of the marker *j*. The π was determined by the experience before build the BayesB model.

Results

We evaluated SELF-model using both real animal and plant datasets. Each dataset was randomly divided into twenty folds by 20-fold cross-validation. Each fold would be the testing set, the remaining nineteen folds would be grouped into the training set. The training set trained the SELF model to predict the GEBVs of individuals in the testing set. The accuracy that we showed in the result section was the mean of prediction accuracy of each testing set, which was quantified with the Pearson correlation. The methods that we compared with SELF included: 1) three base learners: SVR, KRR, ENET; 2) GBLUP; 3) BayesB. Comparison with three base learner was used to verify whether SELF improved the prediction accuracy through an integrated strategy, and the specific

performance of SELF in GS was reflected by comparing with GBLUP and BayesB.

Comparison of prediction accuracy of base learners, GBLUP and BayesB

Firstly, we described the prediction accuracy of three base learners, GBLUP and BayesB for three datasets in Table 2. BayesB and KRR outperformed other methods in three traits, which possessed the best predictive power. The prediction accuracy of GBLUP and ENET was higher than other methods in two traits. The prediction performance of SVR was the worst, and the prediction accuracy of SVR was always lower than one of the other methods. For three base learners, the prediction accuracy of KRR was the highest. However, the prediction accuracy gap between these methods was not significant, and the ability to estimate the GEBVs was comparable.

Comparison of prediction accuracy of SELF and base learners

Figure 2 showed the comparison of the prediction accuracy of the base learners and SELF for nine traits. The red one represents the prediction accuracy of SELF. SELF performed better than all of the base learners for each trait. Especially for CWAL, HT and EMA, the prediction accuracy of SELF improved 9.97%, 7.36% and 6.40% compared to the highest prediction accuracy of base learners respectively. Among the three base learners, the prediction ability of KRR was comparable to SELF in German Holstein dairy cattle dataset.

Comparison of prediction accuracy of SELF, GBLUP and BayesB

Table 3 demonstrated the prediction accuracy of GBLUP, BayesB and SELF for three datasets. For the Chinese Simmental beef cattle dataset, the prediction accuracy of SELF was higher than GBLUP and BayesB evidently, and the average improvement of 11.68% going from SELF to GBLUP. For

the German Holstein daily cattle, except for the trait of MFP, SELF performed better than BayesB and GBLUP. For the loblolly pine dataset, SELF predicted more accurate GEBVs than GBLUP and BayesB, and the improvement was 14.18% for TS compared with GBLUP. Figure 3 presented more intuitive that the SELF performed better than GBLUP and BayesB. Comparing the prediction accuracy of SELF to GBLUP, the average prediction accuracy increased by 7.70% in nine traits.

Discussion

A large number of researches had tried to apply single machine learning methods in genomic prediction [11, 14, 37, 38]. However, the single machine learning methods in the most previous studies only performed well on several traits [13, 14, 38, 39]. Therefore, we proposed a new strategy to utilize machine learning methods in genomic prediction. We implemented the genomic prediction by using a stacking ensemble learning framework integrated three machine learning machine methods to predict GEBVs simultaneously. To examine the prediction ability of SELF, we compared the prediction accuracy of SELF with GBLUP and BayesB using animal and plant datasets with a variety of genetic architecture. Considering the computation time and overfitting, we employed the genotypes derived relationship matrix as the inputs rather than using the genotypes directly[12].

The prediction accuracy of base learners, GBLUP and BayesB

The datasets we utilized in this research included Chinese Simmental Beef Cattle dataset, German Holstein dairy cattle dataset and Loblolly pine dataset. The GBLUP and BayesB were comparable to the single machine learning methods. To prove the results of GBLUP and BayesB of this study were highly reliable, we compared the prediction accuracy of GBLUP and BayesB with Wang et al.,

Zhang et al. and Resende et al. [21, 23, 25]. Wang et al. compared GBLUP with BayesB in the Chinese Simmental Beef Cattle dataset. Zhang et al. and Resende et al. compared the prediction accuracy of different methods on the German Holstein dairy cattle dataset and Loblolly pine dataset, respectively. Overall, the results were consistent. Since the method we used was slightly different from the previous study, the accuracy differed in individual traits. Using a single machine learning method to estimate GEBVs on the three datasets had not been reported. However, a vast quantity of researches had compared the prediction accuracy of the single machine learning method with GBLUP or Bayesian family methods on other populations yet, which could provide a practical reference to evaluate the performance of single machine learning methods. The results of Ghafouri et al. and Long et al. indicated that GBLUP had better prediction accuracy than SVR and RF in most cases and the performance SVR with Gaussian kernel was comparable Bayesian Lasso [38, 40]. It was similar to our results that a single machine learning did not significantly better than GBLUP and Bayes methods.

