JABG. 2024 March, 8(1)17-32
https://doi.org/10.12972/jabng.20240103

Print ISSN 1226-5543
Online ISSN 2586-4297

Research Article

Selection of informative markers using machine learning approaches and genome-wide association studies to improve genomic prediction in Hanwoo cattle: a simulation study

Waruni Ekanayake[1], Phuong Thanh N. Dinh[2], Jun Heon Lee[3], Seung Hwan Lee[3*]

[1]Department of Bio-Big Data, Chungnam National University, Daejeon 34134, Republic of Korea

[2]Department of Bio-AI Convergence, Chungnam National University, Daejeon 34134, Republic of Korea

[3]Division of Animal & Dairy Science, Chungnam National University, Daejeon, 34134, Republic of Korea

***Corresponding author: Seung Hwan Lee,** Division of Animal & Dairy Science, Chungnam National University, Daejeon, Korea, E-mail: slee46@cnu.ac.kr

ABSTRACT

The present study deploys a comparison of Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Genome Wide Association Studies (GWAS) in selecting optimum subsets of single nucleotide polymorphisms (SNPs) to be used in genomic prediction in cattle. The data simulation was carried out for 6,000 animals and 47,841 SNPs which include 43,633 polygenic markers and 4208 quantitative trait loci (QTL) using QMSim software. The genomic prediction was conducted with the best linear unbiased prediction (BLUP) method using the BLUPF90 program. The accuracy of prediction was computed in three different types, namely, Empirical $_{\text{all SNPs}}$, Empirical $_{\text{QTL}}$, and theoretical accuracy, Accuracy $_{\text{PEV}}$. Among the three models, the highest Empirical $_{\text{all SNPs}}$ accuracy 0.79 was derived for GBM followed by 0.77 for XGBoost and 0.76 for GWAS. The Empirical $_{\text{QTL}}$ accuracy was almost equal for all three models. The maximum theoretical accuracy was obtained for GWAS which was 0.93, whereas GBM and XGBoost obtained 0.86 and 0.85 accuracy levels respectively. Our results indicate that all three models comparably performed in genomic predictions; however, subsets selected by both GBM and GWAS reported higher prediction accuracies compared to the whole SNP set. The number of QTL selected as a proportion of the total number of SNPs was superior in GWAS. These observations can be validated using real data which could enable further optimization of the analysis process.

Keywords: Extreme gradient boosting, Genome-wide association studies, Genomic prediction, Gradient boosting machine, Quantitative trait loci

INTRODUCTION

The rapid progression of biotechnologies has inundated the scientific community with vast amounts of genomic data, revolutionizing human medicine, as well as animal and plant sciences. As a result, the utilization of genetic markers, particularly single nucleotide polymorphisms (SNPs), for predicting phenotypes in animals and plants has surged in recent decades. While traditional methods persist, genomic selection— selecting animals and plants based on genomic prediction of phenotypes (Goddard & Hayes, 2007; Meuwissen et al., 2001) —has emerged as a promising approach in breeding programs (Sukhavachana et al., 2022). Genomic selection offers distinct advantages, including enhanced accuracy of estimated breeding values, accelerated genetic gains, reduced generation intervals, and overall cost savings in animal production (Daetwyler et al., 2013; van der Werf, 2013; Wiggans et al., 2017).

Received December 24, 2023 **Revised** April 01, 2024 **Accepted** April 01, 2024

Despite these advantages, the surge in high-dimensional SNP data, coupled with limited observations, has given rise to the 'curse of dimensionality' or the challenge of dealing with the large P (predicting variables) small N (number of samples) problem (Nayeri et al., 2019). Statistical and regression methods, traditionally reliant on univariate hypotheses and independent predictor variables, struggle when the number of predicting variables surpasses the number of samples and when a high correlation exists among predictors (Ayers & Cordell, 2010; Li et al., 2018). Another drawback of dense SNP sets is that they may not yield the anticipated improvement in prediction accuracy. The majority of markers in these sets are phenotypically neutral and the markers that are truly affecting phenotype is a small proportion (Jeong et al., 2020). Thus, using all statistically significant SNPs from dense sets, such as those identified in genome-wide association studies (GWAS), may not necessarily lead to improved prediction accuracies (Ober et al., 2012; Veerkamp et al., 2016).

Studies have highlighted that utilizing pre-selected sets of informative SNPs can alleviate the challenges associated with dense SNP sets (Al Kalaldeh et al., 2019). Berg et al. (2016) observed that selecting markers from multi-breed GWAS has considerably improved prediction accuracies in dairy cattle. Furthermore, genic SNPs (SNPs in or around a gene) from whole-genome sequence (WGS) data enhanced prediction ability in laying chicken in a study by Ni et al. (2017). Additionally, a study with cattle, observed that GWAS-significant markers located within QTL can increase the solidity of genomic prediction (Brøndum et al., 2015). The practice of selecting a subset of genomic variants, also known as feature selection, holds promise in improving genomic prediction accuracy. Genomic feature selection involves extracting the most 'informative' genetic features while eliminating noisy, non-informative, irrelevant, and redundant features (Pudjihartono et al., 2022). Scientists deploy three main categories of feature selection methods: filters, wrappers, and embedded methods (Tadist et al., 2019). Filters (Chi-squared test, information gain, and correlation coefficient scores) use statistical tests to rank and eliminate features based on a predetermined threshold, while wrappers (sequential forward and backward selection, randomized hill climbing, and recursive feature elimination) require the support of machine learning algorithms for evaluation. Filters are computationally efficient but provide a generalized output; however, wrappers are more efficacious and demand higher computational power. Embedded methods, incorporating strengths from both filters and wrappers, include techniques like LASSO regularization, random forests, and elastic net.

To assess the efficacy of feature selection methods in obtaining optimal sets of informative genomic variants, this study employed three approaches: Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Genome-wide Association Studies (GWAS). GBM (Jerome, 2001; Schapire, 2003) and XGBoost (Chen & Guestrin, 2016) are ensemble machine learning algorithms that enhance predictive models by sequentially adding decision trees to minimize prediction errors by employing gradient descent error minimization and boosting techniques (Nayeri et al., 2019). XGBoost, being an optimized form of gradient boosting, offers efficient regularization to avoid overfitting, parallel processing, tree-pruning, the ability to handle sparse data, and adaptation to scale up on multicore machines. The primary criterion for selecting genetic variants in this study was the 'feature importance' value (the contribution of each genetic feature towards a good model prediction) (Johnsen et al., 2023) assigned by the machine learning algorithms. The third SNP selection method, GWAS identifies statistical associations between genetic variants and a trait or the risk for a disease, which is represented by a significance p-value, by testing for deviations in the allele frequency of genetic variants of individuals (Uffelmann et al., 2021). The study compared GBM, XGBoost, and GWAS in selecting informative SNPs and the selected SNPs were then utilized to predict Genomic Estimated Breeding Values (GEBV) using the Best Linear Unbiased Prediction (BLUP) method.

