Animal Science Papers and Reports vol. 40 (2022) no. 1, 45-58
Institute of Genetics and Animal Biotechnology
of the Polish Academy of Sciences, Jastrzębiec, Poland

An evaluation of machine learning for genomic prediction of hairy syndrome in dairy cattle

B. Karacaören*

Department of Animal Science, Faculty of Agriculture, Akdeniz University, Turkey

(Accepted February 2, 2022)

There is a growing body of literature that recognizes the importance of understanding the adaptation of the cattle populations to the climate conditions. Among others heat stress is one of the causes of reduction in feed intake, decreased milk yields, shortened lactation and reduced fertility. Genomic prediction of the phenotype is one of the most frequently stated problem in connection with genome wide association studies (GWAS) and genomic selection. Investigation on genomic prediction of hairy syndrome in cattle is important for our increased understanding of adaptive evolutions due to the climate change. This research examines the emerging role in the context of genomic prediction of hairy and slick condition of the cattle using both Bayesian learning (BL) (Bayesian ridge regression, Bayesian LASSO, Bayes A, Bayes B and Bayes C and machine learning (ML) methods (weighted subspace random forest (wRF), gradient boosted trees (GBT), naïve Bayes (NB) and K-nearest neighbors (KNN) under various experimental designs. The dataset included 99 (37 cases and 62 controls: defined by visual inspection of hairiness) crossbred cows genotyped for 712 222 SNPs. This study set out with the aim of assessing the importance of linkage disequilibrium, population structure, and various SNPs selection process (for only ML) for genomic prediction of hairiness using BL and ML models. The most obvious finding to emerge from this study is the superiority of ML model over BL models for genomic prediction of the phenotype. This study supports evidence from previous observations on beneficial usage of ML model in genetics and genomics research. The relevance of wRF is clearly supported by the current findings. The wRF with GWAS selected SNPs of 15000 gave the best prediction accuracy (with standard error in parenthesis, Area Under Curve=0.998 (0.004)). Despite its relatively small sampling size: these data suggest that ML prediction of hairiness can be achieved through high prediction accuracies hence the finding of this study have a number of important dairy cattle breeding implications for future practice in response to the climate change problem.

*Corresponding author: burakkaracaoren@akdeniz.edu.tr

KEYWORDS Bayesian learning / genomic prediction / hairy syndrome / machine learning

The issue of climate change has received considerable critical attention for negative impacts on cattle production [Nguyen *et al*. 2017]. There is a growing body of literature that recognizes the importance of understanding the adaptation of the cattle populations to the climate conditions [Berry 2018, Freitas *et al*. 2021]. Among others heat stress is one of the cause of reduction in feed intake, decreased milk yields, shortened lactation and reduced fertility [Nguyen *et al*. 2017]. Evidence suggests that adaptive evolutions (as such coat conformation) is among the most important factors for heat tolerance [Cai *et al*. 2021].There is evidence that a single mutation in prolactin (PRL) plays a crucial role in regulating hair length and coat structure in Senepol breed cows [Littlejohn et al. 2014] in connection with better heat tolerance and higher milk yield production. A study on Holstein population with slick hair condition by Dikmen *et al.* [2014] reports higher milk yield productions. Classical breeding/selection applications for introducing PRL mutations to artificially selected populations can play an important role in addressing the issue of the climate change.

However due to longer generation interval of cattle: classical breeding applications might be limited compared with recent molecular breeding applications. Genetic improvement experienced by genomic selection [Meuwissen *et al*. 2001] over the past decade remain unprecedented [Garner *et al.* 2016]. Genomic selection is important for a wide range of scientific and industrial applications in dairy cattle [Wiggans *et al*. 2017] due to longer generation intervals of cattle compared with most of the other farm species.

Genomic prediction of the phenotype is one of the most frequently stated problem in connection with genome wide association studies (GWAS) and genomic selection [Baker *et al*. 2020, Grinberg *et al*. 2020]. Investigation on genomic prediction of hairy syndrome in cattle is important for our increased understanding of adaptive evolutions due to the climate change. Several methods currently exist for the genomic prediction of the phenotypes [Abdollahi-Arpanahi *et al*. 2020]. The choice of the genomic prediction model could be assessed by level of linkage disequilibrium, effect of major genes, marker density, nonlinear interactions of the genes in the population [Baker *et al*. 2020]. The Bayesian regression models [Gianola 2013 Meuwissen *et al.* 2001] has a number of attractive features including: capable of marker specific prior distributions [Baker *et al*. 2020] and incorporation of linkage disequilibrium [Moser *et al*. 2015] for the genomic prediction. Recent advances in machine learning (ML) methods have facilitated investigation of genome wide associations data with binary phenotypes for genomic prediction and classification problems. However, there is a relatively small body of literature that uses ML algorithms for prediction of phenotypes using molecular genetic markers. It has been suggested that [Grinberg *et al.* 2020] ML methods for genomic prediction are independent of genetic model assumptions (additive gene effects, the number of genes and their interactions). One of the most significant current discussions in assumptions of ML is the effect of population stratification [Grinberg *et al*. 2020].