Excellent predictive performance of SELF

Compared to GBLUP, the average prediction accuracy of SELF increased by 7.70% for nine traits, which was significant for animal and plant breeding. Especially for the beef cattle with longer generation intervals, such a considerable prediction accuracy increase will greatly accelerate genetic gain. Actually, building a SELF model to predict a specific problem with high accuracy was very difficult due to there were so much single machine learning methods could be incorporated into the model as the base learner or meta-learner. Thence, we referred to previous studies that using machine learning methods to estimated GEBVs and combined with our

experience to select the candidate base learner. Besides, a single-layer framework or multi-layer framework also should be premeditated when constructing frameworks. Considering the overfitting always accompanied by the machine learning methods in GS and the calculating time of SELF, we determined a single layer stacking framework finally.

Before constructing SELF, RF, SVR, KRR and ENET were chosen as the candidate base learner. Among these, RF, SVR and ENET had been used in previous studies to estimate GEBVs [13, 38, 40-42]. Although KRR used in GS was reported rarely, it had been frequently applied to classification and regression for other majors [31, 43-45]. To make SELF more diversification, ENET was included in the SELF model, and the ENET had performed better than GBLUP in Wang et al. [18]. After we tried to predict GEBVs using four base learners, we determined to drop RF form the SELF due to RF greatly increased the computation time of SELF. Therefore, the final SELF was constructed by SVR, KRR and ENET, in which the base learners constructed different types of models to estimate the GEBVs. SVR and KRR using the kernel function to map the input data into the higher dimensional feature space to predict genomic breeding value with the nonlinear model, and ENET estimated breeding values by a linear model. Generally, it was reasonable to employ different learning algorithms to seek out the relationship between the feature and the target variable[24]. For a regression example (Fig 4), using a stacked ensemble with linear and nonlinear regression would be able to outperform either a single linear or a single nonlinear model significantly. Even if we utilized the best prediction of the linear model and the nonlinear model as the outputs of the integrated model directly without stacking, the integrated model would greatly perform better than the two models. Therefore, the SELF we constructed could learn more

characteristics in different aspects of the input data and performed better than either of the base learners.

In addition, the input data of base learners that we used might be another reason that caused the higher prediction accuracy of SELF. The most of previous studies employed genotypes as the inputs of machine learning methods directly. Nevertheless, there was a prominent problem that the number of markers far bigger than the number of individuals. If we used genotypes with no transformed, the number of variables in the prediction model would be an astronomical figure compared to group size. Because most single machine learning methods did not assume that most markers have no effects like BayesB or LASSO. Although single machine learning methods were able to solve the problem of "big P and small N", stronger overfitting would be inevitable, which also decreased the accurate of the final predicted GEBVs of SELF. Using the genomic relationship matrix as the input data was entirely different, the genomic relationship matrix was a $n \times n$ matrix, which size was determined by the group size n. Therefore, the number of variables in prediction model would consistent with the number of individuals. Although it might reduce the prediction accuracy of base learners, it dramatically reduces the risk of overfitting simultaneously, which led to the SELF could exert the potential to improve prediction accuracy by integrating single machine learning methods.

Conclusion

In conclusion, we proposed a stacking ensemble learning framework integrated SVR, KRR and ENET to predict GEBVs. The excellent performance of SELF in the variety of genetic architecture datasets indicated that SELF with the great potential to improve genomic predictions in other animal

and plant populations.

Abbreviations

CR: Call rate

CW: Carcass weight

CWAL: Crown width along the planting bed

EMA: Eye muscle area

ENET: Elastic net

GBLUP: Genomic best linear unbiased regression

GEBV: Genomic estimated breeding values

HT: Total stem height

HWE: Hardy-Weinberg equilibrium

KRR: Kernel ridge regression

LW: Live weight

MAF: Minor allele frequency

MFP: Milk fat percentage

MY: Milk yield

OLS: Ordinary least squares linear regression

SCS: Somatic cell score

SD: Standard deviation

SELF: Stacking ensemble learning framework

TS: Tree stiffness

SVR: Support vector regression

Acknowledgments

The authors would like to thank Huijiang Gao for proofreading and guidance, and all staff at the experimental cattle unit in Beijing for animal care and sample collection.