This study underscores the significance of feature selection in improving the accuracy of genomic predictions. The comparison of machine learning algorithms with GWAS provides valuable insights into the efficiency and effectiveness of different approaches in selecting informative genomic variants. This research contributes to refining breeding programs by identifying optimal sets of genetic markers for predicting phenotypes in animals.

MATERIALS AND METHODS

The workflow of the analysis is depicted in Figure 1.

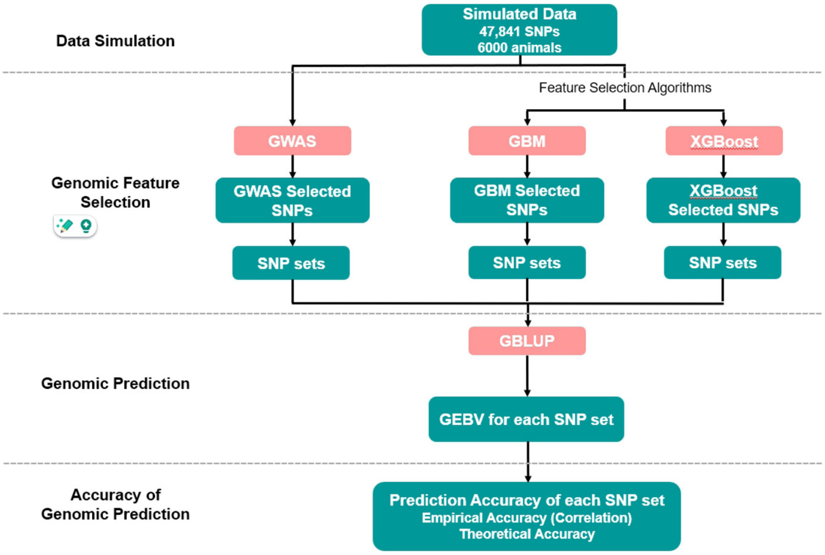

Figure 1. Workflow of the analysis. The overall analysis process consists of four phases namely, data simulation, genomic feature selection, genomic prediction, and calculation of the accuracy of genomic prediction.

Simulation of Data

QMSim 2.0 (Sargolzaei & Schenkel, 2009), a whole genome simulator for livestock was used to generate complex pedigrees and genotype data for cattle, assuming carcass weight as a quantitative trait. Parameters of the cattle population were simulated in five sections such as global parameters, historical populations, subpopulations and generations, genome parameters, and output options with comprehensive information on the simulated populations. First, a historical population with 500 animals in the first generation was simulated and the number of animals was gradually increased up to 1,400 across 50 generations creating a population expansion. An equal probability of being male or female animals was assumed and the mating design was inbreeding. 200 males and 1,200 females from the last historical generation were considered the founders of the recent population. The number of generations in the recent population was 5 and the mating design was inbreeding in which the inbreeding coefficient was 0.088. The proportion of male and female progeny in the current population was equal. Other genetic parameters were specified as an overall heritability of 0.45, QTL heritability of 0.4, and a phenotypic variance of 100. A genome (50k marker-density panel) with 29 chromosomes with differing lengths, with 47,841 SNPs (43,633 polygenic markers and 4,208 QTL) which were designed randomly for each chromosome, was simulated. The polygenic marker and QTL numbers were decided based on the information available on the Animal QTL database (https://www.animalgenome.org/QTLdb/). The population structure and simulation parameters specified are summarized in Table 1. Simulated data were then imputed using the Beagle 5.4 program (Browning et al., 2018).

Table 1. Population structure and simulation parameters.

Parameter	Value
Step 1: Historical Generations (HG)	
Number of generations (size) – phase 1	0 (500)
Number of generations (size) – phase 2	20 (700)
Number of generations (size) – phase 3	30 (1400)
Step 2: Recent Generations	
Number of founder males from the HG	200
Number of founder females form the HG	1200
Number of generations	5
Ratio of males	50%
Mating design	Maximize inbreeding
Replacement ratio for males	30%
Replacement ratio for females	30%
Selection design	High EBV
Culling design	Low EBV
Breeding value estimation method	BLUP
Trait heritability	0.45
QTL heritability	0.40
Phenotypic variance	1.00
Genome	
Number of chromosomes	29
Total length	2,484 cM
Number of markers	43,633
Marker distribution	Random
Number of QTL	4,208
QTL distribution	Random
MAF for markers	Equal
MAF for QTL	Equal
Additive allelic effects for markers	
Additive allelic effects for QTL	Gamma distribution (shape = 0.4)
Rate of missing marker genotypes	0.01
Rate of marker genotyping errors	0.005

Genomic Feature selection

The genomic feature selection procedure was carried out as follows for both GBM and XGBoost algorithms separately. Simulated phenotype and genotype data were prepared and fed into the algorithms and the feature importance values were obtained for each SNP. SNP sets which include different numbers of SNPs were compiled for each algorithm and each SNP set was used for genomic prediction of breeding values.

Feature Selection Algorithms

Gradient Boosting Machine (GBM)

The GBM algorithm was executed using the Python library GradientBoostingRegressor from Scikit-Learn 1.3.2 (Pedregosa et al., 2011). The essential hyperparameters were set as follows: n_estimators (number of boosting stages to perform/number of weak learners) = 5000; learning_rate (optimizes the model performance by scaling the contribution of each weak learner to the final model) = 0.1; max_depth (maximum depth of the individual regression estimators) = 6; and loss (loss function to be optimized) = 'squared error'. Other parameters of the model were set to default values.

In this study the loss function of GBM, mean squared error, was calculated by the following model:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of samples (animals) and y_i and $\hat{y}_i$ are the observed and predicted phenotypes for the i^{th} animal. The model predicts the phenotype, $\hat{y}_i = F_m(x_i)$ based on the following equation.

$$F_m(x_i) = F_{m-1}(x_i) + v \times \gamma_m h_m(x_i)$$

where, for a GBM of M stages, at each stage m ($1 \leq m \leq M$), F_m is an imperfect model

(weak learner), F_{m-1} is another imperfect model which is the predecessor of F_m, h_m is a new estimator, γ_m is the predicted value, v is the learning rate ($0 < v \leq 1$).