Baker *et al.* [2020] studied the prediction of disease statues in dogs using various ML and Bayesian Learning (BL) or genomic prediction/selection models. A significant genomic selection analysis and the discussion on the heat tolerance in dairy cattle was presented by [Carabano *et al*. 2019, Garner *et al*. 2016]. This research examines the emerging role in the context of genomic prediction of hairy and slick condition of the cattle [Littlejohn *et al*. 2014] using both BL and ML methods under various experimental designs including linkage disequilibrium (LD), population stratification (PS) and combinations of LD and PS.

Material and methods

Dataset and Quality Control

A two generation of pedigree was used for sampling the animals. The pedigree consisted 2274 animals with 2 affected sires families. The dataset included 99 (37 cases and 62 controls: defined by visual inspection of hairiness) crossbred cows of mixture of Senepol, Barzona, Red Angus, and Hereford ancestry genotyped for 712.222 SNPs [Littlejohn *et al*. 2014]. We applied different data selection and quality control procedures based on minor allele frequencies (<0.95), calling rate of SNPs (>0.90), Hardy-Weinberg equilibrium (P<1E-07), Linkage Disequilibrium (LD) (r^2>0.7) and genomic relations (inbreeding coefficient>0.09). Genomic relationship matrix and associated inbreeding coefficients were obtained by using genomic relationship matrix [VanRaden 2008] implemented in PLINK [Purcell *et al*. 2007]. Different methods have been proposed [VanRaden 2008] to obtain genomic relationship matrix in connection with genetic relationship among animals. We used

$$G = \frac{(Z)Z'}{2\sum_{j=1}^{m} p_j\ (1-p_j)}$$

equation: where Z is a design matrix equating mean values of the alles to 0 and $2\sum_{j=1}^{m} p_j\ (1-p_j)$ is a scaling factor with second allele frequency of p_j.

There are three main types of study design used to evaluate ML and BL methods: based on LD, population stratification (PS) and combination of LD and PS. Quantification of LD is one of the most common procedures for determining correlated SNPs over genome. Different authors measured the level of LD to obtain decorrelated version of the genotypic data in a variety of ways [Calus and Vandenplas 2018]. Previous studies have based their criteria for square of correlations (for example r^2>0.7) among SNPs [Baker *et al*. 2020, Calus and Vandenplas 2018, Grinberg *et al*. 2020] to obtain LD pruned data [Purcell *et al*. 2007]. A number of techniques have been developed the address PS based on genomic relationship matrix [Zhang *et al*. 2015]. The animals were selected on the bases of the degree of relatedness of their genomic kinship [Purcell *et al.* 2007].

Splitting the data as training (%80 of animals) and testing (%20 of animals) are currently the most popular methods for assessing model performance in ML and BL methods [Baker *et al.* 2020, Grinberg *et al.* 2020]. Area under the curve (AUC) approach was used to capture the accuracies over training and testing procedures by using 10 fold cross validations of ML and BL methods as was defined in [Baker *et al*. 2020]. All analyses were done with the statistical software R [2020] based on associated R code of Baker *et al.* [2020].

Bayesian Learning Models

BL and ML analyses was based on the conceptual framework proposed by [Baker *et al.* 2020]. BL were obtained for Bayesian ridge regression, Bayesian LASSO, Bayes A, Bayes B and Bayes C [Gianola 2013] using the BGLR package [Perez and de los Campos 2014] with 52000 Markov Chain Monte Carlo iterations by 6000 burn-in period.

$$y_i - \sum_{j=1}^{n}(z_{ij}a_j\delta_j) + e_i \qquad (1)$$

in which y_i is the binary phenotypes of the ith animal; z_{ij} is an indicator variable (depends on BL model) for the ith animal, jth SNP locus and kth allel; a_j is marker of locus effects; δj indicates whether SNP has an effect (or not); and e_i is the residual for animal i. BGLR package randomly simulates samples from the desired posterior density by Gibbs sampler with scalar updating [Perez and de los Campos 2014].

Bayesian ridge regression (BRR) assume each regression coefficient in model (1) shrinking towards to zero by common variance using independent Gaussian priors. Bayesian Least Absolute Shrinkage and Selection Operator (LASSO) uses stronger weights [Park and Casella 2008] for penalizing SNPs with small effects by employing Laplace prior distribution. Bayes A employs a scaled t prior distribution for SNP effects [Meuwissen *et al*. 2001] assuming many SNPs of small effect and few of major effect. Bayes B [Meuwissen *et al*. 2001] uses a mixture of two distributions (SNPs with and without effects) for predictions of SNP effects in model (1) assuming many SNPs with zero effect and few of with a t distribution of effects. Bayes Cπ [Habier *et al*. 2011] is comparable with Bayes B except: prediction of π parameter to detect proportion of the SNPs with effects on the phenotype.

