Aquaculture Reports 20 (2021) 100660

Contents lists available at ScienceDirect

Aquaculture Reports

journal homepage: www.elsevier.com/locate/aqrep

Predicting for disease resistance in aquaculture species using machine learning models

Christos Palaiokostas

Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

ARTICLE INFO

Keywords:
Machine learning
Aquaculture
Selective breeding
Disease resistance

ABSTRACT

Predicting disease resistance is one of the most prominent applications of aquaculture selective breeding. Reductions in genotyping costs have allowed the implementation of genomic selection in an abundance of aquaculture species and their related diseases showing promising results. Machine learning (ML) models can be of value for prediction purposes, as suggested by several studies in both plants and livestock. The current study aimed to test the efficiency of various ML models in predicting disease resistance using both simulated and real datasets. More specifically, models like decision trees (DT), support vector machines (SVM), random forests (RF), adaptive boosting (Adaboost) and extreme gradient boosting (XGB) were benchmarked against genomic best linear unbiased prediction for threshold traits backend by Markov chain Monte Carlo (GBLUP-MCMC) both in terms of prediction efficiency and required computational time. Moreover, the model ranking was tested in datasets where the ratio between the two observed phenotypes (resistant vs non-resistant) was unbalanced. Across all tested datasets, XGB ranked first with a slight advantage over GBLUP-MCMC, ranging between 1–4 %. SVM and RF delivered predictions in tight proximity with the ones from XGB and GBLUP-MCMC. In addition, predictions 3–4 % lower compared to GBLUP-MCMC were obtained with Adaboost. On the other hand, the predictions from DT were consistently low (~40 % lower compared to GBLUP-MCMC). All tested ML models had significantly reduced computational requirements than GBLUP-MCMC. In the case of XGB, the computational requirements were reduced more than 20-fold as opposed to GBLUP-MCMC under the settings of the current study. RF delivered both competitive predictions and was highly efficient in terms of the required computational time (~3 min). Overall, the results of the current study suggest that ML models can be valuable tools in aquaculture breeding studies for disease resistance.

1. Introduction

Advancements in sequencing technologies over the last decade have transformed the field of aquaculture breeding and genomics (You et al., 2020). Aquaculture selective breeding has transitioned to the genomics era, at least in the case of major farmed species like Atlantic salmon and Nile tilapia (Houston et al., 2020; Yáñez et al., 2020). Most importantly, it is not uncommon nowadays for selection decisions in aquaculture breeding programs to be guided by genomic information derived either through the usage of single nucleotide polymorphisms (SNPs) arrays (Lhorente et al., 2019) or genotyping by sequencing (GBS) platforms (Barbanti et al., 2020; Robledo et al., 2017). Furthermore, a plethora of research studies in the last five years has demonstrated the value of genomic selection (GS) practices (Meuwissen et al., 2001) in a wide range of aquaculture species, including, amongst others, salmonids, tilapias, carps, bass and oysters (Barría et al., 2018; Besson et al., 2019; Faggion et al., 2019; Gutierrez et al., 2020; Horn et al., 2020; Joshi et al., 2020; Vallejo et al., 2019).

Current knowledge suggests that genomic information is particularly valuable in studying traits related to disease resistance as disease outbreaks in farmed fish tend to be devastating both in economic and welfare aspects (Asche et al., 2009). Since there is a lack of efficient therapeutic agents for various commonly encountered diseases in aquaculture, selective breeding practices can offer solutions (Yáñez et al., 2014). Not surprisingly, the implementation of genomics in the study of disease resistance has a prominent role in aquaculture selective breeding as clearly shown by several recent studies (Aslam et al., 2018, 2020; Boison et al., 2019; Gonen et al., 2015; Robledo et al., 2018).

GS practices are usually considered to be the preferred route of action as resistance to diseases usually resembles a polygenic trait (Houston, 2017). Most common applications of GS usually involve the usage of algorithms that are based on genomic best linear unbiased

E-mail address: christos.palaiokostas@slu.se.

https://doi.org/10.1016/j.aqrep.2021.100660
Received 11 December 2020; Received in revised form 19 February 2021; Accepted 3 March 2021
Available online 13 March 2021

predictor (GBLUP) or its variants like single-step approaches (Lourenco et al., 2020; Misztal et al., 2020) and on Bayesian linear regressions (Gianola, 2013). Equally important, the availability of state-of-the-art software like BLUPF90 (Misztal et al., 2018) or R/BGLR (Pérez and de los Campos, 2014) that are freely distributed allowed the implementation of GS in various aquaculture species and settings where experience so far suggests that GBLUP is a robust choice (Correa et al., 2017; Garcia et al., 2018).

With a few exceptions (Vallejo et al., 2016; Vallejo et al., 2019), the vast majority of published studies to date have assessed the prediction efficiency of GS models for disease resistance based on data of a single generation. The above is mainly due to two reasons: 1. disease challenge experiments have high-costs requirements, and 2. aquaculture breeding programs up to date are relatively new compared to their livestock counterparts, and in many cases, genomic information beyond a single generation is not available. Therefore most of the studies aiming to pick the best performing model for predicting disease resistance have used cross-validation strategies on animals from the same generation to train GS models and minimize the chances of overfitting. However, the aforementioned does not necessarily provide information regarding the model that best predicts future performance, which is the overall aim of selective breeding. In contrast, in equivalent situations in livestock, it is common to train the GS models on multi-generational datasets and perform the validation in the latest generation(s) (Lourenco et al., 2015). It is expected that in the coming years, we will witness more aquaculture-oriented studies on disease resistance using multi-generational datasets.

Even though disease resistance traits can be interpreted and treated in a wide range of manners using well-documented approaches from the field of epidemiology (Saura et al., 2019), a most common approach is where disease resistance is regarded as a binary trait. In such situations, the objective of the tested model is to efficiently classify the animals of each category (resistant vs non-resistant) based on the available genomic information. As a matter of fact, the above is the most common interpretation in aquaculture disease resistance studies. However, limited attention has been placed in the scenario where the phenotypic distribution among resistant and non-resistant animals is skewed towards one or the other category. It should be stressed that in the case of binary traits, some of the most popular breeding software like the BLUPF90 suite or R/BGLR rely on Markov chain Monte Carlo (MCMC). In general, algorithms based on MCMC are computationally demanding and non-prone to parallelization, which can prove to be a significant issue as genotypic datasets continue to increase in size.