After fitting the model to prepared phenotype and genotype data, the model attribute, feature_importances_, was obtained for each SNP. For the GBM algorithm, the impurity-based feature importance as used to evaluate the importance of each SNP concerning the final prediction. Next, all the SNPs were ranked in descending order of the feature importance values. SNPs for which the algorithm computed the importance value were then extracted starting from the SNP with the highest importance value and different subsets of SNPs were compiled using the extracted SNPs.

Extreme Gradient Boosting (XGBoost)

XGBoost algorithm was run with Python programming language via Scikit-Learn Application Programming Interface (API). For the model hyperparameters, n_esimators (number of gradient boosted trees/number of boosting rounds) was set to 5000. Learning_rate (weighing factor for the corrections by new trees) and max_depth (maximum tree depth for base learners) were adjusted to 0.01 and 6 respectively. Unlike in the GBM algorithm, the importance_type (the feature importance type considered to score SNPs) was selected as "gain" which is the average gain across all splits the feature is used in.

The feature selection procedure with XGBoost was similar to that of the GBM algorithm. The SNPs which were assigned an importance value were organized in descending order and different SNP subsets were compiled starting from the SNP with the highest importance value.

Genome-wide Association Studies (GWAS)

Association studies were carried out using a linear mixed model using the GCTA 1.93.2 program (Yang et al., 2011) with the whole set of SNPs and phenotypes and resultant p-values were obtained. The linear regression equation used in GWAS can be written as follows:

$$y = Xb + Zu + g + e$$

where y is an $n \times 1$ vector of observed phenotypes, (n is the sample size and variance $var(y) = V = G\sigma_g^2 + I\sigma_e^2$ where G is the genetic relationship matrix (GRM) and I is the $n \times n$ identity matrix), X is a design matrix assigning records to fixed effects, b is a vector of fixed effects, Z is a design matrix allocating records to SNP effects, u is a vector of SNP effects, g is an $n \times 1$ vector of random effects of the polygenic effect of other SNPs with $g \sim N(0, G\sigma_g^2)$, and e is a vector of random residual effects with $e \sim N(0, I\sigma_e^2)$. GRM was estimated using the GREML module of GCTA 1.93.2 program.

The genome-wide significance threshold was considered as 1.04×10^{-6} and it was calculated using Bonferroni Correction. The SNPs that had p-values smaller than the significance threshold were extracted and a subset was compiled. Moreover, subsets that contained an equal number of SNPs as the subsets compiled using GBM and XGBoost algorithms were also assembled for the comparison of the performance of different feature selection methods.

Genomic Prediction

Estimated breeding values (EBV) were calculated with genomic best linear unbiased prediction (gBLUP) (Clark & van der Werf, 2013) method using BLUPF90 software (Misztal et al., 2018) for each subset of SNPs compiled using GBM and XGBoost algorithms and GWAS method. gBLUP equation used is as follows:

$$y = Xb + Zu + \varepsilon$$

where y is an $n \times 1$ vector of observed phenotypes, n is the sample size, X is a design matrix of order $n \times p$ relating the fixed effects to animals, b is a $p \times 1$ vector of fixed effects, Z is a design matrix of order $n \times q$ which allocates the records in y to the random effects in u, u is a $n \times 1$ vector of random effects (estimated breeding values; EBVs), and ε is an $n \times 1$ vector of residual terms.

Accuracy of Genomic Prediction

The accuracy of genomic prediction with each subset of SNPs compiled using three feature selection methods was calculated using three methods; the empirical accuracy using correlation between true breeding value (TBV) calculated with the effects of all SNPs and EBV ($Empirical_{all\ SNPs}$), the empirical accuracy using correlation between TBV calculated with the effects of only quantitative trait loci (QTL) and GEBV ($Empirical_{QTL}$), and the theoretical accuracy calculated using Prediction Error Variance (PEV) ($Accuracy_{PEV}$). The equations for empirical accuracies are:

$$Empirical_{all\ SNPs} = cor(TBV_{all\ SNPs}, EBV)$$

$$Empirical_{QTL} = cor(TBV_{QTL}, EBV)$$

where $TBV_{all\ SNPs}$ is the true breeding value calculated with the effects of all SNPs, TBV_{QTL} is the true breeding value calculated with the effects of only QTL, and EBV is the estimated breeding value.

The theoretical accuracy was calculated with the following equation:

$$Accuracy_{PEV} = \sqrt{1 - \frac{PEV}{\sigma_a^2}}$$

where PEV is the prediction error variance which is computed by the square of standard error of prediction (SE^2) and σ_a^2 is additive genetic variance.

RESULTS

Genomic Feature Selection

GBM and XGBoost computed corresponding feature importance values for each SNP and in contrast, GWAS provided a significance p-value concerning the SNPs. The distribution of SNPs according to the feature importance values computed by machine learning algorithms is depicted in Figure 2. The larger the feature importance value, the more important the SNP is. As per the results of both GBM and XGBoost, the majority of SNPs have very small or zero importance values (Figure 2A and 2B). GBM model assigned importance values for only 18,070 out of 47,841 SNPs which is 37.8% and XGBoost assigned importance values for 44,332 SNPs and it accounts for 92.7% of the total SNPs. For GBM algorithm, 18,070 SNPs (SNPs for which the feature importance values were computed) were extracted and ranked in descending order. Next, two subsets of SNPs which include 3,553 and 18,070 SNPs were compiled. Subsequently, 44,332 XGBoost selected SNPs were also extracted. After ranking SNPs from the highest importance value to the lowest, six different subsets of SNPS were compiled. Each subset consisted of 3,553, 18,070, 30,000, 35,000, 40,000, and 44,332 number of SNPs respectively. In Figure 2, the red dotted vertical lines represent the number of SNPs in each subset selected by GBM and XGBoost models. Figure 3 shows the resultant $-log_{10}$ transformed p-values for each SNP computed by GWAS with their genomic locations. The results of GWAS indicated that only 3,553 out of 47,841 SNPs (7.42%) have

