MDPI

Article

Systematic Evaluation of Genomic Prediction Algorithms for Genomic Prediction and Breeding of Aquatic Animals

Kuiqin Wang [1], Ben Yang [1], Qi Li [1,2] and Shikai Liu [1,2,*]

1 Key Laboratory of Mariculture, Ministry of Education, College of Fisheries, Ocean University of China, Qingdao 266003, China

2 Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

* Correspondence: liushk@ouc.edu.cn; Tel.: +86-0532-82032595

Abstract: The extensive use of genomic selection (GS) in livestock and crops has led to a series of genomic-prediction (GP) algorithms despite the lack of a single algorithm that can suit all the species and traits. A systematic evaluation of available GP algorithms is thus necessary to identify the optimal GP algorithm for selective breeding in aquaculture species. In this study, a systematic comparison of ten GP algorithms, including both traditional and machine-learning algorithms, was conducted using publicly available genotype and phenotype data of eight traits, including weight and disease resistance traits, from five aquaculture species. The study aimed to provide insights into the optimal algorithm for GP in aquatic animals. Notably, no algorithm showed the best performance in all traits. However, reproducing kernel Hilbert space (RKHS) and support-vector machine (SVM) algorithms achieved relatively high prediction accuracies in most of the tested traits. Bayes A and random forest (RF) better prevented noise interference in the phenotypic data compared to the other algorithms. The prediction performances of GP algorithms in the *Crassostrea gigas* dataset were improved by using a genome-wide association study (GWAS) to select subsets of significant SNPs. An R package, "ASGS," which integrates the commonly used traditional and machine-learning algorithms for efficiently finding the optimal algorithm, was developed to assist the application of genomic selection breeding of aquaculture species. This work provides valuable information and a tool for optimizing algorithms for GP, aiding genetic breeding in aquaculture species.

Keywords: genomic prediction; machine learning; aquatic animals; algorithm comparison; genomic selection

1. Introduction

Nowadays, selective breeding has become essential in the aquaculture industry. The growing demand for aquaculture products has led to an increased interest in improving productive and efficiency traits [1–6]. Traditional breeding approaches usually select individuals based only on phenotype, although molecular marker information is also used in some cases [1,4–6]. Breeding organizations around the world have been using the genomic selection (GS) approach in both crops and livestock [2–4]. In many species, GS achieves a higher selection response than pedigree-based selection (PBLUP) and marker-assisted selection (MAS) [5–8]. Genomic prediction (GP) refers to the prediction process of prediction accuracy during GS [9]. The application of GP requires breeding organizations to genotype a large number of marker loci using high-density marker arrays [10]. In recent years, GP has benefited from the advancement of the genotyping method by utilizing next-generation sequencing and SNP array platforms [11].

Several recent GP studies have been conducted in aquatic animals, mainly focusing on productive and efficiency traits, such as growth and disease resistance. GP potentially performs efficiently in aquatic animals because of the large number of offspring produced by a pair of parents and the use of high-density marker panels to genotype animals [12,13].

GP is also considered an ideal method for complex traits, such as growth traits controlled by many genetic loci with minor effects [14,15]. To date, GP-based studies in aquaculture have mainly focused on productive and efficiency traits [12]. Furthermore, GP for disease resistance has been conducted in large yellow croaker *(Larimichthys crocea)* [16], rainbow trout (*Oncorhynchus mykiss*) [17], hybrid red tilapia (*Oreochromis* spp.) [11], gilthead sea bream *(Sparus aurata)* [18], and Nile tilapia (*O. niloticus*) [19]. GP for growth-related traits has been performed in various species, such as Nile tilapia (*O. niloticus*) [20], Pacific white shrimp (*Litopenaeus vannamei*) [21], and Atlantic salmon (*Salmo salar*) [15]. The aforementioned studies proved that there are no best algorithms for all species and traits. Furthermore, studies that combined genome-wide association study (GWAS) with GP showed promising results. Studies has proved that GP can achieve higher prediction accuracy when using top SNPs with a significant *p*-value selected by GWAS [16,22].

Numerous algorithms have been proposed for GP. Each algorithm is usually suitable for one or more scenarios [23]. GP has faced the problem of "large p small n," where the number of SNPs (p) is much larger than that of individuals (n) because of the rapid progress of next-generation sequencing [24,25], a phenomenon implying the data is highly-dimensional [26]. Non-additive effects have also been commonly observed in both plants and animals [26–28]. These lead to the proposition of non-parametric algorithms, including machine-learning methods, that deal with high-dimensionality data [26,29,30]. Machine-learning methods, such as reproducing kernel Hilbert space (RKHS), support-vector machine (SVM), gradient-boosting machine (GBM), random forest (RF), extreme gradient boosting (Xg-Boost), elastic net, adaptive boosting (AdaBoost), and deep learning, have already been used in selective-breeding programs in both crops and livestock [23,24,26,29,31–33]. It is therefore necessary to test the efficiency of these machine-learning methods when used in aquaculture scenarios. However, only a few studies related to the use of machine-learning methods have been performed in specific aquaculture species [34,35]. A systematic comparison of traditional and machine-learning algorithms for GP can thus provide valuable information to identify the optimal algorithm for specific traits in different aquaculture species.

In this study, ten genomic-prediction algorithms, including both traditional methods and machine-learning methods, were used to conduct a systematic comparison of publicly available genotype and phenotype data from five aquaculture species. This study aimed to (1) compare the genomic-prediction accuracy and computational time of traditional and machine-learning methods, (2) evaluate the prediction accuracy of different algorithms upon inclusion of noise from phenotypic data and using top SNPs with significant *p*-value selected by GWAS, and (3) develop an R package embedded with the commonly used traditional and machine-learning algorithms to assist in selecting the optimal algorithm for specific datasets.