Machine Learning models

Baker *et al.* [2020] proposed two methods for SNP selection (to be used in ML models) is based on (1) ranked P-values from a linear mixed model GWAS analyses [Perdry and Dandine-Roulland 2015] and (2) select SNPs solely based on ranking of mean differences of allele frequencies by comparing cases and controls. The performance of the ML models depended on the genetic architecture of the trait: therefore, different number of SNPs (5 to 15000 SNPs) were used for the ML analyses.

Different ML methods have been proposed [Baker *et al*. 2020] for genomic classification analyses of the binary phenotype: weighted subspace random forest (wRF), gradient boosted trees (GBT), naïve Bayes (NB) and K-nearest neighbors

(KNN). wRF is an method to explore relation among SNPs and binary phenotypes by recursive algorithm that aims to classify animals into clusters by reducing group heterogeneity [Zhao *et al*. 2017]. wRF was implemented by minimum 1000 trees and the square root of the number of features at each tree classification [Baker *et al.* 2020]. While wRF are well defined for tree discovery, associated another model GBT has advantages in terms of smaller mean square errors [Chen *et al*. 2015] with additional hyperparameter tuning for the parameter estimations. Hyperparameters was tuned with 5 fold cross validation based on analyses of [Baker *et al*. 2020]: learning rate (eta)=0.05, minimum loss reduction (gamma)=0.3, subsample ratio of columns when constructing trees=0.8, subsample ratio of training instances=0.8 with 1000 rounds of training. NB [Dimitriadou *et al*. 2017] is one of the most common ML algorithm based on Bayes theorem. One advantages of the NB classification is that it avoids the problem of computational burden compared with the other ML algorithms. The KNN [Kuhn 2008] was used to obtain classification model based on the whole SNP data. The advantages of KNN are that simplicity as the model do not use the training stage. However, there are certain drawbacks associated with the use of NB and KNN in genomic classification research due to huge number of features or inputs (SNPs).

Results and discussion

Quality control thresholds based on minor allele frequencies (<0.05), call rate of genotypes (>0.95) and Hardy Weinberg proportions (P<1E-07) were applied to SNP genotype data. A total of 573922 SNPs was identified after the quality control process. Square of correlations (r2) obtained from pairwise SNPs over windows size of 50 SNPs with a step size of 5 [Grinberg *et al*. 2020] at threshold 0.7 reduced the data to 197.186 SNPs (Fig. 1). After removing lowly correlated (genomic inbreeding coefficient >0.09) animals from the dataset, 61 animals remained to be used in PS analyses.

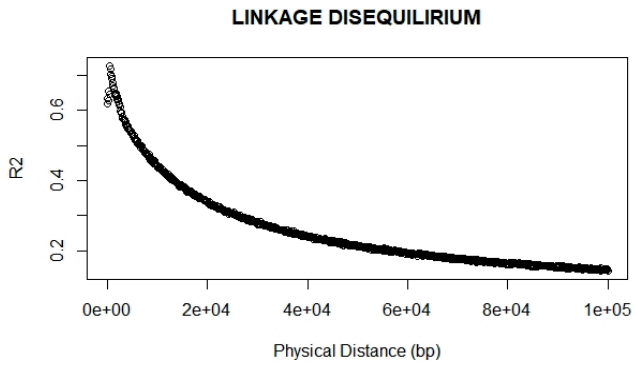


Fig. 1. Decay of linkage disequilibrium over physical distance.

Bayesian Learning analyses

As shown in Figure 2 the Bayesian analyses reported similar prediction accuracies (around to be 0.80) over different models and experimental design. Genotypes are characterized by LD decay with low variation (Fig. 1) over chromosomal base pair

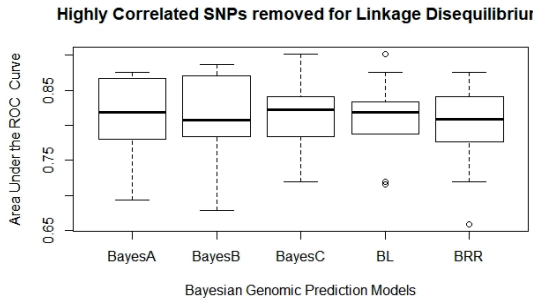


Fig. 2. Prediction accuracies obtained for Bayesian learning models from 10-fold cross validations with removing SNPs using linkage disequilibrium.