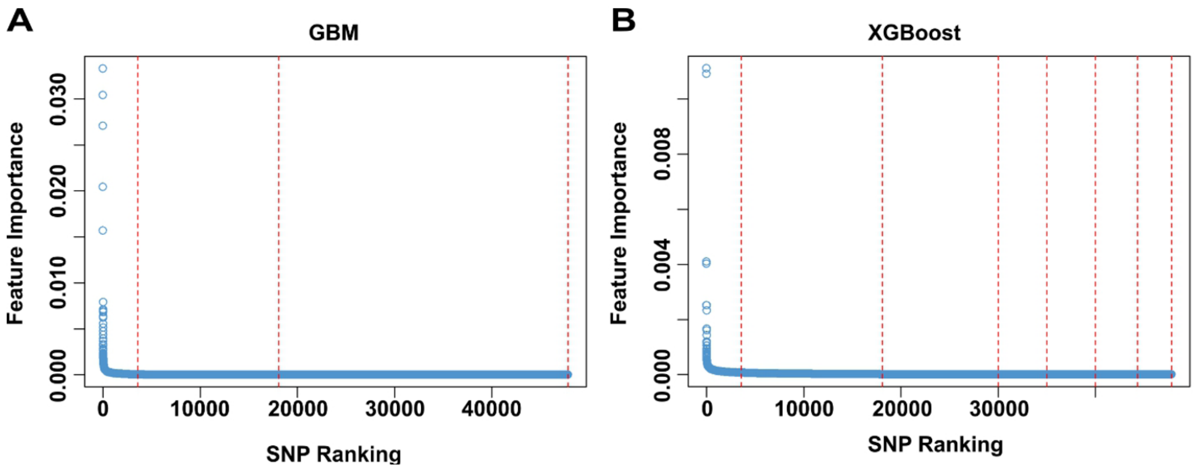

Figure 2. The distribution of SNPs according to the feature importance values computed by machine learning algorithms, GBM and XGBoost. The larger the feature importance values, the more important the SNP is. The red dotted vertical lines represent the number of SNPs in each subset selected by GBM and XGBoost.

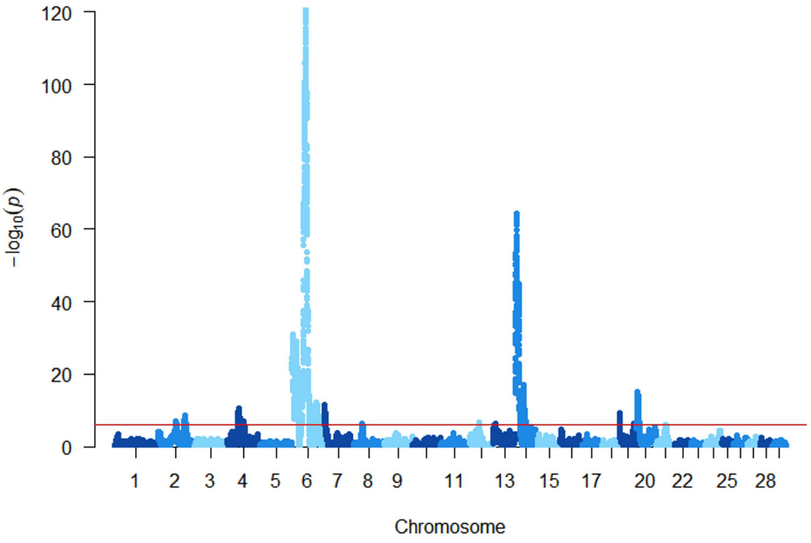

Figure 3. Manhattan plot of the genome-wide significance of each SNP. The x-axis represents the genomic position of each SNP. The y-axis represents the -log$_{10}$ transformed p-values for each SNP computed by GWAS. The red horizontal line represents the genome-wide significance threshold, 1.04×10^{-6}.

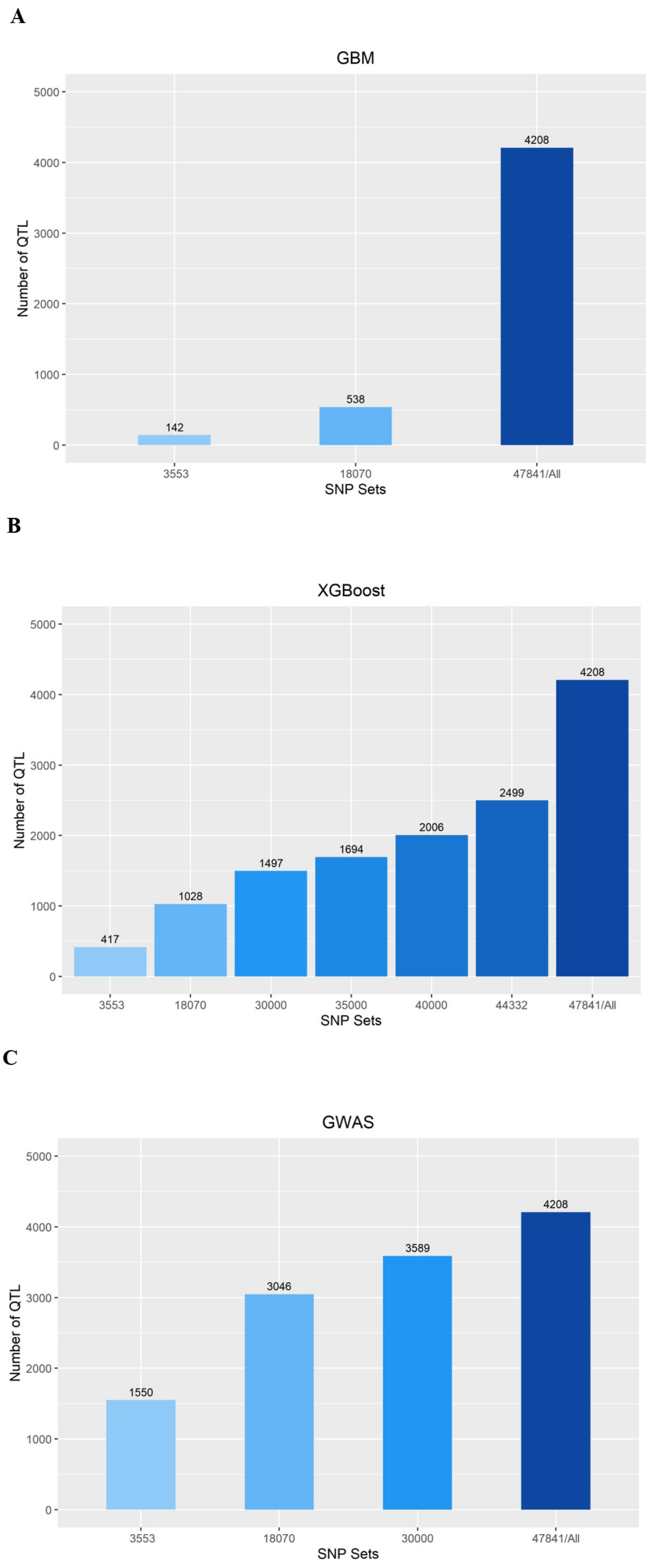

Figure 4. The number of QTL selected in each subset of SNPs by three feature selection methods. Subfigures A, B, and C represent the number of QTL selected by GBM, XGBoost, and GWAS respectively.