2. Materials and Methods

2.1. Datasets

Publicly available datasets from GP studies in five species, including the Pacific oyster (*Crassostrea gigas*) [36], Atlantic salmon (*S. salar*) [37], rainbow trout (*O. mykiss*) [17], Nile tilapia (*O. niloticus*) [20], and common carp (*Cyprinus carpio*) [38], were retrieved from the aforementioned articles for the analyses in this work. The *C. gigas* dataset was generated from a GWAS study aimed at assessing oyster resistance to *Ostreid herpesvirus* (OsHV-1). The *S. salar* dataset was generated from a GP study aimed at assessing genomic-prediction accuracy for salmon resistance to amoebic gill disease. The *O. mykiss* dataset was from a GP study aimed at assessing genomic-prediction accuracy for trout resistance to *Piscirickettsia salmonis*. The *C. carpio* dataset was from a GP study aimed at assessing genomic-prediction accuracy for the resistance of carp to koi herpesvirus disease (KHVD). The *O. niloticus* dataset was from a GP study aimed at assessing genomic-prediction accuracy for growth-related traits, including fillet yield and harvest weight. The missing genotypes of the aforementioned datasets were randomly imputed using the R package "synbreed" (version 0.12-9) [39]. Table 1 outlines the number of SNPs and individuals, and the traits analyzed in

each dataset. Notably, the survival trait is expressed in binary format with 0 and 1 denoting dead and survivor individuals, respectively, whereas the mean gill and amoebic-load trait in the *S. salar* dataset and the weight traits in the *O. mykiss* dataset, *C. carpio* dataset, and *O. niloticus* dataset were continuous traits. Notably, the mean gill refers to how the resistance to amoebic gill disease is defined. It is the mean of the gill-lesion score recorded for both gills. The amoebic load is the qPCR values using *Neoparamoeba perurans* specific primers [37].

Table 1. Datasets used in this study.

Species	Number of Markers	Number of Genotyped Individuals	Trait Analyzed	Data Sources (Accessed on 23 September 2022)
Salmo salar	16,797	1481	Amoebic load, mean gill	https://doi.org/10.1534/g3.118.200075
Oncorhynchus mykiss	26,068	2047	weight Resistance to *P. salmonis*	https://doi.org/10.1534/g3.119.400204
Cyprinus carpio	15,615	1260	Weight Resistance to KHVD	https://doi.org/10.3389/fgene.2019.00543
Oreochromis niloticus	32,306	1125	Weight	https://doi.org/10.1534/g3.119.400116
Crassostrea gigas	23,349	704	Resistance to OsHV-1	https://doi.org/10.1534/g3.118.200113

2.2. Evaluation of Traditional Genomic-Prediction Algorithms

Five traditional genomic-prediction algorithms, including Bayes A (BA), Bayes B (BB), Bayes C (BC), Bayesian ridge regression (BRR), and Bayesian lasso (BL), were fitted to the data, followed by an assessment of their prediction accuracy. The general formula of Bayes A and Bayes B is

$$y_i = \mu + \sum_{j=1}^{h} \beta_j X_{ij} + e \tag{1}$$

where i = 1 . . . for the p individual, j = 1 . . . for the h marker, y_i is the vector of phenotypes to the ith individual, μ stands for the vector of the constant term, β_j is the vector of regression coefficient to be estimated, X_{ij} is the incidence vector to individual i and marker j (0, 1, 2 stands for genotype AA, Aa, and aa), and e represents the random residual vector [2]. Bayes B assumed that only part of the markers contributed to the genetic variance, while Bayes A assumed that the variances of each marker had the same prior distribution. Bayes C belonged to the Bayes Cπ, where parameter π was set to 0.9 [31,40]. The Bayesian ridge-regression algorithm, a Bayesian version of RR-BLUP [3], was proposed by Hsiang [41] and applied in GP by Pérez and de los Campos [42], assuming that all markers have the same genetic variance [3]. The RR-BLUP is fundamentally equivalent to GBLUP [43], meaning that Bayesian ridge regression can represent RR-BLUP and GBLUP. The Bayesian lasso algorithm was used in GP [42,44] by adding a regularization parameter and assuming that the markers follow the Laplace distribution [3]. All five algorithms were implemented in the R package "BGLR" (version 1.1.0) [42]. The iterations were set at 12,000, burn-in at 2000, degree of freedom at 5, and thin at 5. The response-type parameter for binary phenotypes was set as "ordinal." Supplementary Table S1 provides the example codes for the aforementioned traditional algorithms.

2.3. Evaluation of Machine-Learning Algorithms

Five machine-learning algorithms, including artificial neural network (ANN), reproducing kernel Hilbert space (RKHS), support-vector machine (SVM), gradient-boosting machine (GBM), and random forest (RF) were evaluated for their genomic-prediction performance. ANN belongs to neural network algorithms, RKHS and SVM belong to kernel-based algorithms, while GBM and RF belong to decision-tree algorithms [30,45].

ANN was implemented in the R package "brnn" (version 0.8) [46], fitting the simplest two-layer neural network, as described by Foresee and Hagan [47]. The algorithm described

by Nguyen and Widrow [48] was used to assign initial weights, while the Gauss–Newton algorithm was used to perform the optimization. The "neurons" parameter in ANN highlighted the number of neurons in the hidden layer; a larger parameter significantly increased the complexity of the algorithm, thus increasing its computational time [49]. The RKHS algorithm was implemented in the R package "BGLR" (version 1.1.0) [42]. Notably, RKHS is a semi-parametric model replacing the genomic relationship matrix with the general kernel matrix to draw similarities between individuals despite being genetically uncorrelated [23,45]. The "h" parameter in RKHS is responsible for the shape of the probability density function in RKHS. A larger "h" smoothens the probability density function [50]. SVM was implemented in the R package "kernlab" (version 0.9-31) [51]. SVM was originally developed as a classifier and aimed to solve the separation problems of hyperplane having the largest geometric interval that can correctly divide a given training dataset [23,51]. The "epsilon" parameter indicates the tolerance where no penalty is given to constraint violation; a larger "epsilon" allows the existence of more errors. In the same line, the "C" parameter defines the cost of the violation; a larger "C" adds more penalties to the violation [51,52]. The RF algorithm was implemented through the R package "randomForest" (version 4.7-1.1) [53] and adopted a bootstrap resampling method to select a subset of observations to train numbers of decision trees each time. Each decision tree was grown using two-level randomization in the learning process, in which the new data result was decided by the voting score of each tree [26,54]. The "ntree" parameter defines the number of decision trees. However, the tree number should not be too small to achieve a higher prediction accuracy. The "nodesize" parameter defines the size of the trees to be grown. GBM was implemented using the R package "gbm" (version 2.1.8) [55]. It is an improvement on RF because the model in each iteration is updated and the resulting residuals are used to select the next model in a sequential manner [56]. Of note, the sampling of subsets in GBM depends on the weights of previous samples [26]. The "n.trees" parameter is the same as the "ntree" parameter in RF. The "shrinkage" parameter defines the learning rate; smaller learning rates commonly require more trees to be grown. The "interaction.depth" parameter defines the numbers of splits on each tree. A higher "interaction.depth" parameter thus increases the complexity of the GBM algorithm [55,57]. Notably, as the aforementioned hyperparameters significantly affect prediction accuracy, they were tuned to improve the prediction performances [34,58]. Table 2 outlines the hyperparameters tuned for each algorithm. Specifically, Supplementary Table S2 shows the value of hyperparameters set for the five machine-learning algorithms, while Supplementary Table S1 provides the example codes for the machine-learning algorithms.