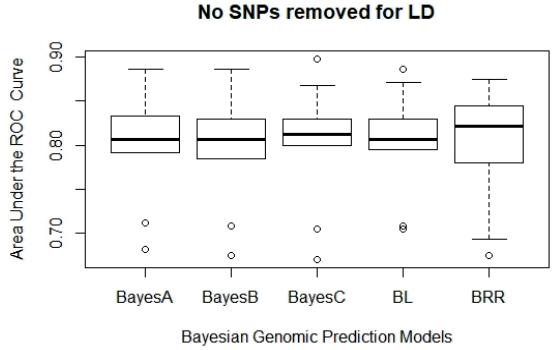


Fig. 3. Prediction accuracies obtained for Bayesian learning models from 10-fold cross validations without removing SNPs using linkage disequilibrium.

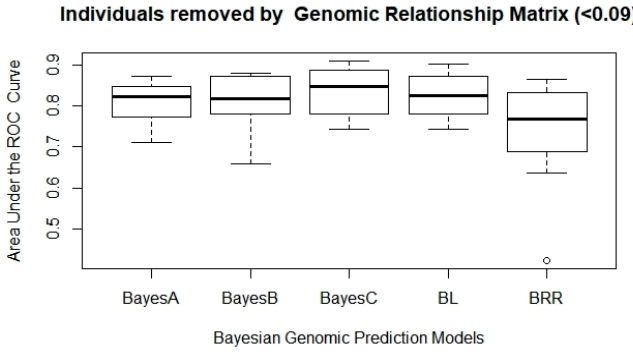


Fig. 4. Prediction accuracies obtained for Bayesian learning models from 10-fold cross validations based on individuals removed by genomic relationship matrix (<0.09).

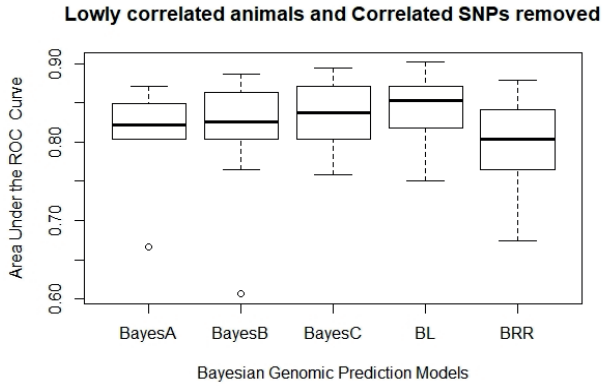


Fig. 5. Prediction accuracies obtained for Bayesian learning models from 10-fold cross validations with removing SNPs using linkage disequilibrium and lowly correlated animals.

locations. To investigate the effect of LD to different BL models, Figure 2 and 3 provides an overview of prediction accuracies with and without LD correction, PS (Fig. 4) and LD correction with PS (Fig. 5). There are little differences in the accuracies of Bayes C compared with other Bayesian models especially for LD pruning (accuracy of Bayes Cπ found to be 0.82) (Fig. 2), PS (accuracy of Bayes C found to be 0.82) (Fig. 4) and LD pruning with PS (accuracy of Bayes C and BL found to be 0.83) experimental designs (Fig. 5).

Machine learning analyses

Table 1 illustrates best prediction accuracies obtained from ML analyses under different experimental settings. There was evidence that LD has an influence on SNPs selection process using GWAS or mean differences among SNPs. A clear benefit of SNPs removal based on LD level for increasement of GP identified (Tab. 1) under various experimental designs. From Table 1 we can see that GWAS based SNPs selection resulted in the highest prediction accuracies over different experimental designs, except with the full genotypic dataset (Table 1: section "No SNPs removed for Linkage Disequilibrium).