Machine learning (ML) tools have been recently in the spotlight, finding applications in numerous real-life situations (Wilmott, 2019). ML algorithms are routinely applied, amongst others, in a wide range of regression and classification problems in practically all sorts of scientific disciplines, with one of their most highlighted application being the study-prediction of human diseases (Myszczynska et al., 2020). In the field of animal breeding, ML algorithms have also been gaining momentum finding applications in a wide range of prediction tasks (Pérez-Enciso, 2017). Even though no single model, whether based on ML or more affiliated with traditional animal breeding, seems to provide optimal predictions for all traits of interest and breeding schemes, ML appears to have a role in the animal breeder's toolbox. Experience gained from both simulation and real data studies suggests that ML models can produce competitive predictions to classical animal breeding models like GBLUP-MCMC (Nayeri et al., 2019). Besides, it should be noted that ML models compared to commonly used animal breeding models usually shine in scenarios where interactions influencing the phenotype of interest exist amongst the model predictors (Howard et al., 2014).

In the current study, ML algorithms were assessed in terms of their efficiency to predict disease resistance in both simulated and real-life aquaculture datasets. More specifically, the prediction efficiency of Decision Trees (DT), Support Vector Machines (SVM), Random Forests (RF) and boosting based approaches like AdaBoost and Extreme Gradient Boosting (XGB) was compared against GBLUP-MCMC. Each model prediction efficiency was also evaluated in situations where the ratio of the two observed phenotypes (resistant vs non-resistant) is unbalanced. Finally, the required computational time for training each ML model was benchmarked against GBLUP-MCMC.

2. Materials and methods

2.1. Simulated datasets

The QMSim software (Sargolzaei and Schenkel, 2009) was used for simulating phenotypic and their corresponding genotypic datasets. The initial historic population consisted of 2,000 generations with a constant size of 10,000 animals. The used parameters for simulating the historic population included equal sex ratio, random mating and discrete generations. Thereafter ten discrete non-overlapping recent generations were simulated using a breeding design often encountered in salmonids. In particular, 100 sires were considered to be uniquely mated with 200 dams in each generation with 30 animals from each family being phenotyped. The heritability of the simulated trait was equal to 0.3 with 300 biallelic and randomly located quantitative trait loci (QTL) affecting the trait. Furthermore, individuals from generation nine and ten (12,000 animals) were genotyped for 9,000 SNPs that were randomly distributed across a genome consisted of 30 chromosomes each of 100 cM in length. In order to simulate a binary phenotypic trait, the animals were assigned into two categories using different thresholds on their true breeding value. The thresholds were chosen in order to simulate for a scenario were the phenotypic distribution amongst the two categories (resistant vs non-resistant) was approximately balanced and another scenario where the percentage of resistant and non-resistant animals was between 20 and 25 % and between 75 and 80 %, respectively. Finally, both simulated scenarios were replicated ten times.

2.2. Carp resistance to koi herpesvirus dataset

A publicly available dataset from Palaiokostas et al. (2019) was used for assessing the efficiency of the ML models in terms of predicting carp resistant to koi herpesvirus disease (KHVD). The dataset consisted of 1, 255 carp juveniles with survival recordings for KHVD that were genotyped for 15,615 SNPs using restriction-site associated DNA sequencing (RAD-seq).

2.3. Baseline predictive efficiency using GBLUP-MCMC

GBLUP-MCMC was applied using the THRGIBBS1F90 module of the BLUPF90 suite (Misztal et al., 2018). The fitted model had the following form:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ was the vector of recorded phenotypes (resistant vs non-resistant). $\mathbf{X}$ and $\mathbf{Z}$ were the incidence matrices relating phenotypes with fixed and random effects. $\mathbf{b}$ represented the vector of the fixed effects (intercept), $\mathbf{u}$ the vector of random animal effects $\sim N(0, \mathbf{G}\ \sigma_g^2)$ with $\mathbf{G}$ corresponding to the genomic relationship matrix (VanRaden, 2008), σ_g^2 the additive genetic variance, $\mathbf{e}$ the vector of residuals $\sim N(0,$ $\mathbf{I}$σe2), $\mathbf{I}$ the identity matrix and σ_e^2 the residual variance. The model parameteres were estimated by Gibbs sampling (1,000,000 iterations; burn-in: 100,000; thin: 100).

2.4. Implementation of machine learning algorithms

An intercept term (known as bias in ML terminology) and the SNP genotypes were used as predictors (known as features in ML terminology) in all the ML models. The response variable in all the tested scenarios was a vector containing the disease resistance status of each animal (coded as 0 and 1 for non-resistant and resistant animals respectively). The Python library scikit-learn v0.22 (Pedregosa et al., 2011) was used for fitting all ML models (DT; RF; SVM; AdaBoost) apart from the XGB. In the latter case, the Python API of XGBOOST v1.2 was utilized. In order to reduce overfitting, appropriate regularization hyperparameters for each model were applied. In the case of DT, the maximum tree depth was restricted to 8. The magnitude of regularization in the case of SVM was controlled through the C parameter using a value of 1. For the ensembles, RF and XGB, a learning rate of 0.1 was used to minimize overfitting in addition to a maximum tree depth of 8. In the case of Adaboost, the maximum tree depth was fixed to 1. Moreover, the ensembles (RF; AdaBoost; XGB) were fitted using a maximum number of 2,000 base estimators. In all cases, the base estimators were decision trees. The feature importance (equivalent to SNP effect) for the used ensembles was plotted in a Manhattan plot form using the R/CMplot library v3.6.2 (Yin et al., 2020). Overall, all the aforementioned hyperparameter values were inferred after 3-fold cross-validation on the training set using the *RandomizedSearchCV* function of scikit-learn. The required computations were performed using Python v3.8, while the corresponding visualization plots were produced using the Seaborn library v0.11 of Python. Finally, an example of Python code for fitting the above models can be found in the Supplementary material (S1.html).

2.5. Model evaluation

The prediction efficiency of each tested model was assessed using receiver operating characteristic (ROC) curves. The models were ranked based on the area under the curve (AUC) metric, which by construction ranges between zero and one, with the latter representing the perfect classifier. In the simulated datasets, animals from the 9th generation were used to train the models, while the animals from the 10th generation served as a test set. On the other hand, since the carp dataset included animals from a single generation a 5-fold cross validation scheme was applied (Fig. 1).

3. Results

3.1. Simulated datasets

The performance of all tested models was based on predictions made on the test set, which was comprised of 6,000 animals from generation 10, while the training was conducted on the parental generation that also contained 6,000 animals. Animals from both generations were genotyped for 9,000 SNPs located randomly across the genome. The extreme gradient boosting (XGB) was the machine learning model that gave the highest AUC score across both scenarios with a mean value of 0.83 (Fig. 2).