Table 2. Hyperparameters tuned in each algorithm.

Algorithm	Hyperparameters Tuned
ANN	neurons: the number of neurons.
GBM	Distribution: the distribution of the data used (squared error, absolute loss, and t-distribution loss, among others) n.trees: the total number of trees to fit shrinkage: a shrinkage parameter applied to each tree in the expansion, also known as the learning rate or step-size reduction interaction.depth: the maximum depth of each tree
RF	ntree: number of trees to grow mtry: Number of variables randomly sampled as candidates at each split nodesize: minimum size of terminal nodes
RKHS	h: bandwidth parameter
SVM	Kernel: the kernel function was used in training and predicting (radial basis, polynomial, Laplacian, and hyperbolic, among others) Epsilon: epsilon in the insensitive-loss function C: cost of constraints violation

2.4. Validation Methods

The prediction accuracies of the continuous traits were evaluated using Pearson correlation coefficient between the observed phenotypes and genomic estimated breeding value (GEBV). The area under the curve (AUC) was used for survival traits (0 and 1 denoted dead and survivor individuals). Hold-out cross-validation was applied by randomly selecting 10% of the samples as the testing population while the remaining 90% represented the training population. The cross-validation was repeated 50 times, a correlation was obtained every time and then an average value of correlation was obtained.

2.5. Effect of Noisy Phenotypic Data on Algorithm Prediction Accuracy

pre-set for each hyperparameter and were then subjected to all possible combinations for different hyperparameters. The validation method of the grid-search adopted the one described in Section 2.4 in the Materials and Methods. The combination of hyperparameters that achieved the highest prediction accuracy was used for the subsequent predictions. To evaluate the hyperparameter tuning function in R package "ASGS", the survival trait in *C. gigas* was used to compare the AUC scores of tuning hyperparameters manually and automatically. The comparison was conducted under different significant marker densities.

3. Results

3.1. Prediction Accuracies of the Algorithms

Figure 1 shows the prediction accuracy evaluated by both Pearson correlation coefficient and AUC. The standard deviation is shown in Supplementary Table S1. Notably, eight traits in five species were used to conduct the comparison. Nearly all the algorithms showed higher prediction accuracy in one or more traits. Bayes A performed well in the weight trait of *O. niloticus* (0.520) and the resistance to KHVD trait of *C. carpio* (0.734). Bayes B performed well in the weight trait of *O. niloticus* (0.506), resistance to KHVD trait of *C. carpio* (0.731), and resistance to *P. salmonis* trait of *O. mykiss* (0.810). Both the Bayesian lasso and Bayesian ridge regression performed well in the resistance to *P. salmonis* trait of *O. mykiss* at 0.807 and 0.811, respectively. Moreover, GBM performed well in the weight trait of *C. carpio* (0.395), while RF performed well in the mean gill trait of *S. salar* (0.332). Of note, RKHS performed well in the weight trait of *C. carpio* (0.392), the weight trait of *O. mykiss* (0.483), the amoebic-load trait of *S. salar* (0.356), mean gill trait of *S. salar* (0.319), and resistance to KHVD trait of *C. carpio* (0.732). SVM performed well in the weight traits of *C. carpio* (0.406) and *O. mykiss* (0.489). Notably, RKHS and SVM achieved relatively high prediction accuracy in 5 out of 8 traits. Moreover, there was a significant difference between the poorly and well-performed algorithms in each trait. For instance, the highest prediction accuracy of the survival trait in *C. carpio* was 33% higher than the lowest prediction accuracy, while that of the amoebic-load trait in *S. salar* was 8%. Notably, all algorithms achieved AUC scores near 0.5 in the *C. gigas* dataset.
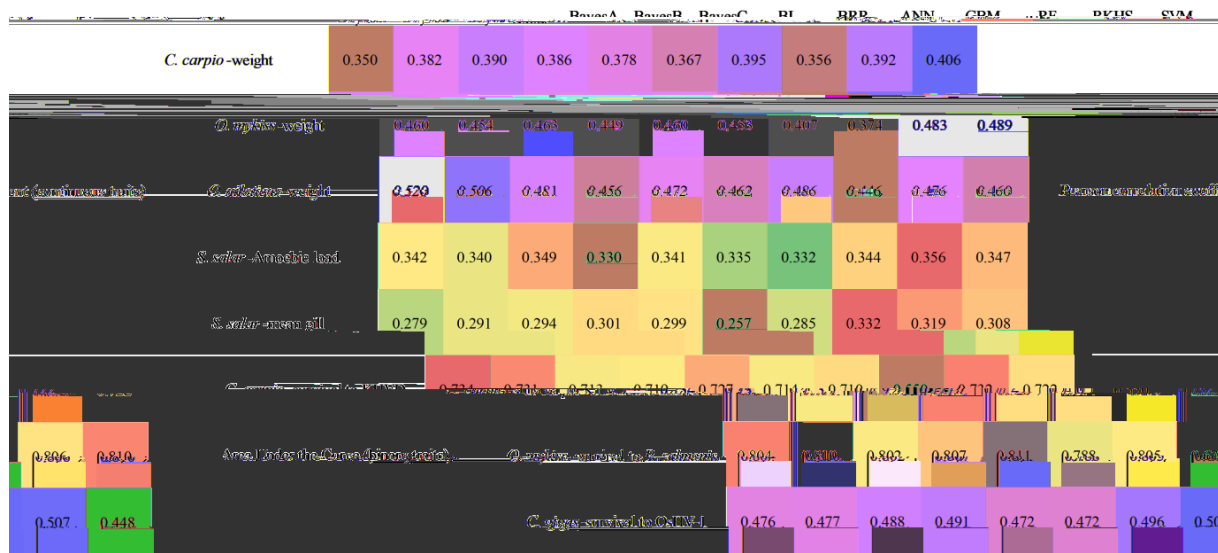

Figure 1. Prediction accuracies of the ten genomic-prediction algorithms analyzing various traits in five aquaculture species. The Pearson correlation coefficient and area under the curve (AUC) were used to evaluate the prediction accuracy of algorithms in each of the traits. Green denotes the lowest prediction accuracy in each trait, while red denotes the highest prediction accuracy.