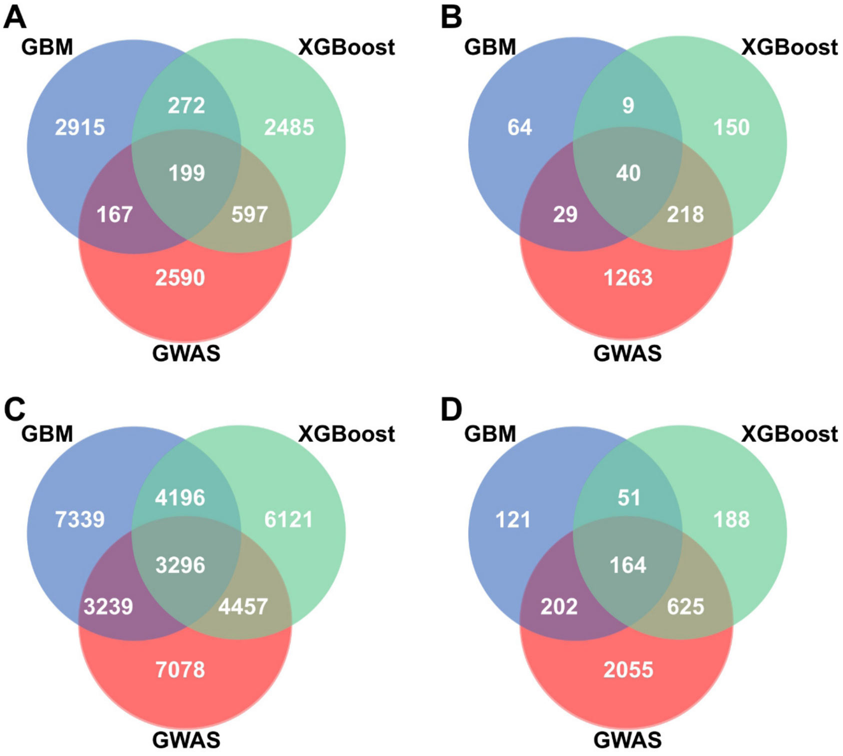

Figure 5. The number of SNPs and QTL mutually identified by three feature selection methods with 3,553 and 18,070 subsets. (A) Mutual SNPs in the 3,553 subset, (B) Mutual QTL in the 3,553 subset, (C) Mutual SNPs in the 18,070 subset, (D) Mutual QTL in the 18,070 subset.

p-values below the genome-wide significance threshold (1.04 × 10^{-6}). After ranking SNPs in ascending order of the p-value, two SNP subsets which contain 18,070 and 30,000 SNPs with the lowest p-values, were also compiled.

The numbers of QTL selected in each subset of SNPs by three feature selection methods, GBM, XGBoost, and GWAS, are illustrated in Figure 4. Out of 4,208 QTL included in the simulation data, there were 142 and 538 QTL included in the GBM selected subsets of 3,553 and 18,070 SNPs respectively (Figure 4A). On the other hand, each subset of 3,553, 18,070, 30,000, 35,000, 40,000, and 44,332 XGBoost selected SNPs accounted for 417, 1,028, 1,497, 1,694, 2,006, and 2,499 QTL respectively (Figure 4B). GWAS selected subsets 3,553, 18,070, and 30,000 subsets included 1,550, 3,046, and 3,589 QTL respectively (Figure 4C). Figure 5 demonstrates the number of SNPs and QTL mutually identified by each feature selection method with 3,553 and 18,070 SNPs sets. With 3,553 and 18,070 subsets, all three methods have commonly selected 199 and 3,296 SNPs and 40 and 164 QTL respectively.