Author Contributions

HJG and JYL conceived and designed the study. ML and BXA performed statistical analyses and wrote the paper. ML, JM, XQW wrote the code.TPC, BXA, XHD, LLD and JM participated in data analyses. LPZ, LYX and XG participated in the design of the study and contributed to acquisition of data. All authors read and approved the final manuscript.

Funding

This work was supported by funds from the National Natural Science Foundations of China (31872975) and the Program of National Beef Cattle and Yak Industrial Technology System (CARS-37).

Availability of data and materials

Chinese Simmental Beef Cattle dataset: Data is available from the Dryad Digital Repository: DOI:10.5061/dryad.4qc06.

German Holstein dairy cattle dataset: data: Data can be obtained at https://www.g3journal.org/content/5/4/615.supplemental.

Loblolly pine dataset: The quality-controlled genotypes can get at https://www.genetics.org /highwire/filestream/412827/field_highwire_adjunct_files/1/FileS1.zip and the complete phenot ypes at https://www.genetics.org/highwire/filestream/412827/field_highwire_adjunct_files/4/File S4.xlsx.

Ethics approval and consent to participate

Animal experiments were approved by the Science Research Department of the Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (CAAS) (Beijing, China). There was no use of human participants, data or tissues.

Consent for publication

Not applicable

Conflict of interest

The authors certify that there is no actual or potential conflict of interest with this article.

Figure legends

Fig. 1.The data flow of stacking ensemble learning framework for genomic prediction, from original data to the base learners, creating metadata for the meta-learner. G genotypes derived genomic relationship matrix, SVR support vector regression, KRR kernel ridge regression, ENET elastic net, OLS ordinary least squares linear regression.

Fig. 2. The comparison of prediction accuracy of SVR (blue violet), KRR (dodger blue), ENET

(dark orange) for nine traits. **a** for live weight. **b** for carcass weight. **c** for eye muscle area. **d** for milk yield. **e** for milk fat percentage. **f** for somatic cell score. **g** for total stem height. **h** for crown width along the planting beds. **i** for tree stiffness.

Fig. 3. The comparison of prediction accuracy of SELF (red), GBLUP (dodger blue) and BayesB (dark orange) for three datasets. **a** for Chinese Simmental Beef Cattle dataset. **b** for German Holstein dairy cattle dataset. **c** for Loblolly pine dataset. LW live weight, CW carcass weight, EMA eye muscle area, MY milk yield, MFP milk fat percentage, SCS somatic cell score, HT total stem height, CWAL crown width along the planting beds, TS tree stiffness. GBLUP genomic best linear unbiased prediction, SELF a stacking ensemble learning framework.

Fig 4. A simple regression example. Using a stacked ensemble with linear and nonlinear regression will be able to outperform either a single linear or a nonlinear model significantly.

Tables

Table 1. Descriptive statistics of phenotype data used in genomic prediction

Dataset	Trait	N[a]	h^2	Mean	SD
Beef cattle	LW	1216	0.53	505.26	70.76
	CW	1216	0.44	271.36	45.65
	EMA	1117	0.57	85.21	13.32
Dairy cattle	MY	5024	0.95	370.79	641.60
	MFP	5024	0.94	-0.06	0.28
	SCS	5024	0.88	102.32	11.73
Loblolly pine	HT	861	0.31	20.30	5372.89
	CWAL	861	0.27	2.44	745.29
	TS	910	0.37	0.10	1.49

N[a] number of the animal with phenotypes, h^2 heritability, SD standard deviation.

LW live weight, CW carcass weight, EMA eye muscle area, MY milk yield, MFP milk fat percentage, SCS somatic cell score, HT total stem height, CWAL crown width along the planting beds, TS tree

stiffness

Table 2. Prediction accuracy of SVR, KRR, ENET, GBLUP and BayesB for three datasets

Dataset	Trait	SVR	KRR	ENET	GBLUP	BayesB
	LW	0.274	**0.283**	0.276	0.256	0.265
Beef cattle	CW	0.307	**0.315**	**0.315**	0.292	0.282
	EMA	0.280	0.281	0.2851	**0.292**	0.281
	MY	0.764	**0.781**	0.762	0.768	0.767
Dairy cattle	MFP	0.796	0.828	0.797	0.832	**0.855**
	SCS	0.706	0.751	0.722	**0.752**	0.731
	HT	0.340	0.352	**0.366**	0.349	0.365
Loblolly pine	CWAL	0.352	0.359	0.369	0.384	**0.400**
	TS	0.397	0.407	0.398	0.366	**0.418**