Notably, the tested ensembles (DT, RF, Adaboost, XGB) provide estimates regarding the importance of each feature. With the exception of RF the rest of the ensembles performed as well variable selection by assigning values of zero to certain features. In the case of DT and Adaboost approximately 94 % and 92 % of the respective features had a value of zero. On the other hand in the case of XGB approximately 9% of the features were estimated of having an effect of zero (Fig. 3). Nevertheless, minor differences were observed between the best performing models (XGB, SVM, RF). In particular, the AUC score from XGB was 1–2 % higher than the equivalent of SVM and RF (Fig. 4). The aforementioned ML models slightly outperformed GBLUP-MCMC with an AUC 1–4 % higher than the latter. Adaboost performed slightly worse (3–4 %) than GBLUP-MCMC. On the other hand, the performance of DT was consistently low, with an average AUC of 0.54. It is important to note that an AUC score of 0.50 is expected merely by chance.

The standard deviation of the AUC score among replicates ranged between 0.01 and 0.03 for all the tested models. The lowest standard deviation was found in the case of DT (0.01), while the highest was observed in the case of GBLUP-MCMC (0.03). The best performing model (XGB) had a standard deviation of 0.02.

Two different scenarios were tested in the current study in terms of the phenotypic distribution of animals characterized as resistant or susceptible. More specifically, the model performance was tested in cases where the two recorded phenotypic categories had approximately an equal number of observations and in cases were the phenotypic distribution was skewed towards non-resistant animals. In the former case, the percentage of disease-resistant animals ranged between 42–47%, while in the latter case, it ranged between 11–23 %. The above ratios were consistent amongst the training and test sets. Overall, the model ranking was not affected by the ratio of resistant to non-resistant animals, with differences in terms of AUC scores being between 0.005 and 0.02 for each tested model. In the case of XGB that had overall the highest performance, the difference of AUC score was 0.008 among the two scenarios (Fig. 4).

3.2. Carp resistance to koi herpes virus

Model performance was inferred by following a 5-fold cross-validation scheme consisting of sets of 1,004 animals for training and 251 animals for validation purposes. The percentage of resistant animals amongst the training and validation sets ranged between 33–37%. Overall, the ranking of models was the same as in the case of the simulation datasets.

However, the differences between the best performing ML models and GBLUP-MCMC were minimal. Amongst all tested models, the XGB

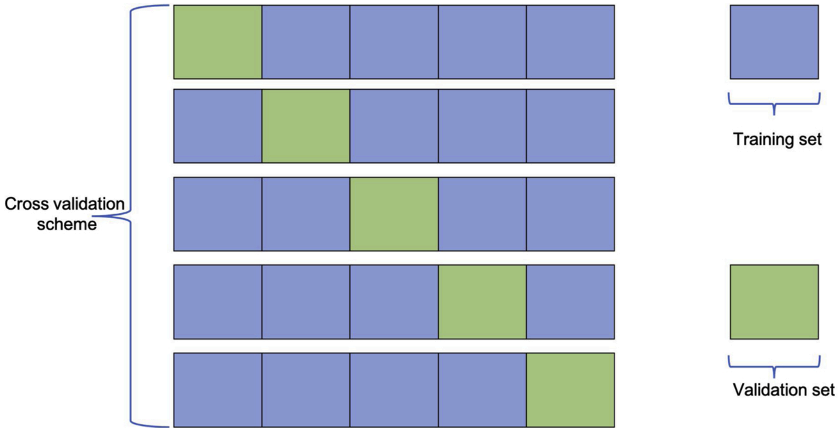

Fig. 1. Cross validation scheme.

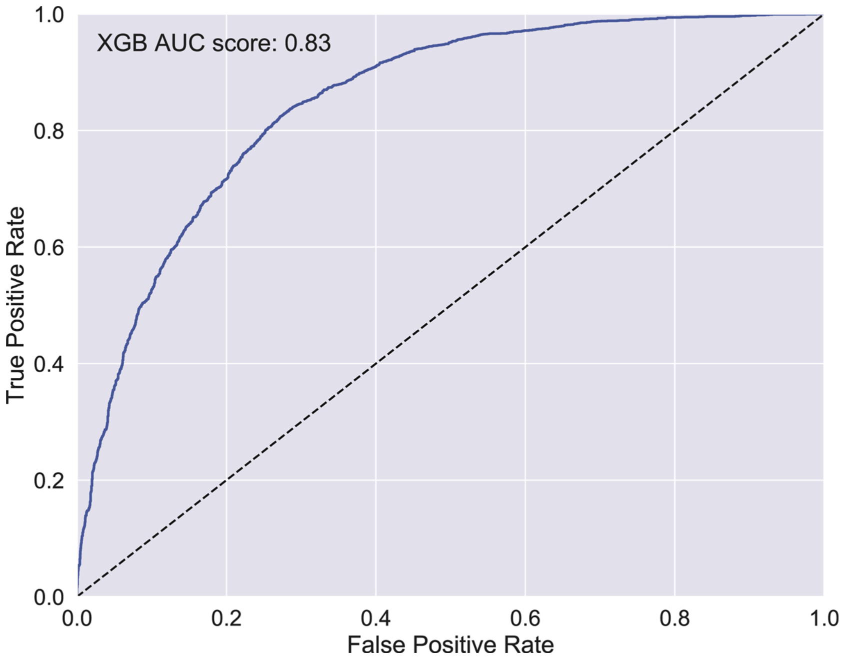

Fig. 2. Evaluation of the extreme gradient boosting (XGB) model with ROC curve. Mean AUC score derived from the simulation datasets.

provided the highest AUC score of 0.74. The above was only slightly higher compared to GBLUP-MCMC where an AUC score of 0.73 was recorded. The latter was practically equivalent to the AUC scores of SVC and RF. Adaboost performed slightly worse than GBLUP-MCMC, with the corresponding AUC scores being approximately 4% lower. As in the case of the simulated datasets, DT had the lowest performance with AUC score of 0.57. Finally, the standard deviation of the recorded AUC scores ranged between 0.02 – 0.03, with the lowest value observed in the case of GBLUP-MCMC (Fig. 5).

3.3. *Hyperparameter tuning - computational time*

The number of available hyperparameters for the ML models ranged between 5 – 18. Adaboost had the lowest number of hyperparameters, while XGB the highest. The magnitude of influencing the predictive ability of each ML by hyperparameter tuning varied substantially amongst the tested models (Table 1). Hyperparameter tuning had a more profound effect in the case of Adaboost, were fixing the maximum allowed depth of the underlying DT classifiers to 1 resulted in 40–50 % increase of the AUC score. On the other hand, changing the hyperparameter values from the default ones in the case of SVM resulted in worse predictions.