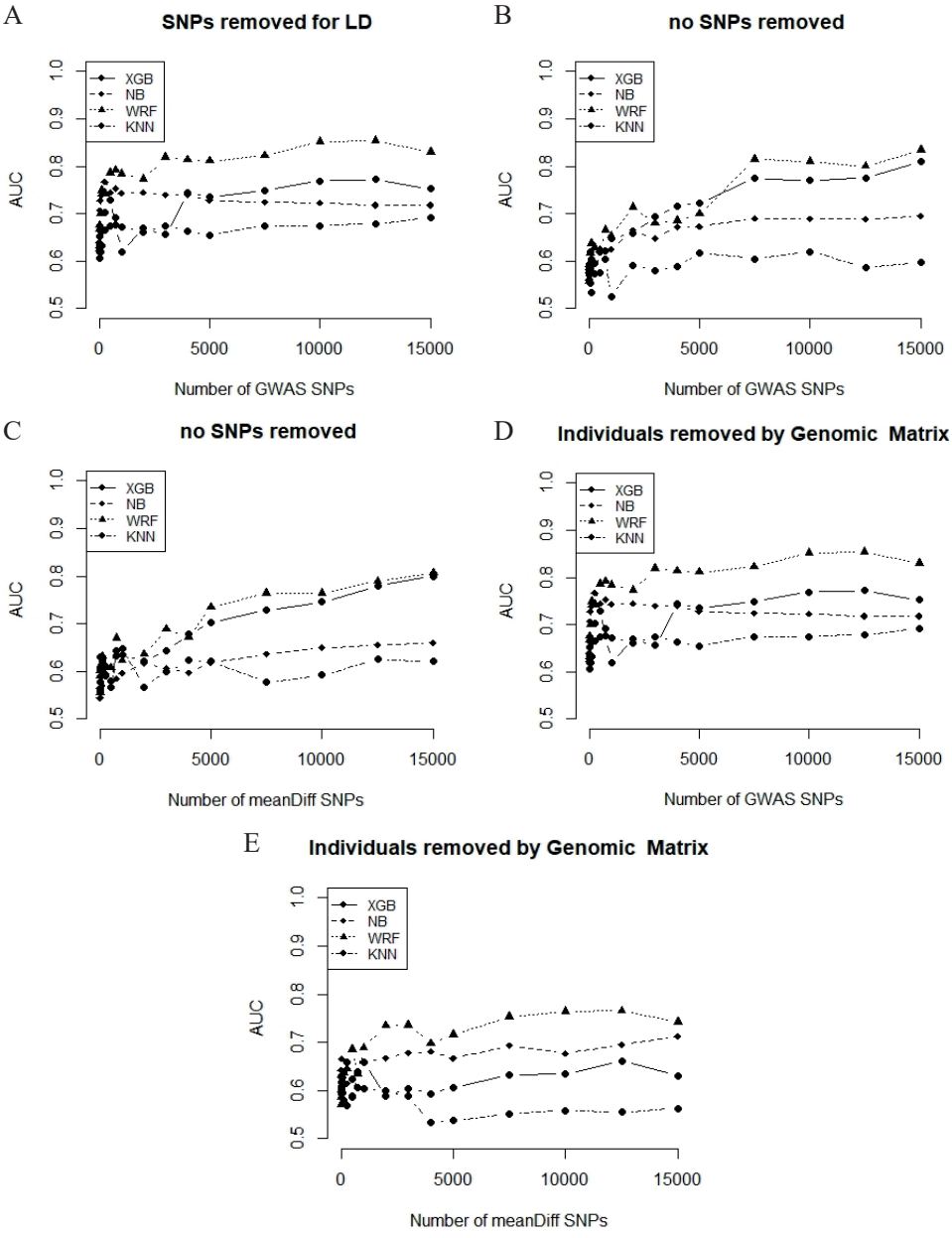
Means for model predictions over all folds reported for SNPs removed for LD, with full genotypic data, SNPs selected by GWAS and mean differences of allelic frequencies, animals removed by genomic relationship matrix and combinations of various experimental designs given in Figure 6: (A) Model comparison with LD pruning (r^2>0.7) using GWAS SNPs. (B) Model comparison without LD pruning using GWAS SNPs; (C) Model comparison without LD pruning using mean allelic differences of SNPs. (D) Model comparison with individuals removed by genomic relationship matrix (genomic inbreeding coefficient >0.09) using GWAS SNPs. (E) Model comparison with individuals removed by genomic relationship matrix (genomic inbreeding coefficient >0.09) using mean allelic differences of SNPs. (F)

Table 1. Best performing machine learning models obtained from different experimental settings over 10 fold cross validation procedure

Model	Feature Selection	No. SNPs	AUC (SD)
No SNPs removed for Linkage Disequilibrium			
wRF	GWAS	15000	0.807 (0.104)
	meanDiff	15000	0.834 (0.107)
GBT	GWAS	15000	0.800 (0.110)
	meanDiff	15000	0.810 (0.130)
NB	GWAS	15000	0.660 (0.074)
	meanDiff	15000	0.695 (0.091)
KNN	GWAS	1000	0.648 (0.069)
	meanDiff	10000	0.621(0.104)
Highly Correlated SNPs removed for Linkage Disequilibrium (r^2>0.7)			
wRF	GWAS	15000	0.998 (0.004)
	meanDiff	15000	0.807 (0.004)
GBT	GWAS	12500	0.994 (0.006)
	meanDiff	15000	0.800 (0.110)
NB	GWAS	100	0.734 (0.058)
	meanDiff	15000	0.660 (0.074)
KNN	GWAS	5000	0.649 (0.109)
	meanDiff	1000	0.648 (0.069)
Lowly Correlated Individuals removed by using Genomic Relationship Matrix (>0.09)			
wRF	GWAS	12500	0.854 (0.108)
	meanDiff	12500	0.766 (0.173)
GBT	GWAS	12500	0.772 (0.099)
	meanDiff	12500	0.661 (0.138)
NB	GWAS	250	0.767 (0.061)
	meanDiff	15000	0.712 (0.076)
KNN	GWAS	15000	0.692 (0.157)
	meanDiff	25	0.631 (0.110)
Lowly Correlated Individuals (<0.09) and highly correlated SNPs (r^2>0.7) removed			
wRF	GWAS	15000	0.922 (0.044)
	meanDiff	15000	0.804 (0.099)
GBT	GWAS	15000	0.958 (0.027)
	meanDiff	7500	0.707 (0.172)
NB	GWAS	250	0.780 (0.060)
	meanDiff	15000	0.751 (0.086)
KNN	GWAS	1000	0.694 (0.125)
	meanDiff	250	0.639 (0.100)