The accuracy was calculated by the Pearson correlation. LW live weight, CW carcass weight, EMA eye muscle area, MY milk yield, MFP milk fat percentage, SCS somatic cell score, HT total stem height, CWAL crown width along the planting beds, TS tree stiffness. SVR support vector regression, KRR kernel ridge regression, ENET elastic net, GBLUP genomic best linear unbiased prediction

Table 3. Prediction accuracy of base learners and SELF for three datasets

Dataset	Trait	GBLUP	BayesB	SELF
	LW	0.256	0.265	**0.299**
Beef cattle	CW	0.292	0.282	**0.334**
	EMA	0.292	0.281	**0.303**
	MY	0.768	0.767	**0.783**
Dairy cattle	MFP	0.832	**0.855**	0.832
	SCS	**0.752**	0.731	**0.752**
	HT	0.349	0.365	**0.393**
Loblolly pine	CWAL	0.384	0.400	**0.406**
	TS	0.366	**0.418**	**0.418**

The accuracy was calculated by the Pearson correlation. LW live weight, CW carcass weight, EMA eye muscle area, MY milk yield, MFP milk fat percentage, SCS somatic cell score, HT total stem height, CWAL crown width along the planting beds, TS tree stiffness. SVR support vector regression, KRR kernel ridge regression, ENET elastic net, GBLUP genomic best linear unbiased prediction

Reference

1. Meuwissen THE, Hayes B, Goddard M: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**(4):1819-1829.
2. Tong H, Küken A, Nikoloski Z: **Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth**. *Nature communications* 2020, **11**(1):1-9.
3. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y: **Genomic selection in plant breeding: methods, models, and perspectives**. *Trends in plant science* 2017, **22**(11):961-975.
4. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: Progress and challenges**. *Journal of dairy science* 2009, **92**(2):433-443.
5. Heffner EL, Sorrells ME, Jannink JL: **Genomic selection for crop improvement**. *Crop Science* 2009, **49**(1):1-12.
6. De Los Campos G, Gianola D, Allison DB: **Predicting genetic predisposition in humans: the promise of whole-genome markers**. *Nature Reviews Genetics* 2010, **11**(12):880-886.
7. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J: **Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties**. *PLoS one* 2013, **8**(4):e61318.
8. Alpaydin E: **Introduction to machine learning**: MIT press; 2020.
9. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J: **Applications of machine learning methods to genomic selection in breeding wheat for rust resistance**. *The plant genome* 2018, **11**(2):1-15.
10. Montesinos-López OA, Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J: **Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits**. *G3: Genes, Genomes, Genetics* 2018, **8**(12):3829-3840.
11. Jubair S, Domaratzki M: **Ensemble supervised learning for genomic selection**. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2019*. IEEE: 1993-2000.
12. Gianola D, Okut H, Weigel KA, Rosa GJ: **Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat**. *BMC genetics* 2011, **12**(1):87.
13. Ogutu JO, Piepho H-P, Schulz-Streeck T: **A comparison of random forests, boosting and support vector machines for genomic selection**. In: *BMC proceedings: 2011*. Springer: S11.
14. Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, Juliana P, Singh R: **A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding**. *G3: Genes, Genomes, Genetics* 2019, **9**(2):601-618.
15. Polikar R: **Ensemble based systems in decision making**. *IEEE Circuits and systems*

magazine 2006, **6**(3):21-45.