3.2. Prediction Accuracy of Algorithms Affected by Noise in Phenotypic Data

All algorithms exhibited a declining trend in accuracy as the amount of noisy data increased (Figure 2). Well-performing algorithms significantly changed upon adding a

different amount of noise to the phenotype data. Notably, Bayes A and RF performed better in more situations when noise was added. Bayes A outperformed the other algorithms when 10%, 40%, and 50% of noise were added. Similarly, RF outperformed the other algorithms when 30%, 40%, 60%, and 70% of noise were added.
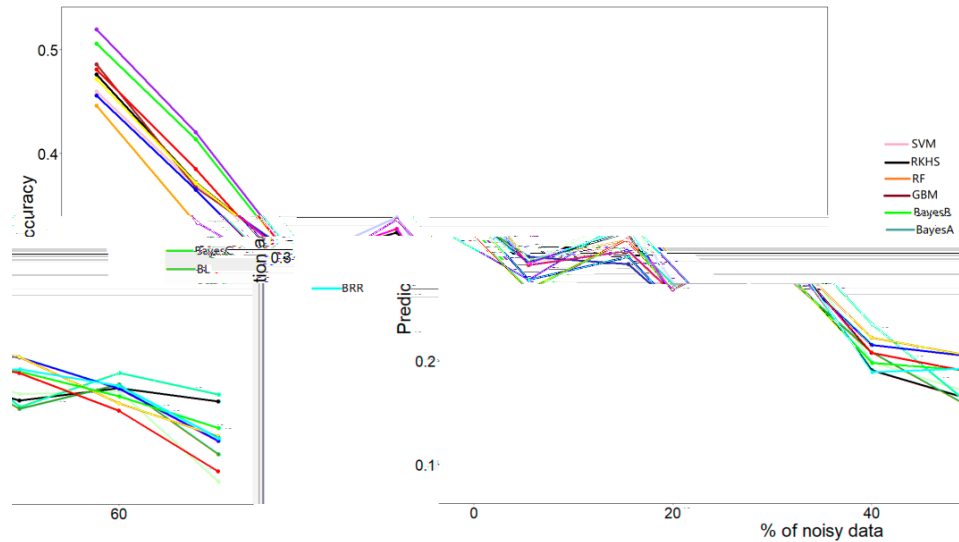
Figure 2. Prediction accuracy of each algorithm when noisy data is added to the phenotype. The *O. niloticus* weight-trait dataset was used. The percentage of noisy data included in the phenotype data and the prediction efficiency evaluated by Pearson correlation coefficient are represented in the X and Y axes, respectively.

### 3.3. Prediction Efficiency of Algorithms in When Using SNPs Selected by GWAS

The results are shown in Figure 3. Notably, all algorithms achieved higher AUC scores when using top SNPs compared to the case when using all SNPs. The highest AUC scores for most algorithms were achieved based on 4500 SNPs. All the algorithms achieved low AUC scores when using all SNPs in the *C. gigas* dataset.
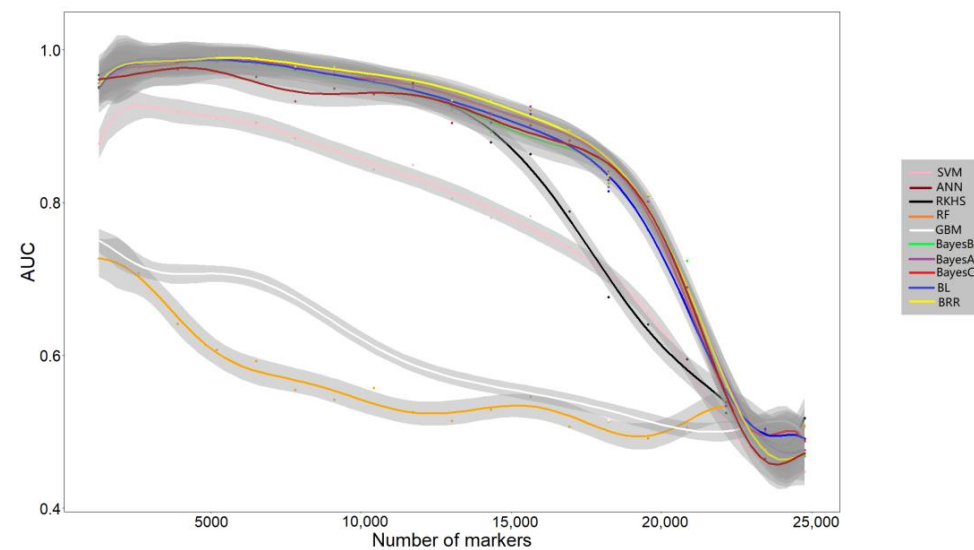
Figure 3. Prediction efficiency of each algorithm when using GWAS to select subsets of significant SNPs. The dataset of *C. gigas* was used. The number of markers used and the prediction efficiency evaluated by AUC (area under the curve) are represented in the X and Y axes, respectively. The grey area indicates 90% confidence interval.