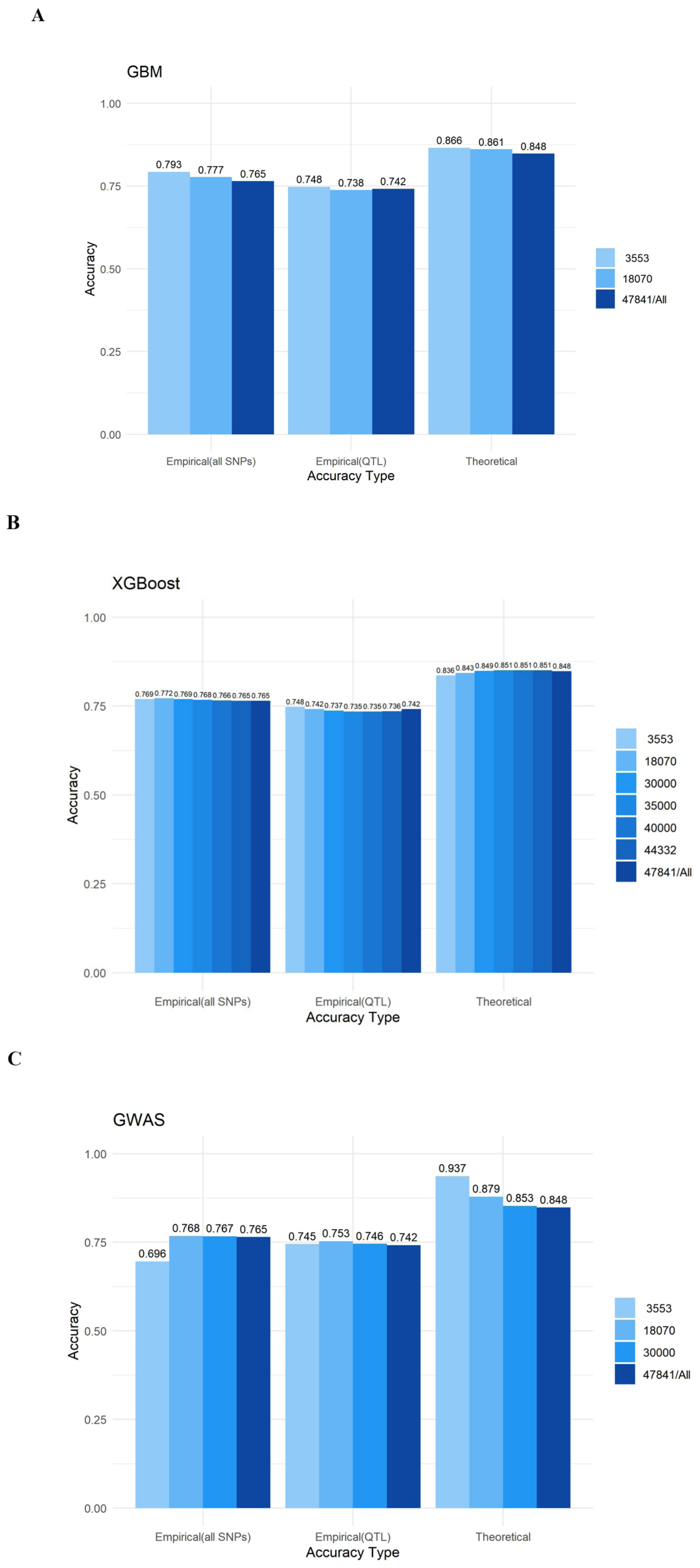

Figure 6. Three types of prediction accuracies, Empirical$_{\text{all SNPs}}$, Empirical$_{\text{QTL}}$, and Accuracy$_{\text{PEV}}$ computed with different SNP subsets selected by three SNP selection methods, GBM, XGBoost, and GWAS.

Prediction of Genomic Breeding Values and Accuracies of Prediction

Prediction of genomic breeding values was carried out with each subset of SNPs selected by GBM, XGBoost, and GWAS using the BLUP method. Prediction accuracies calculated by three approaches were then compared between three SNP selection methods and different SNP subsets. The three types of prediction accuracies, $Empirical_{all\ SNPs}$, $Empirical_{QTL}$, and $Accuracy_{PEV}$ (theoretical accuracy) computed with different SNP subsets selected by three SNP selection methods (GBM, XGBoost, and GWAS) are illustrated in Figure 6. Importantly, accuracies did not display significant differences with different SNP selection methods and SNP sets. $Empirical_{all\ SNPs}$, $Empirical_{QTL}$, and $Accuracy_{PEV}$ calculated using the whole set of SNPs (47,841) were 0.76, 0.74, and 0.84 respectively. The accuracy values for three different types of accuracies and different SNP sets were as follows.

Empirical accuracy calculated with all SNPs ($Empirical_{all\ SNPs}$)

For the $Empirical_{all\ SNPs}$ accuracy, the GBM model with 3,553 SNPs set was observed to have the highest accuracy of 0.79. However, the respective subset of XGBoost showed a slightly lower value of 0.76 and GWAS exhibited the lowest of 0.69. Interestingly, all three methods showed similar accuracies as the 18,070 subset. Moreover, for the XGBoost model, SNP subsets from 30,000 to 44,332, and the whole SNP set displayed the same accuracy of 0.76. Remarkably, for both GBM and XGboost algorithms, at least one subset of SNPs displayed a slightly higher accuracy than the SNP set with all SNPs (47,841). XGBoost subsets from 30,000 to 44,332 SNPs and GWAS subsets of 18,070 and 30,000 SNPs provided similar values for accuracy (0.76).

Empirical accuracy calculated with only QTL ($Empirical_{QTL}$)

The highest value of $Empirical_{QTL}$ accuracy of 0.75 was observed with GWAS selected 18,070 SNPs set. Notably, GBM selected 18,070 SNP set and XGBoost selected SNP sets of 30,000, 35,000, 40,000, and 44,332 displayed the lowest accuracy of 0.73. The $Empirical_{QTL}$ accuracy of GBM 18,070 subset was slightly lower than the accuracy calculated with all SNPs. However, for the XGBoost algorithm, both 18,070 and all SNPs sets provided equal accuracies (0.74). Noticeably, for $Empirical_{QTL}$ accuracy, SNP subsets provided similar values with the whole SNP set.

Theoretical Accuracy ($Accuracy_{PEV}$)

The 3,553 subset of SNPs selected by GWAS was observed to have the highest theoretical accuracy which is 0.93. On the other hand, the lowest value (0.83) was recorded with the XGBoost selected 3,553 subset. Importantly, except for the XGBoost method, both other methods (especially the GWAS method) showed higher theoretical accuracies with SNP subsets than that with all SNPs.

Comparison of Three Types of Accuracies

Interestingly, for all feature selection methods and all SNP sets, the theoretical accuracy ($Accuracy_{PEV}$) was higher than both types of empirical accuracies. For the GBM algorithm, the highest value of accuracy (0.86) was demonstrated for the theoretical accuracy with 3,553 subset and the 18,070 subset for $Empirical_{QTL}$ accuracy recorded the lowest value (0.73). Accuracies of XGBoost selected SNP sets were recorded as follows. Both 18,070, 30,000, subsets and all SNPs (47,841) showed the similar highest accuracies (0.84) with theoretical accuracy, and in contrast, subsets from 30,000 to 44,332 were observed to have the lowest values (0.73) with $Empirical_{QTL}$ accuracy. The theoretical accuracy of the 3,553 subset showed the highest accuracy (0.93) for GWAS SNP sets and the same subset gave the lowest (0.69) for $Empirical_{all\ SNPs}$ accuracy.

DISCUSSION

This research aimed to compare the accuracy of genomic breeding values computed using different subsets of SNPs selected by three feature selection methods: GBM, XGBoost, and GWAS, using simulation data. Numerous research suggests that feature selection procedures that utilize machine learning and statistical approaches can eliminate uninformative SNPs and decrease the complexity of genomic data which in turn maximizes the efficiency of genomic predictions (Pudjihartono et al., 2022; Tadist et al., 2019). Although GBM and XGBoost are commonly used feature selection methods, GWAS is not conventionally considered a feature selection method in the context of machine learning. Instead, GWAS is a statistical approach used in genomics to identify associations between genetic markers (SNPs) and a trait or risk for a disease (Witte, 2010). However, the genetic loci that are identified by GWAS as statistically associated with a trait or disease can be further studied to better understand the underlying biology of the condition. Thus, the results of GWAS can lead to the detection of genetic markers that are associated with a trait or disease.