16. Thomas GD: **Machine learning research: Four current directions**. *Artificial Intelligence, Magazine* 1997, **18**(4):97-136.
17. Wang G, Hao J, Ma J, Jiang H: **A comparative assessment of ensemble learning for credit scoring**. *Expert systems with applications* 2011, **38**(1):223-230.
18. Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y: **Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection**. *Applied Soft Computing* 2019, **77**:188-204.
19. Sun W, Trevor B: **A stacking ensemble learning framework for annual river ice breakup dates**. *Journal of Hydrology* 2018, **561**:636-650.
20. Yi H-C, You Z-H, Wang M-N, Guo Z-H, Wang Y-B, Zhou J-R: **RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information**. *BMC bioinformatics* 2020, **21**(1):1-10.
21. Zhang Z, Erbe M, He J, Ober U, Gao N, Zhang H, Simianer H, Li J: **Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix**. *G3: Genes, Genomes, Genetics* 2015, **5**(4):615-627.
22. Yin L, Zhang H, Zhou X, Yuan X, Zhao S, Li X, Liu X: **KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters**. *Genome Biology* 2020, **21**(1):1-22.
23. Resende MF, Muñoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M: **Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.)**. *Genetics* 2012, **190**(4):1503-1510.
24. Kyriakides G, Margaritis KG: **Hands-On Ensemble Learning with Python**. In.; 2019.
25. Wang X, Miao J, Chang T, Xia J, An B, Li Y, Xu L, Zhang L, Gao X, Li J: **Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle**. *PloS one* 2019, **14**(2):e0210442.
26. Ogutu JO, Schulz-Streeck T, Piepho H-P: **Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions**. In: *BMC proceedings: 2012*. Springer: S10.
27. VanRaden PM: **Efficient methods to compute genomic predictions**. *Journal of dairy science* 2008, **91**(11):4414-4423.
28. Vapnik V: **The nature of statistical learning theory**: Springer science & business media; 2013.
29. Müller AC, Guido S: **Introduction to machine learning with Python: a guide for data scientists**: " O'Reilly Media, Inc."; 2016.
30. Li H: **Statistical Learning Methods (Second Edition)**. In.: Tsinghua University Press; 2019.
31. Douak F, Melgani F, Benoudjit N: **Kernel ridge regression with active learning for wind speed prediction**. *Applied energy* 2013, **103**:328-340.
32. Exterkate P, Groenen PJ, Heij C, van Dijk D: **Nonlinear forecasting with many predictors using kernel ridge regression**. *International Journal of Forecasting* 2016, **32**(3):736-753.
33. He J, Ding L, Jiang L, Ma L: **Kernel ridge regression classification**. In: *2014 International Joint Conference on Neural Networks (IJCNN): 2014*. IEEE: 2263-2267.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: **Scikit-learn: Machine learning in Python**. *the*

Journal of machine Learning research 2011, **12**:2825-2830.

35. Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix**. *Genetics research* 2009, **91**(1):47-60.
36. Meuwissen TH, Solberg TR, Shepherd R, Woolliams JA: **A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value**. *Genetics Selection Evolution* 2009, **41**(1):2.
37. Lenz PR, Nadeau S, Mottet MJ, Perron M, Isabel N, Beaulieu J, Bousquet J: **Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce**. *Evolutionary applications* 2020, **13**(1):76-94.
38. Long N, Gianola D, Rosa GJ, Weigel KA: **Application of support vector regression to genome-assisted prediction of quantitative traits**. *Theoretical and applied genetics* 2011, **123**(7):1065.
39. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J: **Applications of machine learning methods to genomic selection in breeding wheat for rust resistance**. *The plant genome* 2018, **11**(2).
40. Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A: **Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation**. *Animal Production Science* 2017, **57**(2):229-236.
41. González-Recio O, Rosa GJ, Gianola D: **Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits**. *Livestock Science* 2014, **166**:217-231.
42. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics**. *Nature Reviews Genetics* 2015, **16**(6):321-332.
43. Avron H, Kapralov M, Musco C, Musco C, Velingker A, Zandieh A: **Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees**. In: *International Conference on Machine Learning: 2017*. 253-262.
44. Chang X, Lin S-B, Zhou D-X: **Distributed semi-supervised learning with kernel ridge regression**. *The Journal of Machine Learning Research* 2017, **18**(1):1493-1514.
45. Naik J, Satapathy P, Dash P: **Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression**. *Applied Soft Computing* 2018, **70**:1167-1188.