Model comparison with LD pruning (r^2>0.7) and individuals removed by genomic relationship matrix (genomic inbreeding coefficient >0.09) using GWAS SNPs. (G) Model comparison with individuals removed by genomic relationship matrix (genomic inbreeding coefficient >0.09) using mean allelic differences of SNPs. (F) Model comparison with LD pruning (r^2>0.7) and individuals removed by genomic relationship matrix (genomic inbreeding coefficient >0.09) using mean allelic differences of SNPs.

From Table 1 it can be seen that wRF resulted highest accuracies under most of the experimental settings. As shown in Figure 6, the wRF with GWAS selected SNPs of 15000 gave the best prediction accuracy (AUC=0.998(0.004)). LD pruning and



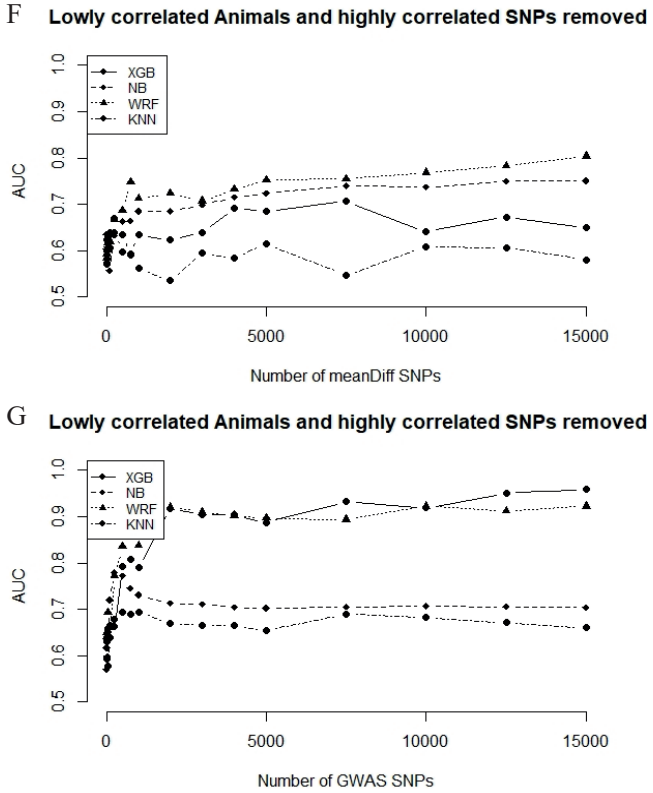


Fig. 6. Prediction accuracies obtained from 10-fold cross validation with models trained from 5 to 15000 SNPs under various experimental designs.

removal of lowly correlated animals increased the AUC values of wRF and GBT (Tab. 1 and Fig. 6). Prediction accuracies did not improve as more SNPs were included to the NB and KNN (Fig. 6).

In reviewing the literature, no data was found on genomic prediction of hairiness using ML and BL methods in dairy cattle. The present study was designed to determine the effect of genetic architecture by using different number of SNPs, SNP selection criteria's and various ML and BL model for take into underlying gene actions for predicting the phenotype. BL models with different prior distributions [Gianola 2013] was designed to determine the effect of various genetical architecture scenarios: ranging from polygenic model to major gene effects. In accordance with the increased prediction accuracies obtained from Bayes Cπ (Fig. 1) previous studies [Dikmen *et al.* 2013, Dikmen *et al*. 2014, Karacaören 2016, Littlejohn *et al*. 2014] have shown that some polygenes are associated with hairiness and heat stress. This outcome is slightly contrary to that of [Baker *et al*. 2020] who did not find prediction accuracy differences

over different BL models. In addition, the prediction accuracies in this investigation were higher compared with the BL results of [Baker *et al*. 2020]. There are several possible explanations for this result. The observed increase in prediction accuracies in general (and particularly for Bayes Cπ) could be attribute to conjugation of genetic architecture of our phenotype and corresponding BL model assumptions. Baker *et al*. [2020] showed that their phenotype showed polygenic inheritance and did not match with the prior assumptions of their BL models.