The computational time for fitting each model was benchmarked on an iMac (macOS 10.14.2) 4,2 GHz Intel Core i7 with 64 GB 2667 MHz DDR4 of RAM using the simulated datasets. All ML models required substantially less computational time compared to GBLUP-MCMC (Fig. 6) for fitting and prediction purposes. More specifically, the computational time for running GBLUP-MCMC was in the magnitude of hours (~4 h), while on the other hand, the computational time for ML ranged from seven seconds (DT) to approximately 30 min (SVM). Regarding the group of best performing ML models (RF, SVM, XGB), RF required approximately 3 min for completion, while SVM and XGB required approximately 30 and 10 min, respectively.

4. Discussion

The ability to predict disease resistance using genomic information in aquaculture species has attracted considerable research efforts (Elaswad and Dunham, 2018). In the current study, various ML models were evaluated in terms of their efficiency to detect disease-resistant animals through their genomic profile. Overall, promising results were obtained with the derived predictions of the best performing ML models, being in close proximity or even higher than the equivalent ones from GBLUP-MCMC. Even though no prior applications of ML in aquaculture breeding are available at the moment, encouraging results have been documented in both empirical and simulated datasets from plants and livestock (Montesinos-López et al., 2019; Waldmann, 2018). In the aforementioned cases, ML models performed at least equally well and even surpassed in certain scenarios the prediction efficiency of GS models in various regression tasks applied to continuous traits.

Traditionally the performance of various GS models for regression tasks in aquaculture species is mainly evaluated based on the so-called accuracy metric, which is, in fact, the Pearson correlation coefficient between the predicted values and the true breeding values (in case of a simulated dataset) or the phenotypic recordings (in case of empirical data) of the validation-test dataset (usually adjusted for fixed effects). Interestingly, it was recently pointed out that reliance solely on the correlation coefficient can result in a non-optimal model selection (Waldmann, 2019). The usage of the above accuracy term is the most common approach also for binary traits (e.g. resistant vs non-resistant) even though the definition of correlation, in this case, could be deemed somewhat problematic. However, the accuracy term is also commonly encountered in a broad literature of various classification problems where it denotes the number of cases predicted successfully out of the whole prediction attempts.

Nevertheless, it can be argued that none of the above definitions-usages of accuracy is optimal for binary traits. More specifically, the usage of accuracy for evaluating either GS or ML model performance in binary traits with a skewed ratio among the two observed phenotypic categories conveys limited practical value. Elaborating on the latter in a former study of genetic resistance of sea bream to pasteurellosis (Palaiokostas et al., 2016) where the percentage of resistant animals was approximately only 5%, a naïve classifier always predicting for a non-resistant animal would have achieved an accuracy of approximately 0.95. Model assessment was performed in the current study with ROC curves using the AUC metric. Through the simultaneous usage of false and true positive rate, ROC curves are less sensitive compared to accuracy in cases where the numbers of the two observed categories are not

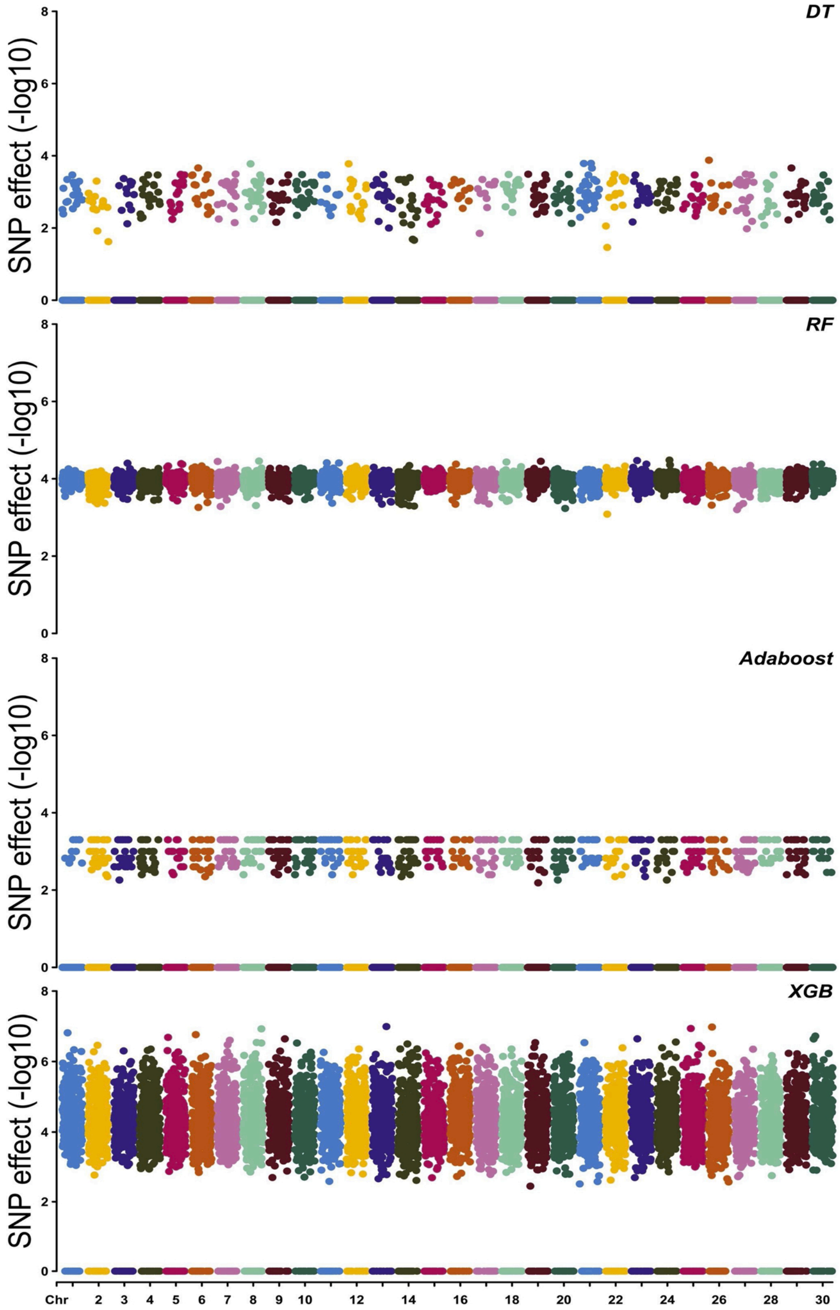

Fig. 3. Feature importance for each predictor SNP of the simulated dataset estimated from decision trees (DT), random forests (RF), adaptive boosting (Adaboost) and extreme gradient boosting (XGB).

balanced. Moreover, ROC curves have been already utilized in both human and animal genetic studies (Razgour et al., 2019; Tsairidou et al., 2014; Wray et al., 2010).