Figures

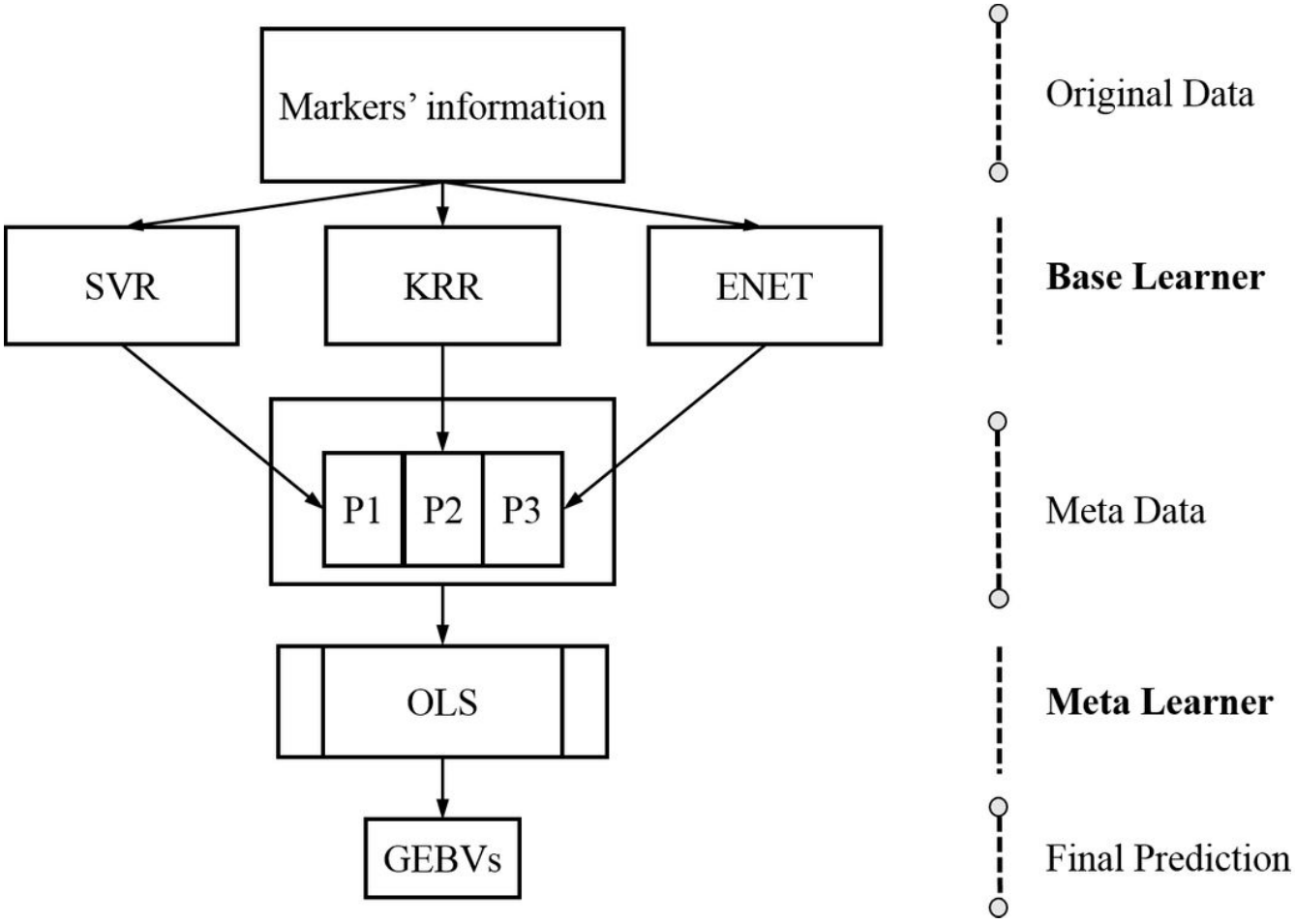

Figure 1

The data flow of stacking ensemble learning framework for genomic prediction, from original data to the base learners, creating metadata for the meta-learner. G genotypes derived genomic relationship matrix, SVR support vector regression, KRR kernel ridge regression, ENET elastic net, OLS ordinary least squares linear regression.