To investigate the association between the selected number of SNPs chosen by the GWAS or mean differences: we reported the prediction accuracies of 10-fold cross validation process from ML analyses (Tab. 1). We observed (Fig. 3) that the prediction accuracy of the SNPs obtained from GWAS (based on P values) found to be higher, however the finding of the current study do not support the results of [Baker *et al*. 2020]. It has been suggested that SNPs filtering using mean differences should be advantageous [Baker *et al*. 2020] in case/control design but this was not the case in our results (Tab. 1). This inconsistency may be due to differences between genetical architecture underlying the phenotypes of the current study and the phenotype of [Baker *et al.* 2020]. Our results corroborate the ideas of Karacaoren [2016] and Littlejohn *et al*. [2014] who reported mutations and associated SNPs with the hairiness in dairy cattle. Hence, it could conceivably be hypothesised that when there is a major gene in association with the phenotype: SNPs selection using results of GWAS could be more informative.

This study set out with the aim of assessing the importance of LD for genomic prediction of hairiness using BL and ML models. Consistent with the literature [Baker *et al*. 2020, Grinberg *et al*. 2020] this research found that usage of LD pruned data with wRF and GBT increased the prediction accuracies (Tab. 1). Generally, prediction accuracies are higher when using wRF and GBT compared with NB and KNN (Tab. 1). These results are in accord with recent studies indicating that NB and KNN might have misclassification problem [Baker *et al.* 2020] with huge number of SNPs. It can be seen from the data in Table 1 that the NB and KNN reported significantly different results under different SNP selection criteria's more than the other wRF and GBT. However, with successive decreases in number of SNPs (Figure 3) the prediction accuracies found to be higher in NB and KNN compared with wRF and GBT in all experimental designs (Tab. 1). This finding is partly consistent with that of Srivastaka *et al.* [2021] who compared the genomic predictive performance of RF, GBT and support vector machines for various phenotypes of Hanwoo cattle. From Table 1 we can see that NB with only 100 GWAS selected SNPs (in linkage equilibrium) resulted with 0.734 (0.058) prediction accuracy. When genomic relationships were considered, the prediction accuracy of KNN was found to be 0.631 (0.110) with only 25 SNPs selected based on mean differences of cases and controls (Fig. 3). Taken together, these results suggest that both NB and KNN could be useful in genomic prediction with smaller number of informative SNPs in expense of lower prediction accuracies compared with wRF and GBT.

Increased prediction accuracies in the ML models corroborates with the results of [Abdollahi-Arpanahi *et al*. 2020]. A possible explanation for better results obtained by wRF and GBT might be related with the genetic architecture of the hairiness. Since hairiness was found to be associated with small number of SNPs [Karacaoren 2016, Littlejohn *et al*. 2014], with possibly nonlinear gene actions, wRF found to be the superior model in terms of prediction accuracy. These results are in agreement with Grinberg *et al.* [2020] findings which showed the superiority of random forest under different experimental designs. Further these results reflect those of Li *et al*. [2018] who also found higher accuracies by random forest algorithms due to captured nonlinear gene actions in Brahman cattle. There are similarities between the increased prediction accuracies obtained by wRF and those described by Fitzpatrick *et al.* [2021] employed various random forest algorithms for associating genomic information into climate change impact assessments.

The present study was designed to determine the effect of various experimental design to prediction accuracies of BL and ML models for genomic prediction of hairiness in dairy cattle including LD pruning, PS and LD and PS. The most obvious finding to emerge from this study is the superiority of ML model over BL models for genomic prediction of the phenotype. The results of this investigation show the importance of LD pruning in GP using ML. This study supports evidence from previous observations [Nicholls *et al*. 2020] on beneficial usage of ML model in genetics and genomics research. The relevance of wRF is clearly supported by the current findings. Despite its relatively small sampling size: these data suggest that ML prediction of hairiness can be achieved through high prediction accuracies hence the finding of this study have a number of important dairy cattle breeding implications for future practice in response to the climate change problem.