The results of the current study, including both simulated and empirical datasets, demonstrated that ML models could be successfully applied in classification problems relevant to breeding. According to the current results, the ranking of the tested models was not affected in the cases were an unbalanced distribution amongst the two observed phenotypes was used. In all tested scenarios, XGB was the model that ranked first though its advantage compared to GBLUP-MCMC was only slight, ranging between 1–4%. Furthermore, competitive predictions, as opposed to GBLUP-MCMC, were also obtained using SVM and RF. Notably, examples exist in the literature with applications of SVM, RF or similar ensemble learning algorithms using decision trees (e.g. Bayesian additive regression trees) in genomic selection studies on plants and livestock where the recorded prediction metric was of the same magnitude as with GBLUP-MCMC (Ogutu et al., 2011; Waldmann, 2016)

Even though no application of XGB in aquaculture selective breeding seems to have been documented as of now in the literature, the results of the current study coupled with the fact that it is one of the most powerful ML algorithms (Géron, 2019) suggest that it could be a valuable tool in future genetics studies of disease resistance in aquaculture. Interestingly, XGB was amongst the best performing models in terms of prediction efficiency for either sire conception rate in Holstein bulls or for simulated datasets (Abdollahi-Arpanahi et al., 2020). Furthermore, in the latter case, XGB ranked first in scenarios where the trait of interest was primarily controlled by non-additive genetic effects. On the other hand, XGB was outperformed by RF in terms of prediction efficiency for body weight in Brahman cattle (Li et al., 2018).

Notably, as is the case for most of the ML algorithms, XGB is

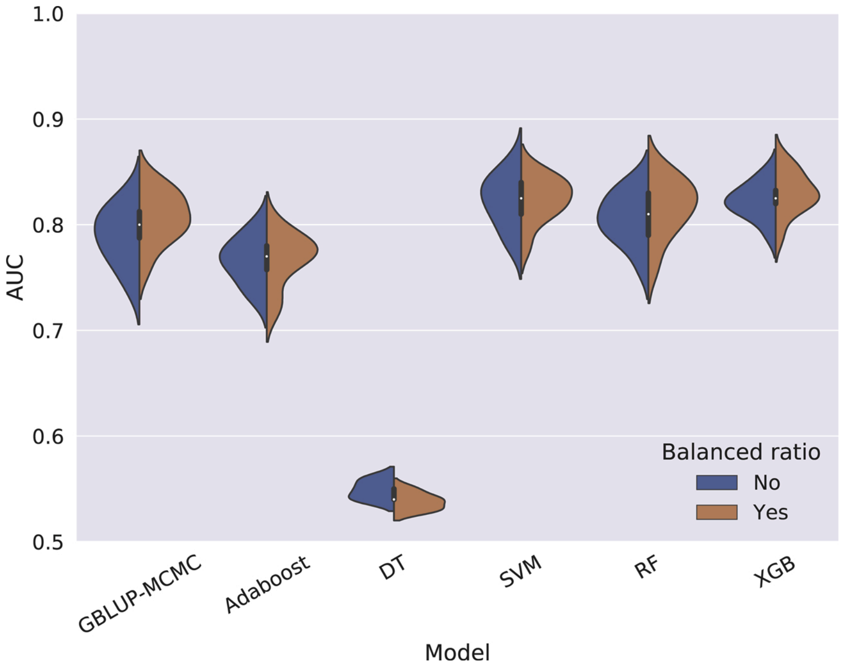

Fig. 4. Comparison of machine learning models with GBLUP-MCMC-MCMC based on their area under curve (AUC) score. The models were evaluated on simulation datasets with either balanced or skewed ratio of disease resistant vs non-resistant animals.

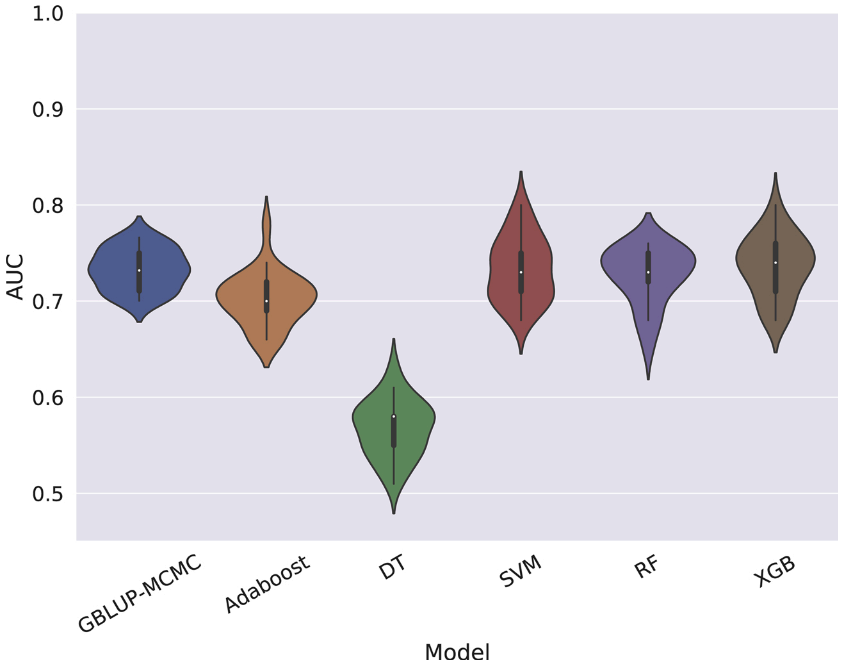

Fig. 5. Prediction of carp resistant to KHV using machine learning models and GBLUP-MCMC-MCMC.

Table 1
Predicted improvements (%) by fine-tuning respective hyperparameter values of each ML model.

Model	No. of hyperparameters	No. of hyperparameters tuned	Improvement over default %
DT	13	3	< 0
Adaboost	5	1	40 - 50
SVM	12	1	< 0
RF	14	4	5 - 10
XGB	18	5	8 - 12

particularly prone to overfitting, especially in datasets where the number of features (SNPs in the current case) far surpasses the number of observations. As such, XGB requires the *a priori* setting of regularization hyperparameters, which in the current case was achieved primarily by using the hyperparameters of learning rate and the maximum number of estimators. Generally speaking, the former parameter restricts the magnitude of the weight the algorithm assigns to each feature, while the latter refers to the maximum allowed number of base estimators that the algorithms uses. It should be stressed that in the current study, a non-extensive search for hyperparameter tuning was performed, so more effective predictions than the ones documented cannot be excluded.