3.4. The Performance of Algorithms in Computational Time

Figure 4 shows the required computational time of 50 times cross-validation of each algorithm. The hyperparameters set for each machine-learning algorithm are shown in Supplementary Table S2. Notably, the computational time of different algorithms varied significantly. RF and GBM, both belonging to the decision-tree algorithm, were highly time-consuming, especially when the size of forest trees was large. ANN and SVM performed better than the time-consuming traditional Bayes algorithms and had comparable or even better prediction accuracies than the traditional algorithms in some traits. RKHS and SVM, the two optimal algorithms (Figure 1), varied significantly in computational time. The computational time of RKHS (approximately 35.5 h) was twice that of SVM (approximately 15.2 h).
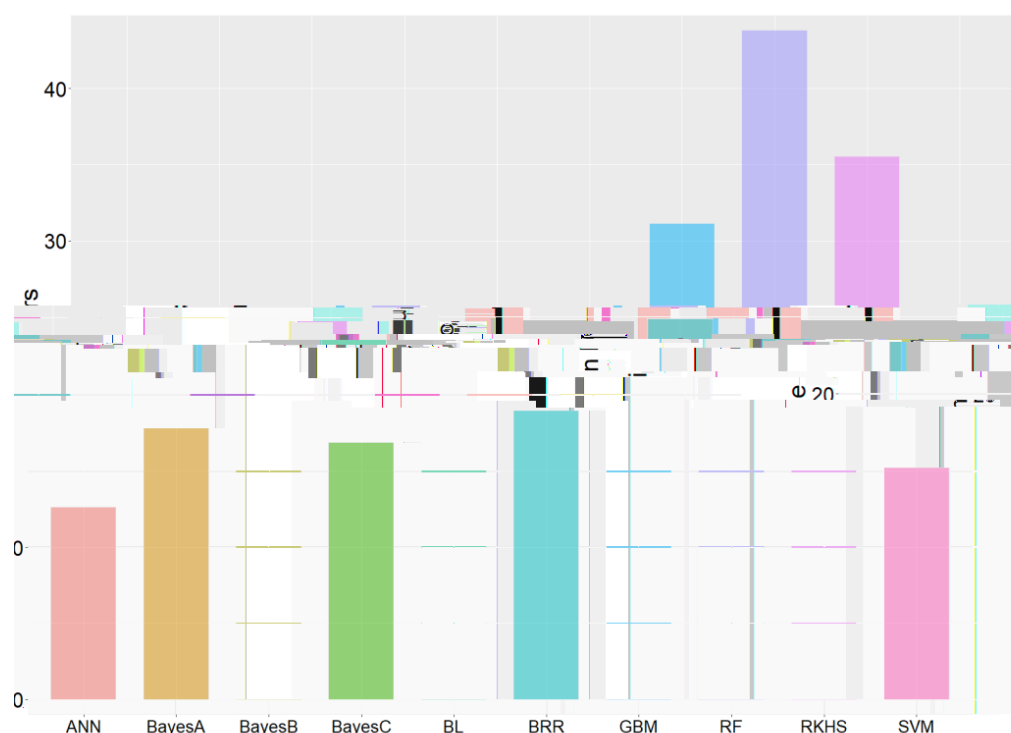
Figure 4. Comparison of the required computational time for fitting each of the ten algorithms. The *O. mykiss* survival trait dataset was used for this analysis.

3.5. Development of R Package "ASGS"

The R package "ASGS" was developed to assist in the efficient use of the GP algorithms. The input data comprised the genotypic, phenotypic, and fraction testing data. The 50 times cross-validation was implemented to obtain the mean prediction accuracy. The output data comprised a model-fitting situation and prediction accuracy for both the training and testing populations. Auto hyperparameter adjusting through the grid-search method was implemented for machine-learning algorithms to achieve a higher prediction accuracy. The R package "ASGS" can overcome the complex hyperparameter adjusting problem in the machine-learning algorithms. Figure 5 shows the comparison of five machine-learning algorithms when tuning the hyperparameters manually and using R package "ASGS". Notably, the auto-hyperparameter-adjusting function achieved similar or even higher (in ANN and GBM) AUC scores compared to tuning hyperparameters manually. Users of the "ASGS" package in R will not have to adjust the hyperparameters independently. The "ASGS" package in R includes built-in example data using the *O. mykiss* dataset, as described earlier, to provide examples for users on how to run it. The package has been released to GitHub (https://github.com/Kuiqin/ASGS, accessed on 23 September 2022).

Figure 5. A comparison of AUC (area under the curve) scores of manual hyperparameter adjustment and automatic hyperparameter adjustment of five machine-learning algorithms. The *C. gigas* survival trait dataset was used for this analysis.

4. Discussion

The extensive use of genomic selection in both livestock and crops has led to a series of algorithms, though none of the algorithms suit all the species and traits under study [58]. Studies that have applied GP algorithms in aquaculture species have mainly focused on comparison of traditional algorithms [62,63]. To date, no studies have compared the machine-learning algorithms with traditional algorithms in a variety of aquatic animals. A systematic assessment of available GP algorithms in aquatic animals is much needed. In this study, we retrieved publicly available genotype and phenotype data from five aquatic species to systematically compare ten GP algorithms, including both Bayesian and machine-learning methods. Notably, each algorithm outperformed others in one or more traits. Studies that have applied GP algorithms in livestock and crops also suggest that there is no best algorithm for every trait in all species [3,29,31,64,65].

Herein, RKHS and SVM achieved relatively high prediction accuracy in five out of eight traits, suggesting that RKHS and SVM are potentially optimal algorithms for more traits. Notably, SVM required much less computational time than RKHS. These findings collectively suggest that breeding organizations should try using SVM for less-well-known traits because of its relatively high prediction accuracy and reduced computational time.

However, the prediction accuracy of any algorithm is affected by many factors, for example, the population structure and the degree of non-additive effects [31]. Breeding organizations should thus choose specific algorithms according to the characteristics of the data because each algorithm fits specific types of data structures. For example, Bayesian methods are suitable for traits more affected by dominant effects, while machine-learning methods are suitable for traits more affected by non-dominant effects [3,4,14,29,31,65].

Notably, all algorithms achieved low AUC scores (near 0.5) when using all SNPs in the *C. gigas* dataset. This phenomenon may be attributed to the small individual number and the existence of a large number of SNPs with no effects or negative effects. Therefore, using all SNPs to conduct GP is not suitable for the *C. gigas* dataset. Herein, we used the top SNPs with significant *p*-values selected by GWAS to conduct GP in the *C. gigas* dataset. The highest AUC scores for most algorithms were achieved based on 4500 SNPs. By reducing SNPs with no effects with negative effects, previous studies also proved that the prediction accuracy can be improved when using the top SNPs selected by GWAS [16,22].