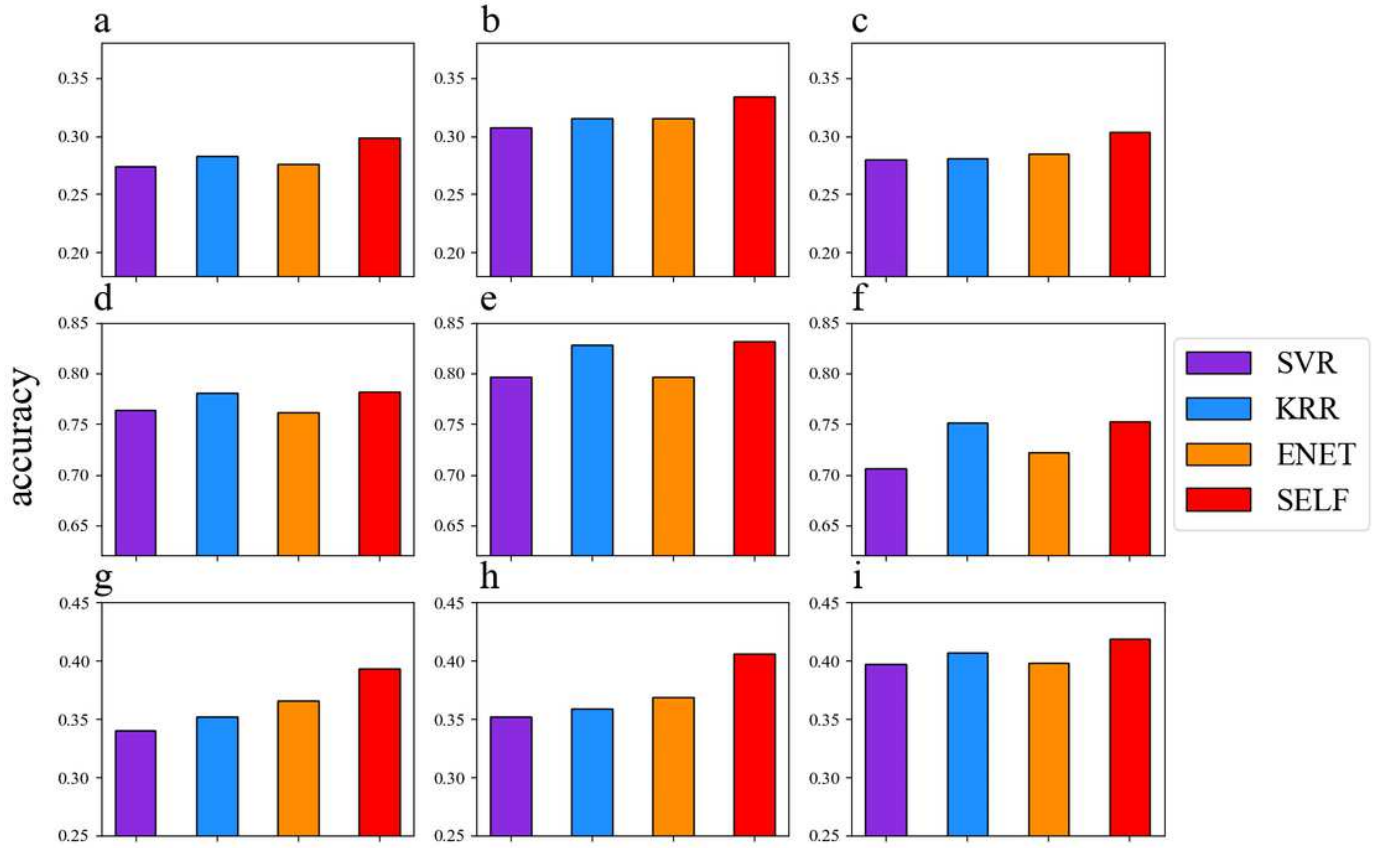

Figure 2

The comparison of prediction accuracy of SVR (blue violet), KRR (dodger blue), ENET (dark orange) for nine traits. a for live weight. b for carcass weight. c for eye muscle area. d for milk yield. e for milk fat percentage. f for somatic cell score. g for total stem height. h for crown width along the planting beds. i for tree stiffness.

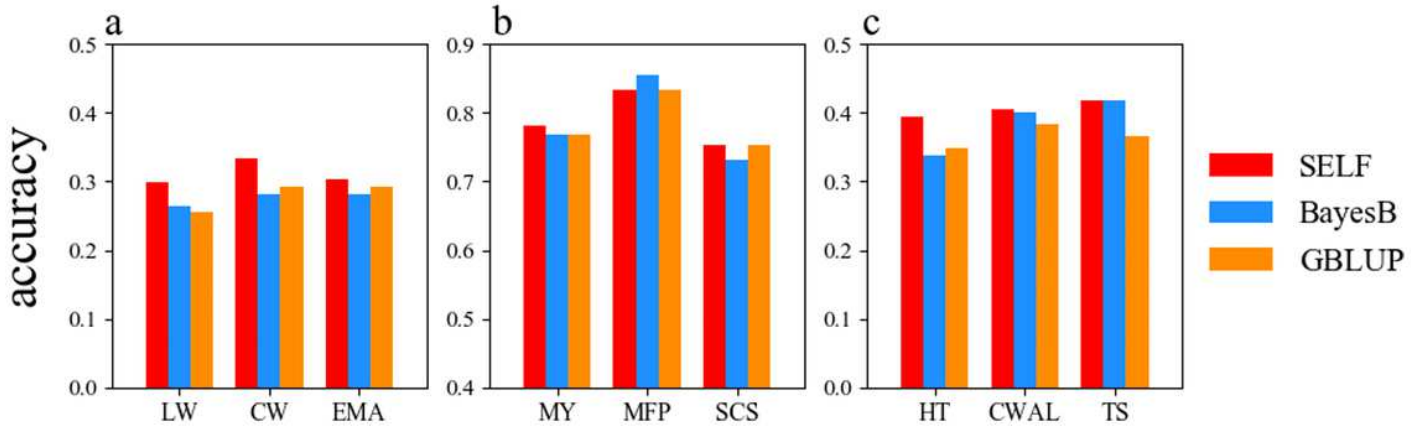

Figure 3

The comparison of prediction accuracy of SELF (red), GBLUP (dodger blue) and BayesB (dark orange) for three datasets. a for Chinese Simmental Beef Cattle dataset. b for German Holstein dairy cattle dataset. c for Loblolly pine dataset. LW live weight, CW carcass weight, EMA eye muscle area, MY milk yield, MFP milk fat percentage, SCS somatic cell score, HT total stem height, CWAL crown width along the planting beds, TS tree stiffness. GBLUP genomic best linear unbiased prediction, SELF a stacking ensemble learning framework.