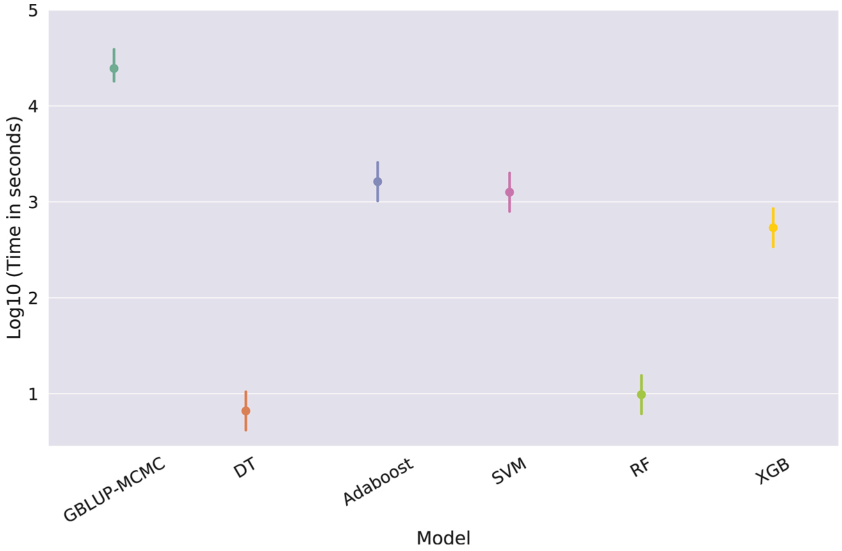

Fig. 6. Required computational time for fitting machine learning models and GBLUP-MCMC.

From all the tested ML models, hyperparameter fine-tuning had the most substantial effect in the case of Adaboost, where setting a single hyperparameter resulted in 40–50 % increase of the AUC score. On the other extreme, changing hyperparameter values from the default ones resulted in worse predictions in the case of SVM, indicating that fine-tuning hyperparameters in ML is a far from trivial task. Especially in the case of models with a high number of hyperparameters like XGB, an exhaustive search would be deemed particularly difficult and time-consuming.

Interestingly, XGB, Adaboost and RF are ensemble learning algorithms relying on aggregating the outcomes of base estimators (e.g. weak learners like DT) following different optimization routes (Biau and Scornet, 2016; Friedman, 2001) like bagging or pasting. In all three cases, the most common base estimator is the DT, with the fundamental idea being that through aggregating across the outcomes of several simple estimators, the prediction efficiency of the model can be improved compared to the equivalent of a single estimator (Géron, 2019). Even though gaining a full picture of the exact internal optimization route for each of the ensemble models is most challenging, it was evident from the acquired results that substantial differences exist in terms of the magnitude of variable selection. On one extreme, Adaboost performed predictions after zeroing the values of approximately 92 % of the available features (SNP genotypes), while on the other extreme, in the case of XGB, approximately 91 % of the available features had a non-zero value. As such, it could be hypothesized that in the case of complex traits where the underlying genetic architecture is polygenic, the strong variable selection performed with Adaboost might not be optimal. Nevertheless, since investigating the internals of ML models was not the primary focus of this study, future research would be advised in order to validate or not the above hypothesis.

An enormous amount of research has taken place in the field of animal breeding aiming to find the optimal model for most accurate and reliable predictions under all cases. Nevertheless, practical experience in line with the famous “no free lunch theorem” in mathematics suggests that this aim is probably unrealistic. Constraining our focus on the task of predicting disease resistance in aquaculture and taking into consideration the wide variation of the underlying genetic mechanisms involved in various diseases, it is doubtful that a single model, whether from the GS or ML, will be optimal for all cases. However, it is fair to state that GBLUP-MCMC is a robust approach, as was also clearly shown in the current study. Moreover, even though a slight advantage was observed in favour of the ML models compared to GBLUP-MCMC, it could easily be the case that the situation reverses in other datasets. Nevertheless, a significant advantage of the tested ML models lies in substantial reductions of computational time compared to GBLUP-MCMC in terms of model fitting. It should be stressed that the above mainly refers to comparisons in the case of binary traits when GBLUP relies on MCMC. Furthermore, taking into account that some of the ML models contain a large number of hyperparameters an extensive fine-tuning could easily require substantial computational time. In addition, it should be stressed that a certain aspect a subjectivity exists in the whole argument of computational efficiency as the number of iterations of the MCMC can vary depending on each specific application. In the current study, a relatively high number of iterations was used as in the case of binary traits, the mixing of the MCMC is slow. Nevertheless, despite the above, it is still apparent that ML, mainly due to parallelization of the assigned tasks, clearly outperform MCMC based algorithms in terms of computational efficiency. Notably, more substantial differences could be expected between the two classes in the case of using high-performance computing (HPC). Significant reductions of the required computational time were recorded in other studies as well. More specifically, SVM was shown to outperform various GS models in terms of computational efficiency (Montesinos-López et al., 2019). In the current study, 20-fold and above reductions of the computational time were observed between XGB and GBLUP-MCMC. Taking into account the fact that genotyping efforts in the field of aquaculture breeding continuously increase the availability of tools with the ability to produce accurate predictions within a reasonable time is highly attractive.

Overall, it is important to stress that several simplifications were applied in the current study. Foremost, the model evaluation was conducted on the basis of disease resistance being simplified as a binary trait. Even though this approach is appealing from a practical perspective, it could be argued that genetic resistance against a disease is a far more complicated process. As such, future studies, including information regarding the resilience and tolerance of the host against pathogens, can shed additional light and contribute to expediting the genetic progress through selective breeding (Knap and Doeschl-Wilson, 2020). Moreover, the performed simulations considered the genetic architecture of the trait as purely additive. Even though the latter has repeatedly proven to be a reliable approximation, it could well be the case that various interactive effects amongst the determining genetic components play an essential role in disease resistance. Interestingly, ML models usually shine in detecting non-linear patterns and interactions. Finally, even though an extensive range of popular ML models was tested here, the most highlighted category of deep neural networks (DNN) was intentionally not included as it would require a separate study on its own. DNN have various applications in the field of genomics (Eraslan et al., 2019) and could also be of high value in aquaculture breeding.