Of note, the adjustment of the hyperparameters in machine-learning methods significantly affects the prediction accuracy and computational time (Materials and Methods 2.3 and the study conducted by Azodi [58]), while the traditional methods are unaffected. This phenomenon is attributed to the ability of the hyperparameters to affect the complexity of machine-learning algorithms. In this study, the "neurons" parameter in ANN, "ntree" in RF, and "n.trees" in GBM were positively correlated with the computational time [46,53–55,57]. For instance, the computational time for ANN lessened when the "neurons" parameter in ANN was set to 1. However, the computational demands of ANN grew tremendously when a larger "neurons" parameter was set. RF requires substantially less computational time when using Python library scikit-learn [34]. In contrast, RF proved to be highly time-consuming when using the R package "randomForest". This phenomenon may be attributed to different RF implementation processes by the two packages. Moreover, RF appeared to be more time-consuming because the number of trees set in RF (1000) was larger than those in GBM (500). Notably, there is a certain degree of subjectivity in the argument of computational time because the number of iterations of the Markov chain Monte Carlo (MCMC) in the Bayesian algorithms varies depending on specific situations [34]. Machine-learning algorithms are promising in reducing computational time compared to the traditional Bayesian algorithms, especially when parallelization is used.

Overfitting and underfitting are the two main causes of low prediction accuracy of a model. Overfitting is caused by the inclusion of too much useless information in the models, insufficient size of the training dataset, or the high complexity of the model. Underfitting is caused by a simplified and hence poorly fitted model to predict the multi-dimensional data sufficiently [66–68]. The complexity of machine-learning algorithms is also affected by their hyperparameters. For example, the "neurons" parameter in ANN positively correlates with the algorithm complexity [46]. To obtain a higher prediction accuracy, it is best to try to reduce the noise in the dataset and choose the most suitable algorithm according to the complexity of the training dataset. Breeding organizations should also choose different algorithms according to the data quality. For instance, decision-tree algorithms can achieve higher prediction accuracy in datasets that contain much noise. Notably, ANN could not function when noise was included, possibly because the dataset used was too small for training ANN when noise was included.

Numerous factors, including the number of individuals and SNPs, the degree of dominant effects and additive effects [4,14], and the existence of epistasis and environmental effects [69,70], can influence the prediction accuracy of an algorithm. Both traditional and machine-learning algorithms have their advantages in overcoming these situations. For instance, traditional algorithms are more likely to achieve a higher prediction accuracy when a trait is more affected by dominant effects [4,14]. Similarly, machine-learning algorithms tend to achieve a higher prediction accuracy in cases when epistasis and environmental effects are considered [69,70]. This study mainly focused on comparing GP algorithms in aquaculture species, without exploring the detailed characteristics of the various algorithms

(for example, how the hyperparameters affect the structure and prediction accuracies of machine-learning algorithms). Future studies should therefore aim to explore more detailed characteristics of the algorithms, and benchmark the GP algorithms in cases of the existence of non-addictive effects.

Notably, the adjustment of hyperparameters in machine-learning algorithms is quite time-consuming [58]. It has been shown that by using "ASGS" package in R, five machine-learning algorithms can achieve similar prediction accuracies compared to tuning the hyperparameters manually (Figure 5). Of note, there is a grid-search function in Python library scikit-learn [71]. However, there is no grid-search function in R. In order to assist the efficient use of the machine-learning algorithms in R, the grid-search function was compiled. The R package "ASGS" has combined the grid-search function with cross validation. When using the "ASGS" package in R, the hyperparameters will not have to be adjusted independently. Moreover, the "ASGS" package in R is user-friendly. Users will not have to write the codes because they have been encapsulated within the package.

In conclusion, none of the ten representative GP algorithms used to analyze datasets of eight traits in five species achieved the highest prediction accuracy in all the traits. However, RKHS and SVM achieved relatively high prediction accuracy in five out of eight traits, suggesting their suitability as the optimal algorithms for more traits. Notably, SVM required much less computational time than RKHS. The prediction accuracy of each algorithm was affected by the inclusion of noise in the phenotypic data. Bayes A and RF were the better algorithms when noise was included. The prediction accuracies of GP algorithms in *C. gigas* dataset were optimized by using GWAS to select subsets of significant SNPs. Furthermore, ANN and SVM outperformed the time-consuming traditional Bayes algorithms. The decision-tree algorithms and RKHS were proved to be highly time-consuming. The relationship between the complexity of the algorithms and data influences the prediction accuracy. Adapting complex algorithms to rather simple data causes overfitting, while adapting the simple algorithm to complex data causes underfitting. This work provides valuable information on the prediction efficiencies of the currently available GP algorithms and a useful tool for assisting in choosing the optimal algorithm for selective breeding of aquaculture species.