The results from the SNP selection process revealed that GBM assigned positive feature importance values to a comparatively smaller number of SNPs than XGBoost. XGBoost applies a more formalized regularization with Lasso (L1) and Ridge (L2) regularization which controls overfitting (Mancin et al., 2022). Thus, XGBoost is expected to perform better than GBM and other ensemble machine learning techniques. Furthermore, another difference between the mechanisms of the two algorithms is the method each algorithm uses to calculate feature importance. GBM calculates feature importance based on Gini Importance or Mean Decrease in Impurity (MDI), which is the total reduction of the criterion brought by that feature. Specifically, Gini Importance calculates the importance of each SNP based on the sum of the number of splits across all trees that include the SNP weighted by the number of samples it splits (Nembrini et al., 2018). SNPs with higher Gini Importance have a higher influence on the final prediction of the phenotype. On the other hand, the feature importance type in XGBoost applied for this study is based on “gain”, which is the relative contribution of the corresponding feature to the model (Zheng et al., 2017). The above aspects may have led the two algorithms to select differing numbers of SNPs, however, a much more thorough analysis will be useful for a sound understanding of the scenario.

Linkage disequilibrium (LD) plays an important role in the efficiency of genomic prediction. In genomic prediction, it is assumed that markers are in LD with QTL thus, identifying DNA markers that are associated with QTL and using them in the prediction of phenotypes will improve the prediction accuracy (Chen et al., 2023). Therefore, one of the objectives of this study was to identify QTL commonly selected by all methods (Figures 5B and 5D) assuming that these QTL are in LD with DNA markers associated with the phenotypes. The number of QTL selected by three methods with each SNP subset substantially differs (Figure 4). It is vital to note that GWAS has selected the highest number of QTL (3,046) compared to GBM (538) and XGBoost (1,028) with 18,070 subset. There may be several reasons for GWAS to select more QTL than machine learning algorithms. First, GWAS is specifically designed for detecting the variant-trait association across the genome with larger sample sizes for higher statistical power, however, machine learning methods need even larger samples to achieve a comparable statistical power (Witte, 2010). Second, when detecting associations, GWAS does not rely on assumptions about the underlying genetic architecture, on the other hand, machine learning methods might sometimes focus on specific features or genomic regions. However, since the genetic variants identified by GWAS have smaller effect sizes and explain only a fraction of the phenotypic variation, more sophisticated techniques such as fine mapping are required to identify the true causal mutations (Pudjihartono et al., 2022).

The true breeding value (TBV) of an animal is the genetic potential of that animal or the real value of the animal for breeding (Oldenbroek et al., 2014). Since, knowing TBV is impossible in real situations, our objective is to evaluate how accurate our estimation of the TBV (EBV) is, using the correlation between TBV and EBV. However, given that this study uses simulation data, TBV is available and can be directly used to estimate the accuracy of predicted breeding value. Accordingly, prediction accuracy was calculated using three procedures. Apart from the common method of calculating the empirical accuracy ($Empirical_{all\ SNPs}$), which is readily outputted by simulation, empirical accuracy was also calculated with the TBV computed using only QTL effects, *Empirical*$_{QTL}$. QTL are the specific regions on the genome that contain genes regulating the traits (Paudel et al., 2020). Thus. focusing on these regions and using specific information about these regions, can lead to more precise predictions about the trait of interest.

The accuracy estimates indicate that concerning *Empirical*$_{all\ SNPs}$ accuracy, the 3,553 and 18,070 SNP subsets selected by machine learning models perform equally or slightly better than the whole set of SNPs compared to the respective subset selected by GWAS. Li et al. (2018) suggested that machine learning methods perform better in selecting SNPs since these methods capture non-linear relationships and interactions between SNPs more effectively, resulting in smaller residual variance, increased genetic variance, and heritability.

Noticeably, the GWAS genome-wide significant SNP set delivered the highest prediction accuracy (0.93) out of all selection methods and SNP sets. Even though this SNPs set contained the smallest number of SNPs of all sets (3,553) the highest accuracy can be attributed to the fact that, the specific subset has the highest proportion of QTL (43.6%) compared to the other subsets. Although the effect of each QTL is comparatively smaller, the aggregate of all QTL effects makes up the total genetic variance of a quantitative trait (Morgante et al., 2018). Thus, using a higher proportion of QTL in the BLUP prediction model may have been the reason for the higher accuracy of subsets than all SNPs set. Furthermore, it is important to note that theoretical accuracies produced higher values than empirical accuracies for all subsets. The theoretical accuracy is calculated with the prediction error variance (PEV) additive genetic variance (σ_a^2). PEV directly measures the variance of the difference between predicted and true breeding values. Hence, if the predicted values are closer to the true values with low variability, the PEV will be small, leading to a higher accuracy. Conversely, the accuracy derived from correlation measures the linear relationship between predicted and true breeding values, which may not be able to capture the variance of the predictions. PEV's ability to capture non-linear relationships between predicted and true values, correlation's sensitivity to the variables, outliers, and also the genetic architecture of the traits might be among the other reasons for the PEV-based accuracy to be higher than correlation-based accuracy.

According to the results, it can be concluded that the 3,553 subset selected by GWAS was the best SNP subset for genomic prediction among the SNP subsets considered in this study. However, with all three feature selection methods, prediction accuracy did not considerably differ with SNP subsets resulting from feature selection. The observed outcome could be owing to diverse reasons. First, even though machine learning methods can capture the non-linearities between multiple variants, a larger number of SNPs may have caused overfitting of models leading to lower accuracies (Elgart et al., 2022). Second, non-causal variants that are on LD with causal variants may provide redundant information resulting in low improvement in predictions (Uffelmann et al., 2021). High polygenicity of the trait, dominance and epistatic effects, and population structure are some of the other factors that may have played a considerable role in deciding the effectiveness of the prediction. Hence, it is essential to carefully consider the characteristics and the genetic architecture of the traits and the population structure when deciding on genomic prediction models. Furthermore, this research was conducted using only simulation data and it is worthwhile to test this study model with real Hanwoo data to investigate in what ways the results will be comparable to this study.