5. Conclusions

The results of the present study suggest that ML can be valuable tools in aquaculture breeding studies that aim to predict disease-resistant animals. XGB was the model that ranked first, conveying a slight advantage over GBLUP-MCMC that ranged between 1–4%. Furthermore, SVM and RF delivered competitive predictions as well. The application of solely DT is not recommended as low predictions were obtained consistently in all tested datasets. Finally, in terms of required computational time, all ML models clearly outperformed GBLUP-MCMC.

Declaration of Competing Interest

The author declares no conflict of interest.

Acknowledgments

The author acknowledges support from the SLU Aquaculture Platform. The author expresses his gratitude to the editor and two anonymous reviewers for their constructive comments and suggestions to improve this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aqrep.2021.100660.

References

Abdollahi-Arpanahi, R., Gianola, D., Peñagaricano, F., 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol. 52, 12. https://doi.org/10.1186/s12711-020-00531-z.

Asche, F., Hansen, H., Tveteras, R., Tveterås, S., 2009. The salmon disease crisis in Chile. Mar. Resour. Econ. 24, 405–411. https://doi.org/10.1086/mre.24.4.42629664.

Aslam, M.L., Carraro, R., Bestin, A., Cariou, S., Sonesson, A.K., Bruant, J.-S., et al., 2018. Genetics of resistance to photobacteriosis in gilthead sea bream (*Sparus aurata*) using 2b-RAD sequencing. BMC Genet. 19, 43. https://doi.org/10.1186/s12863-018-0631-x.

Aslam, M.L., Boison, S.A., Lillehammer, M., Norris, A., Gjerde, B., 2020. Genome-wide association mapping and accuracy of predictions for amoebic gill disease in Atlantic salmon (*Salmo salar*). Sci. Rep. 10, 1–9. https://doi.org/10.1038/s41598-020-63423-8.

Barbanti, A., Torrado, H., Macpherson, E., Bargelloni, L., Franch, R., Carreras, C., et al., 2020. Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. Mol. Ecol. Resour. 20, 795–806. https://doi.org/10.1111/1755-0998.13144.

Barría, A., Christensen, K.A., Yoshida, G.M., Correa, K., Jedlicki, A., Lhorente, J.P., et al., 2018. Genomic predictions and genome-wide association study of resistance against *Piscirickettsia salmonis* in coho salmon (*Oncorhynchus kisutch*) using ddRAD sequencing. G3 Bethesda (Bethesda) 8, 1183–1194. https://doi.org/10.1534/g3.118.200053.

Besson, M., Allal, F., Chatain, B., Vergnet, A., Clota, F., Vandeputte, M., 2019. Combining individual phenotypes of feed intake with genomic data to improve feed efficiency in Sea Bass. Front. Genet. 10, 219. https://doi.org/10.3389/fgene.2019.00219.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Boison, S., Ding, J., Leder, E., Gjerde, B., Bergtun, P.H., Norris, A., et al., 2019. QTLs associated with resistance to cardiomyopathy syndrome in Atlantic Salmon. J. Hered. 110, 727–737. https://doi.org/10.1093/jhered/esz042.

Correa, K., Bangera, R., Figueroa, R., Lhorente, J.P., Yáñez, J.M., 2017. The use of genomic information increases the accuracy of breeding value predictions for sea louse (*Caligus rogercresseyi*) resistance in Atlantic salmon (*Salmo salar*). Genet. Sel. Evol. 49, 15. https://doi.org/10.1186/s12711-017-0291-8.

Elaswad, A., Dunham, R., 2018. Disease reduction in aquaculture with genetic and genomic technology: current and future approaches. Rev. Aquac. 10, 876–898. https://doi.org/10.1111/raq.12205.

Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F.J., 2019. Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. 20, 389–403. https://doi.org/10.1038/s41576-019-0122-6.

Faggion, S., Vandeputte, M., Chatain, B., Gagnaire, P.-A., Allal, F., 2019. Population-specific variations of the genetic architecture of sex determination in wild European sea bass *Dicentrarchus labrax* L. Heredity (Edinb) 122, 612–621. https://doi.org/10.1038/s41437-018-0157-z.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Garcia, A.L.S., Bosworth, B., Waldbieser, G., Misztal, I., Tsuruta, S., Lourenco, D.A.L., 2018. Development of genomic predictions for harvest and carcass weight in channel catfish 06 Biological Sciences 0604 Genetics. Genet. Sel. Evol. 50, 66. https://doi.org/10.1186/s12711-018-0435-5.

Géron, A., 2019. Hands-On Machine Learning with Scikit-learn, Keras, and TensorFlow. O'Reilly Media.

Gianola, D., 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194, 573–596. https://doi.org/10.1534/genetics.113.151753.

Gonen, S., Baranski, M., Thorland, I., Norris, a, Grove, H., Arnesen, P., et al., 2015. Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (*Salmo salar*). Heredity (Edinb) 1–10. https://doi.org/10.1038/hdy.2015.37.

Gutierrez, A.P., Symonds, J., King, N., Steiner, K., Bean, T.P., Houston, R.D., 2020. Potential of genomic selection for improvement of resistance to ostreid herpesvirus in Pacific oyster (*Crassostrea gigas*). Anim. Genet. 51, 249–257. https://doi.org/10.1111/age.12909.

Horn, S.S., Ruyter, B., Meuwissen, T.H.E., Moghadam, H., Hillestad, B., Sonesson, A.K., 2020. GWAS identifies genetic variants associated with omega-3 fatty acid composition of Atlantic salmon fillets. Aquaculture 514, 734494. https://doi.org/10.1016/j.aquaculture.2019.734494.

Houston, R.D., 2017. Invited Review Future directions in breeding for disease resistance in aquaculture species. Bras. Zootec 46, 545–551. https://doi.org/10.1590/S1806-92902017000600010.

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., et al., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 1–21. https://doi.org/10.1038/s41576-020-0227-y.

Howard, R., Carriquiry, A.L., Beavis, W.D., 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes, Genomes, Genet. 4, 1027–1046. https://doi.org/10.1534/g3.114.010298.

Joshi, R., Skaarud, A., de Vera, M., Alvarez, A.T., Ødegård, J., 2020. Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (*Oreochromis niloticus*). Aquaculture 516, 734641. https://doi.org/10.1016/j.aquaculture.2019.734641.

Knap, P.W., Doeschl-Wilson, A., 2020. Why breed disease-resilient livestock, and how? Genet. Sel. Evol. 52, 1–18. https://doi.org/10.1186/s12711-020-00580-4.