References

1. Liu, Y.; Lu, S.; Liu, F.; Shao, C.; Zhou, Q.; Wang, N.; Li, Y.; Yang, Y.; Zhang, Y.; Sun, H.; et al. Genomic Selection Using BayesCπ and GBLUP for Resistance Against *Edwardsiella tarda* in Japanese Flounder (*Paralichthys olivaceus*). *Mar. Biotechnol.* **2018**, *20*, 559–565. [CrossRef] [PubMed]
2. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
3. Mota, R.R.; Silva, F.F.e.; Guimarães, S.E.F.; Hayes, B.; Fortes, M.R.S.; Kelly, M.J.; Guimarães, J.D.; Penitente-Filho, J.M.; Ventura, H.T.; Moore, S. Benchmarking Bayesian genome enabled-prediction models for age at first calving in Nellore cows. *Livest. Sci.* **2018**, *211*, 75–79. [CrossRef]
4. Xu, Y.; Liu, X.; Fu, J.; Wang, H.; Wang, J.; Huang, C.; Prasanna, B.M.; Olsen, M.S.; Wang, G.; Zhang, A. Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Commun.* **2020**, *1*, 100005. [CrossRef]
5. Heffner, E.L.; Lorenz, A.J.; Jannink, J.-L.; Sorrells, M.E. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Sci.* **2010**, *50*, 1681–1690. [CrossRef]
6. Bernardo, R.; Yu, J. Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Sci.* **2007**, *47*, 1082–1090. [CrossRef]
7. Albrecht, T.; Wimmer, V.; Auinger, H.-J.; Erbe, M.; Knaak, C.; Ouzunova, M.; Simianer, H.; Schön, C.-C. Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* **2011**, *123*, 339. [CrossRef]
8. 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 Genet.* **2015**, *5*, 615–627. [CrossRef]
9. Crossa, J.; Pérez, P.; Hickey, J.; Burgueño, J.; Ornella, L.; Cerón-Rojas, J.; Zhang, X.; Dreisigacker, S.; Babu, R.; Li, Y.; et al. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* **2014**, *112*, 48–60. [CrossRef]
10. Wang, X.; Xu, Y.; Hu, Z.; Xu, C. Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* **2018**, *6*, 330–340. [CrossRef]
11. Sukhavachana, S.; Tongyoo, P.; Massault, C.; McMillan, N.; Leungnaruemitchai, A.; Poompuang, S. Genome-wide association study and genomic prediction for resistance against *Streptococcus agalactiae* in hybrid red tilapia (*Oreochromis* spp.). *Aquaculture* **2020**, *525*, 735297. [CrossRef]
12. Liu, G.; Dong, L.; Gu, L.; Han, Z.; Zhang, W.; Fang, M.; Wang, Z. Evaluation of Genomic Selection for Seven Economic Traits in Yellow Drum (*Nibea albiflora*). *Mar. Biotechnol.* **2019**, *21*, 806–812. [CrossRef] [PubMed]
13. Yue, G.H. Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish Fish.* **2014**, *15*, 376–396. [CrossRef]
14. Yang, J.; Mezmouk, S.; Baumgarten, A.; Buckler, E.S.; Guill, K.E.; McMullen, M.D.; Mumm, R.H.; Ross-Ibarra, J. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genet.* **2017**, *13*, e1007019. [CrossRef] [PubMed]
15. Tsai, H.-Y.; Hamilton, A.; Tinch, A.E.; Guy, D.R.; Gharbi, K.; Stear, M.J.; Matika, O.; Bishop, S.C.; Houston, R.D. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genom.* **2015**, *16*, 969. [CrossRef]
16. Zhao, J.; Bai, H.; Ke, Q.; Li, B.; Zhou, Z.; Wang, H.; Chen, B.; Pu, F.; Zhou, T.; Xu, P. Genomic selection for parasitic ciliate *Cryptocaryon irritans* resistance in large yellow croaker. *Aquaculture* **2021**, *531*, 735786. [CrossRef]
17. Barria, A.; Marín-Nahuelpi, R.; Cáceres, P.; López, M.E.; Bassini, L.N.; Lhorente, J.P.; Yáñez, J.M. Single-Step Genome-Wide Association Study for Resistance to *Piscirickettsia salmonis* in Rainbow Trout (*Oncorhynchus mykiss*). *G3 Genes Genomes Genet.* **2019**, *9*, 3833–3841. [CrossRef]
18. Palaiokostas, C.; Ferraresso, S.; Franch, R.; Houston, R.D.; Bargelloni, L. Genomic Prediction of Resistance to Pasteurellosis in Gilthead Sea Bream (*Sparus aurata*) Using 2b-RAD Sequencing. *G3 Genes Genomes Genet.* **2016**, *6*, 3693–3700. [CrossRef]
19. Lu, S.; Zhu, J.; Du, X.; Sun, S.; Meng, L.; Liu, S.; Fan, G.; Wang, J.; Chen, S. Genomic selection for resistance to *Streptococcus agalactiae* in GIFT strain of *Oreochromis niloticus* by GBLUP, wGBLUP, and BayesCπ. *Aquaculture* **2020**, *523*, 735212. [CrossRef]
20. Yoshida, G.M.; Lhorente, J.P.; Correa, K.; Soto, J.; Salas, D.; Yáñez, J.M. Genome-Wide Association Study and Cost-Efficient Genomic Predictions for Growth and Fillet Yield in Nile Tilapia (*Oreochromis niloticus*). *G3 Genes Genomes Genet.* **2019**, *9*, 2597–2607. [CrossRef]
21. Wang, Q.; Yu, Y.; Yuan, J.; Zhang, X.; Huang, H.; Li, F.; Xiang, J. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. *BMC Genet.* **2017**, *18*, 45. [CrossRef] [PubMed]
22. Luo, Z.; Yu, Y.; Xiang, J.; Li, F. Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. *Aquaculture* **2021**, *539*, 736620. [CrossRef]
23. Howard, R.; Carriquiry, A.L.; Beavis, W.D. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. *G3 Genes Genomes Genet.* **2014**, *4*, 1027–1046. [CrossRef] [PubMed]
24. Waldmann, P. Genome-wide prediction using Bayesian additive regression trees. *Genet. Sel. Evol.* **2016**, *48*, 42. [CrossRef] [PubMed]
25. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **2013**, *9*, e1003264. [CrossRef]
26. Li, B.; Zhang, N.; Wang, Y.-G.; George, A.W.; Reverter, A.; Li, Y. Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. *Front. Genet.* **2018**, *9*, 237. [CrossRef]

27. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99*, 323–329. [CrossRef]
28. Goddard, M. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* **2009**, *136*, 245–257. [CrossRef]
29. Roorkiwal, M.; Rathore, A.; Das, R.R.; Singh, M.K.; Jain, A.; Srinivasan, S.; Gaur, P.M.; Chellapilla, B.; Tripathi, S.; Li, Y.; et al. Genome-Enabled Prediction Models for Yield Related Traits in Chickpea. *Front. Plant Sci.* **2016**, *7*, 1666. [CrossRef]
30. Nayeri, S.; Sargolzaei, M.; Tulpan, D. A review of traditional and machine learning methods applied to animal breeding. *Anim. Health Res. Rev.* **2019**, *20*, 31–46. [CrossRef]
31. Neves, H.H.R.; Carvalheiro, R.; Queiroz, S.A. A comparison of statistical methods for genomic selection in a mice population. *BMC Genet.* **2012**, *13*, 100. [CrossRef] [PubMed]
32. Grinberg, N.F.; Orhobor, O.I.; King, R.D. An evaluation of machine-learning for predicting phenotype: Studies in yeast, rice, and wheat. *Mach. Learn.* **2020**, *109*, 251–277. [CrossRef] [PubMed]
33. Ma, W.; Qiu, Z.; Song, J.; Li, J.; Cheng, Q.; Zhai, J.; Ma, C. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **2018**, *248*, 1307–1318. [CrossRef] [PubMed]
34. Palaiokostas, C. Predicting for disease resistance in aquaculture species using machine learning models. *Aquac. Rep.* **2021**, *20*, 100660. [CrossRef]
35. Bargelloni, L.; Tassiello, O.; Babbucci, M.; Ferraresso, S.; Franch, R.; Montanucci, L.; Carnier, P. Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream. *Aquac. Rep.* **2021**, *20*, 100661. [CrossRef]
36. Gutierrez, A.P.; Bean, T.P.; Hooper, C.; Stenton, C.A.; Sanders, M.B.; Paley, R.K.; Rastas, P.; Bryrom, M.; Matika, O.; Houston, R.D. A Genome-Wide Association Study for Host Resistance to Ostreid Herpesvirus in Pacific Oysters (*Crassostrea gigas*). *G3 Genes Genomes Genet.* **2018**, *8*, 1273–1280. [CrossRef] [PubMed]
37. Robledo, D.; Matika, O.; Hamilton, A.; Houston, R.D. Genome-Wide Association and Genomic Selection for Resistance to Amoebic Gill Disease in Atlantic Salmon. *G3 Genes Genomes Genet.* **2018**, *8*, 1195–1203. [CrossRef]
38. Palaiokostas, C.; Vesely, T.; Kocour, M.; Prchal, M.; Pokorova, D.; Piackova, V.; Pojezdal, L.; Houston, R.D. Optimizing Genomic Prediction of Host Resistance to Koi Herpesvirus Disease in Carp. *Front. Genet.* **2019**, *10*, 543. [CrossRef] [PubMed]
39. Wimmer, V.; Albrecht, T.; Auinger, H.-J.; Schön, C.-C. Synbreed: A framework for the analysis of genomic prediction data using R. *Bioinformatics* **2012**, *28*, 2086–2087. [CrossRef]
40. Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinform.* **2011**, *12*, 186. [CrossRef]
41. Hsiang, T.C. A Bayesian View on Ridge Regression. *J. R. Stat. Soc. Ser. D Stat.* **1975**, *24*, 267–268. [CrossRef]
42. Pérez, P.; de los Campos, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **2014**, *198*, 483–495. [CrossRef] [PubMed]
43. Habier, D.; Fernando, R.L.; Dekkers, J.C.M. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* **2007**, *177*, 2389–2397. [CrossRef] [PubMed]
44. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
45. González-Recio, O.; Rosa, G.J.M.; Gianola, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* **2014**, *166*, 217–231. [CrossRef]
46. Rodriguez, P.P.; Gianola, D. brnn: Bayesian Regularization for Feed-Forward Neural Networks. 2020. Available online: https://CRAN.R-project.org/package=brnn (accessed on 23 September 2022).
47. Foresee, F.D.; Hagan, M.T. Gauss-Newton approximation to Bayesian learning. In Proceedings of the International Conference on Neural Networks (ICNN'97), Houston, TX, USA, 12 June 1997; pp. 1930–1935.
48. Nguyen, D.; Widrow, B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In Proceedings of the 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 17–21 June 1990; pp. 21–26.
49. Koumakis, L. Deep learning models in genomics; are we there yet? *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1466–1473. [CrossRef] [PubMed]
50. Jones, M.C.; Marron, J.S.; Sheather, S.J. A Brief Survey of Bandwidth Selection for Density Estimation. *J. Am. Stat. Assoc.* **1996**, *91*, 401–407. [CrossRef]
51. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. Kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
52. Chih-Wei, H.; Chih-Jen, L. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [CrossRef]
53. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
54. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
55. Greenwell, B.; Boehmke, B.; Cunningham, J.; Developers, G. Gbm: Generalized Boosted Regression Models. Available online: https://CRAN.R-project.org/package=gbm (accessed on 23 September 2022).
56. John Lu, Z.Q. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2010**, *173*, 693–694. [CrossRef]
57. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]

58. Azodi, C.B.; Bolger, E.; McCarren, A.; Roantree, M.; de los Campos, G.; Shiu, S.-H. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 Genes Genomes Genet.* **2019**, *9*, 3691–3702. [CrossRef] [PubMed]

59. Tang, Y.; Liu, X.; Wang, J.; Li, M.; Wang, Q.; Tian, F.; Su, Z.; Pan, Y.; Liu, D.; Lipka, A.E.; et al. GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant Genome* **2016**, *9*, 1–9. [CrossRef] [PubMed]

60. Huang, M.; Liu, X.; Zhou, Y.; Summers, R.M.; Zhang, Z. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience* **2019**, *8*, giy154. [CrossRef]

61. Yang, B.; Zhai, S.; Zhang, F.; Wang, H.; Ren, L.; Li, Y.; Li, Q.; Liu, S. Genome-wide association study toward efficient selection breeding of resistance to *Vibrio alginolyticus* in Pacific oyster, *Crassostrea gigas*. *Aquaculture* **2022**, *548*, 737592. [CrossRef]

62. Song, H.; Dong, T.; Yan, X.; Wang, W.; Tian, Z.; Sun, A.; Ying, D.; Zhu, H.; Hu, H. Genomic selection and its research progress in aquaculture breeding. *Rev. Aquac.* **2022**, *in press*. [CrossRef]

63. Song, H.; Hu, H. Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species. *Evol. Appl.* **2022**, *15*, 578–590. [CrossRef]

64. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]

65. Pérez-Rodríguez, P.; Gianola, D.; González-Camacho, J.M.; Crossa, J.; Manès, Y.; Dreisigacker, S. Comparison between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. *G3 Genes Genomes Genet.* **2012**, *2*, 1595–1605. [CrossRef] [PubMed]

66. Andrews, J.L. Addressing overfitting and underfitting in Gaussian model-based clustering. *Comput. Stat. Data Anal.* **2018**, *127*, 160–171. [CrossRef]

67. Yu, H.; Wu, Y.; Niu, L.; Chai, Y.; Feng, Q.; Wang, W.; Liang, T. A method to avoid spatial overfitting in estimation of grassland above-ground biomass on the Tibetan Plateau. *Ecol. Indic.* **2021**, *125*, 107450. [CrossRef]

68. Ord, K. Data adjustments, overfitting and representativeness. *Int. J. Forecast.* **2020**, *36*, 195–196. [CrossRef]

69. Ali, M.; Zhang, L.; DeLacy, I.; Arief, V.; Dieters, M.; Pfeiffer, W.H.; Wang, J.; Li, H. Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *Crop J.* **2020**, *8*, 866–877. [CrossRef]

70. Millet, E.J.; Kruijer, W.; Coupel-Ledru, A.; Alvarez Prado, S.; Cabrera-s; rqu68(E.J.;)L3(Alvar)09(ube(A.;)-3S3(Coupel-hch)-251)]T68(