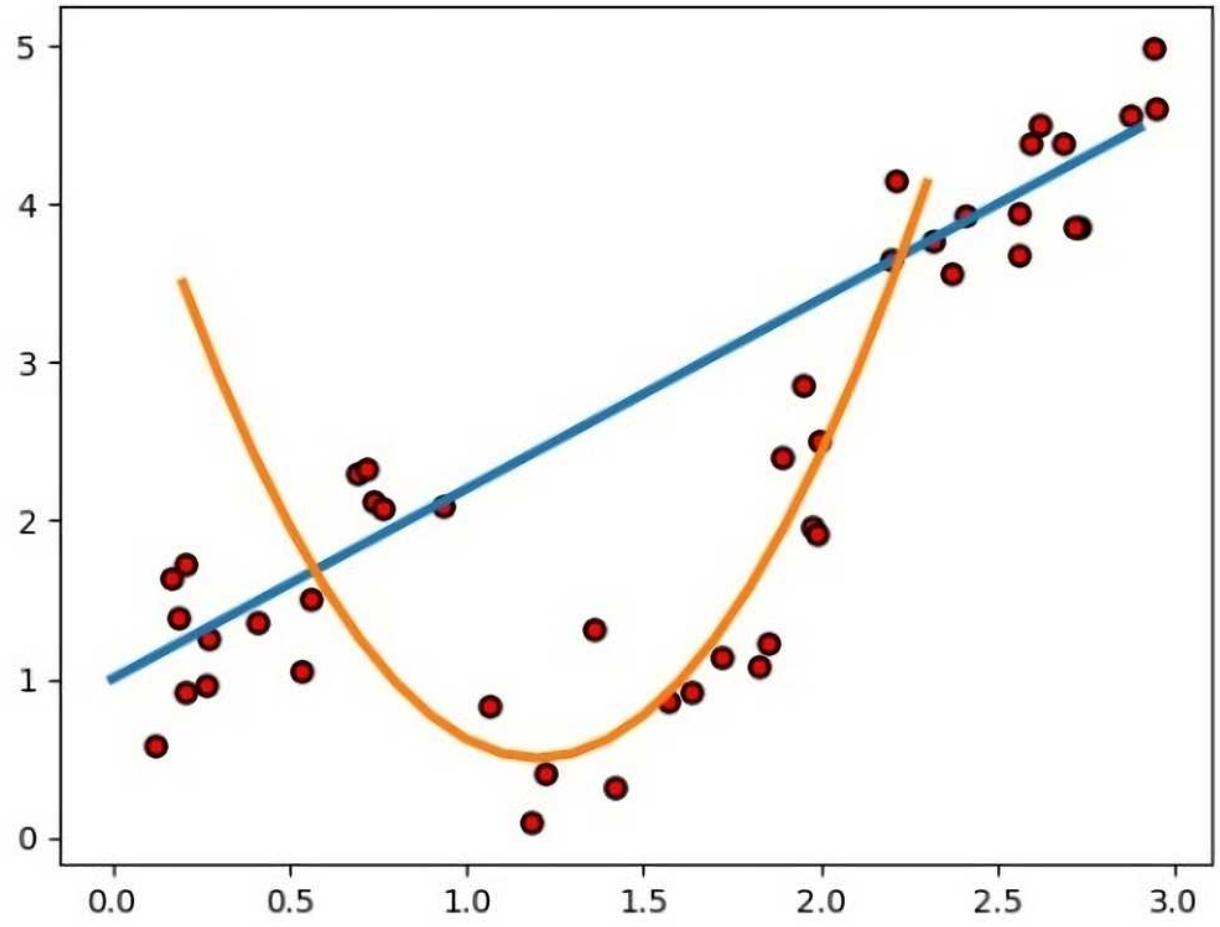

Figure 4

A simple regression example. Using a stacked ensemble with linear and nonlinear regression will be able to outperform either a single linear or a nonlinear model significantly.