Lhorente, J.P., Araneda, M., Neira, R., Yáñez, J.M., 2019. Advances in genetic improvement for salmon and trout aquaculture: the Chilean situation and prospects. Rev. Aquac. 11, 340–353. https://doi.org/10.1111/raq.12335.

Li, B., Zhang, N., Wang, Y.-G., George, A.W., Reverter, A., Li, Y., 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9, 237. https://doi.org/10.3389/fgene.2018.00237.

Lourenco, D.A.L., Tsuruta, S., Fragomeni, B.O., Masuda, Y., Aguilar, I., Legarra, A., et al., 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93, 2653–2662. https://doi.org/10.2527/jas.2014-8836.

Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., Misztal, I., 2020. Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. Genes (Basel) 11, 790. https://doi.org/10.3390/genes11070790.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Misztal, I., Tsuruta, Shogo, Lourenco, Daniela, Aguilar, I., Legarra, A., Vitezica, Zulma, 2018. Manual for BLUPF90 Family of Programs. Univ. Georg., Athens, USA. Available at: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf (Accessed 12 July 2018).

Misztal, I., Lourenco, D., Legarra, A., 2020. Current status of genomic evaluation. J. Anim. Sci. 98, 1–14. https://doi.org/10.1093/jas/skaa101.

Montesinos-López, O.A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al., 2019. 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, Genet. 9, 601–618. https://doi.org/10.1534/g3.118.200998.

Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B., Neil, D., Saffari, A., Mead, R., et al., 2020. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat. Rev. Neurol. 16. https://doi.org/10.1038/s41582-020-0377-8.

Nayeri, S., Sargolzaei, M., Tulpan, D., 2019. A review of traditional and machine learning methods applied to animal breeding. Anim. Heal. Res. Rev. 20, 31–46. https://doi.org/10.1017/S1466252319000148.

Ogutu, J.O., Piepho, H.P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings (BioMed Central) 1–5. https://doi.org/10.1186/1753-6561-5-S3-S11.

Palaiokostas, C., Ferraresso, S., Franch, R., Houston, R.D., Bargelloni, L., 2016. Genomic prediction of resistance to pasteurellosis in Gilthead Sea Bream (*Sparus aurata*) using 2b-RAD sequencing. G3 Bethesda (Bethesda) 6, 3693–3700. https://doi.org/10.1534/g3.116.035220.

Palaiokostas, C., Vesely, T., Kocour, M., Prchal, M., Pokorova, D., Piackova, V., et al., 2019. Optimizing genomic prediction of host resistance to Koi herpesvirus disease in carp. Front. Genet. 10, 543. https://doi.org/10.3389/fgene.2019.00543.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pérez, P., de los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198, 483–495. https://doi.org/10.1534/genetics.114.164442.

Pérez-Enciso, M., 2017. Animal Breeding learning from machine learning. J. Anim. Breed. Genet. 134, 85–86. https://doi.org/10.1111/jbg.12263.

Razgour, O., Forester, B., Taggart, J.B., Bekaert, M., Juste, J., Ibáñez, C., et al., 2019. Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. Proc. Natl. Acad. Sci. U. S. A. 116, 10418–10423. https://doi.org/10.1073/pnas.1820663116.

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R., 2017. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev. Aquac. 0, 1–13. https://doi.org/10.1111/raq.12193.

Robledo, D., Matika, O., Hamilton, A., Houston, R.D., 2018. Genome-wide association and genomic selection for resistance to amoebic gill disease in Atlantic Salmon. G3 Bethesda (Bethesda) 8, 1195–1203. https://doi.org/10.1534/g3.118.200075.

Sargolzaei, M., Schenkel, F.S., 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25, 680–681. https://doi.org/10.1093/bioinformatics/btp045.

Saura, M., Carabaño, M.J., Fernández, A., Cabaleiro, S., Doeschl-Wilson, A.B., Anacleto, O., et al., 2019. Disentangling genetic variation for resistance and endurance to scuticociliatosis in turbot using pedigree and genomic information. Front. Genet. 10, 539. https://doi.org/10.3389/fgene.2019.00539.

Tsairidou, S., Woolliams, J.A., Allen, A.R., Skuce, R.A., McBride, S.H., Wright, D.M., et al., 2014. Genomic prediction for tuberculosis resistance in dairy cattle. PLoS One 9. https://doi.org/10.1371/journal.pone.0096728.

Vallejo, R.L., Leeds, T.D., Fragomeni, B.O., Gao, G., Hernandez, A.G., Misztal, I., et al., 2016. Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. Front. Genet. 7, 96. https://doi.org/10.3389/fgene.2016.00096.

Vallejo, R.L., Cheng, H., Fragomeni, B.O., Shewbridge, K.L., Gao, G., MacMillan, J.R., et al., 2019. Genome-wide association analysis and accuracy of genome-enabled breeding value predictions for resistance to infectious hematopoietic necrosis virus in a commercial rainbow trout breeding population. Genet. Sel. Evol. 51, 47. https://doi.org/10.1186/s12711-019-0489-z.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Waldmann, P., 2016. Genome-wide prediction using Bayesian additive regression trees. Genet. Sel. Evol. 48, 42. https://doi.org/10.1186/s12711-016-0219-8.

Waldmann, P., 2018. Approximate Bayesian neural networks in genomic prediction. Genet. Sel. Evol. 50, 1–9. https://doi.org/10.1186/s12711-018-0439-1.

Waldmann, P., 2019. On the use of the pearson correlation coefficient for model evaluation in genome-wide prediction. Front. Genet. 10, 899. https://doi.org/10.3389/fgene.2019.00899.

Wilmott, P., 2019. Machine learning: an applied mathematics introduction. Panda Ohana 1–219.

Wray, N.R., Yang, J., Goddard, M.E., Visscher, P.M., Kimberly, R., 2010. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 6, e1000864 https://doi.org/10.1371/journal.pgen.1000864.

Yáñez, J.M., Houston, R.D., Newman, S., 2014. Genetics and genomics of disease resistance in salmonid species. Front. Genet. 5, 415. https://doi.org/10.3389/fgene.2014.00415.

Yáñez, J.M., Joshi, R., Yoshida, G.M., 2020. Genomics to accelerate genetic improvement in tilapia. Anim. Genet. 51, 658–674. https://doi.org/10.1111/age.12989.

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al., 2020. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. bioRxiv. https://doi.org/10.1101/2020.08.20.258491, 2020.08.20.258491.

You, X., Shan, X., Shi, Q., 2020. Research advances in the genomics and applications for molecular breeding of aquaculture animals. Aquaculture 526, 735357. https://doi.org/10.1016/j.aquaculture.2020.735357.