CONCLUSION

This study compared the effectiveness of feature selection methods GBM, XGBoost, and GWAS in selecting optimum subsets of SNPs, using simulated data of 6000 animals and 47,841 SNPs. The results suggest that all three methods performed equally well in selecting informative SNPs and the genomic predictions using SNP subsets provided similar prediction accuracies to those using the whole set of SNPs. Importantly, GWAS and GBM-selected subsets of 3,553 SNPs outperformed the whole SNP set. Therefore, it is concludable that a subset of informative markers selected by feature selection methods can result in a similar or higher genomic prediction accuracies compared to the whole set of markers.

CONFLICT OF INTERESTS

No potential conflict of interest relevant to this article is reported.

ACKNOWLEDGEMENTS

This research is funded by the project, deep learning modeling for genomic prediction to increase accuracy of breeding value in Korean and Israeli cattle populations (2022K1A3A1A31093393) and supported by the National Research Foundation of Korea.

REFERENCES

Al Kalaldeh, M., Gibson, J., Duijvesteijn, N., Daetwyler, H. D., MacLeod, I., Moghaddar, N., Lee, S. H., & van der Werf, J. H. J. (2019). Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. *Genetics Selection Evolution*, *51*(1), 32. https://doi.org/10.1186/s12711-019-0476-4

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol*, *34*(8), 879-891. https://doi.org/10.1002/gepi.20543

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107-4116. https://doi.org/https://doi.org/10.3168/jds.2014-9005

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet*, *103*(3), 338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Chen, Z.-Q., Klingberg, A., Hallingbäck, H. R., & Wu, H. X. (2023). Preselection of QTL markers enhances accuracy of genomic selection in Norway spruce. *BMC Genomics*, *24*(1), 147. https://doi.org/10.1186/s12864-023-09250-3

Clark, S. A., & van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods Mol Biol*, *1019*, 321-330. https://doi.org/10.1007/978-1-62703-447-0_13

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347-365. https://doi.org/10.1534/genetics.112.147983

Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., Lin, H. J., Raffield, L., Gao, Y., Chen, H., de Vries, P., Lloyd-Jones, D. M., Lange, L. A., Peloso, G. M., Fornage, M., Rotter, J. I., Rich, S. S., Morrison, A. C., Psaty, B. M., . . . the, N. s. T.-O. i. P. M. C. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology*, *5*(1), 856. https://doi.org/10.1038/s42003-022-03812-z

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323-330. https://doi.org/https://doi.org/10.1111/j.1439-0388.2007.00702.x

Jeong, S., Kim, J.-Y., & Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Scientific Reports*, *10*(1), 19653. https://doi.org/10.1038/s41598-020-76759-y

Jerome, H. F. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Johnsen, P. V., Strümke, I., Langaas, M., DeWan, A. T., & Riemer-Sørensen, S. (2023). Inferring feature importance with uncertainties with application to large genotype data. *PLOS Computational Biology*, *19*(3), e1010963. https://doi.org/10.1371/journal.pcbi.1010963

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

Mancin, E., Mota, L. F. M., Tuliozi, B., Verdiglione, R., Mantovani, R., & Sartori, C. (2022). Improvement of Genomic Predictions in Small Breeds by Construction of Genomic Relationship Matrix Through Variable Selection. *Front Genet*, *13*, 814264. https://doi.org/10.3389/fgene.2022.814264

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Morgante, F., Huang, W., Maltecca, C., & Mackay, T. F. C. (2018). Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. *Heredity*, *120*(6), 500-514. https://doi.org/10.1038/s41437-017-0043-0

Nayeri, S., Sargolzaei, M., & Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. *Anim Health Res Rev*, *20*(1), 31-46. https://doi.org/10.1017/s1466252319000148

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, *34*(21), 3711-3718. https://doi.org/10.1093/bioinformatics/bty373

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., Stricker, C., Gianola, D., Schlather, M., Mackay, T. F. C., & Simianer, H. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster. *PLOS Genetics*, *8*(5), e1002685. https://doi.org/10.1371/journal.pgen.1002685

Paudel, D., Dhakal, S., Parajuli, S., Adhikari, L., Peng, Z., Qian, Y., Shahi, D., Avci, M., Makaju, S. O., & Kannan, B. (2020). Chapter 38 - Use of quantitative trait loci to develop stress tolerance in plants. In D. K. Tripathi, V. Pratap Singh, D. K. Chauhan, S. Sharma, S. M. Prasad, N. K. Dubey, & N. Ramawat (Eds.), *Plant Life Under Changing Environment* (pp. 917-965). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-818204-8.00048-5

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform*, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, *25*(5), 680-681. https://doi.org/10.1093/bioinformatics/btp045

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear Estimation and Classification* (pp. 149-171). Springer New York. https://doi.org/10.1007/978-0-387-21579-2_9

Sukhavachana, S., Senanan, W., Tunkijjanukij, S., & Poompuang, S. (2022). Improving genomic prediction accuracy for harvest traits in Asian seabass (Lates calcarifer, Bloch 1790) via marker selection. *Aquaculture*, *550*, 737851. https://doi.org/https://doi.org/10.1016/j.aquaculture.2021.737851

Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., & Zahi, A. (2019). Feature selection methods and genomic big data: a systematic review. *Journal of Big Data*, *6*(1), 79. https://doi.org/10.1186/s40537-019-0241-0

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

van der Werf, J. (2013). Genomic selection in animal breeding programs. *Methods Mol Biol*, *1019*, 543-561. https://doi.org/10.1007/978-1-62703-447-0_26

Veerkamp, R. F., Bouwman, A. C., Schrooten, C., & Calus, M. P. L. (2016). Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. *Genetics Selection Evolution*, *48*(1), 95. https://doi.org/10.1186/s12711-016-0274-1

Wiggans, G. R., Cole, J. B., Hubbard, S. M., & Sonstegard, T. S. (2017). Genomic Selection in Dairy Cattle: The USDA Experience. *Annu Rev Anim Biosci*, *5*, 309-327. https://doi.org/10.1146/annurev-animal-021815-111422

Witte, J. S. (2010). Genome-wide association studies and beyond. *Annu Rev Public Health*, *31*, 9-20 24 p following 20. https://doi.org/10.1146/annurev.publhealth.012809.103723

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, *88*(1), 76-82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zheng, H., Yuan, J., & Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*, *10*(8).

AUTHORS INFORMATION

Seung Hwan Lee: https://orcid.org/0000-0003-1508-4887

Waruni Ekanayake: https://orcid.org/0000-0003-2988-0315

Phuong Thanh N. Dinh://orcid.org/0000-0002-3057-0210

Jun Heon Lee://orcid.org/0000-0003-3996-9209