REFERENCES

1. ABDOLLAHI-ARPANAHI R., GIANOLA D., PEÑAGARICANO, F., 2020 - Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. ***Genetics Selection Evolution*** 52, 1, 1-15.
2. BAKER L.A., MOMEN M., CHAN K., BOLLIG N., LOPES F.B., ROSA G.J.M., TODHUNTER R.J., BINVERSIE E.E., SAMPLE S.J., MUIR P., 2020 - Bayesian and machine learning models for genomic prediction of anterior cruciate ligament rupture in the canine model. ***G3: Genes, Genomes, Genetics*** 10, 8, 2619-2628.
3. BERRY D.P., 2018 - Symposium review: Breeding a better cow-Will she be adaptable? ***Journal of Dairy Science*** 101, 4, 3665-3685.
4. CAI C., HUANG B., QU K., ZHANG J., LEI C., 2021 - A novel missense mutation within KRT75 gene strongly affects heat stress in Chinese cattle. ***Gene*** 768, 145294.
5. CALUS, M.P.L., VANDENPLAS J., 2018 - SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. ***Genetics Selection Evolution*** 50, 1, 1-11.
6. CARABAÑO M.J., RAMON M., MENENDEZ-BUXADERA A., MOLINA A., DIAZ C., 2019 - Selecting for heat tolerance. ***Animal Frontiers*** 9, 1, 62-68.

7. DIKMEN S., COLE J.B., NULL D.J., HANSEN P.J., 2013 - Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. ***PloS One*** 8, 7, e69202.
8. DIKMEN S., KHAN F.A., HUSON H.J., SONSTEGARD T.S., MOSS J.I., DAHL G.E., HANSEN P.J., 2014 - The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. ***Journal of Dairy Science*** 97, 9, 5508-5520.
9. FITZPATRICK M.C., CHHATRE V.E., SOOLANAYAKANAHALLY R.Y., KELLER S.R., 2021 - Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. ***Molecular Ecology Resources*** 21, 2749-2765
10. FREITAS P.H., WANG Y., YAN P., OLIVERIA H.R., SCHENKEL F.S., ZHANG Y., XU Q., BRITO L.F., 2021 - Genetic Diversity and Signatures of Selection for Thermal Stress in Cattle and Other Two Bos Species Adapted to Divergent Climatic Conditions. ***Frontiers in Genetics*** 12, 102.
11. GARNER J.B., DOUGLAS M.L., WILLIAMS S.O., WALES W.J., MARETT L.C., NGUYEN T.T.T., REICH C.M., HAYES B.J., 2016 - Genomic selection improves heat tolerance in dairy cattle. ***Scientific Reports*** 6, 34114.
12. GIANOLA D., 2013 - Priors in whole-genome regression: the Bayesian alphabet returns. ***Genetics*** 194, 3, 573-596.
13. GRINBERG N.F., ORHOBOR O.I., KING, R.D., 2020 - An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. ***Machine Learning*** 109, 2, 251-277.
14. HABIER D., FERNANDO R.L., KIZILKAYA K., GARRICK D.J., 2011 - Extension of the Bayesian alphabet for genomic selection. ***BMC Bioinformatics*** 12, 1, 1-12.
15. KARACAOREN B., 2016 - Investigations on Genetic Architecture of Hairy Loci in Dairy Cattle by Using Single and Whole Genome Regression Approaches. ***Asian-Australasian journal of Animal Sciences*** 29, 7, 938.
16. KUHN M., 2008 - Building predictive models in R using the caret package. ***Journal of Statistical Software*** 28, 1, 1-26.
17. 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. ***Frontiers in Genetics*** 9, 237.
18. LITTLEJOHN M.D., HENTY K.M., TIPLADY K., JOHNSON T., HARLAND C., LOPDELL T., SHERLOCK R.G., LI W., LUKEFAHR S.D., SHANKS B.C., GARRICK D.J., SNELL R.G., SPELMAN R.J., DAVIS S.R., 2014 - Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. ***Nature Communications*** 5, 1, 1-8.
19. 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.
20. MOSER G., LEE S.H., HAYES B.J., GODDARD M.E., WRAY N.R., VISSCHER P.M., 2015 - Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. ***PLoS Genetics*** 11, 4, e1004969.
21. NGUYEN T.T., BOWMAN P.J., HAILE-MARIAM M., NIEUWHOF G.J., HAYES B.J., PRYCE J.E., 2017 - Implementation of a breeding value for heat tolerance in Australian dairy cattle. ***Journal of Dairy Science*** 100, 9, 7362-7367.
22. NICHOLLS H.L., JOHN C.R., WATSON D.S., MUNROE P.B., BARNES M.R., CABRERA C.P., 2020 - Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. ***Frontiers in Genetics*** 11, 350.
23. PARK T., CASELLA G., 2008 - The Bayesian lasso. ***Journal of the American Statistical Association*** 103, 482, 681-686.
24. PEREZ P., de los CAMPOS G., 2014 - Genome-wide regression and prediction with the BGLR statistical package. ***Genetics*** 198, 2, 483-495.

25. PURCELL S., NEALE B., TODD-BROWN K., THOMAS L., FERREIRA M.A., BENDER D., MALLER J., SKLAR P., de BAKKER P.I., DALY M. J., SHAM P.C., 2007 - PLINK: a tool set for whole-genome association and population-based linkage analyses. ***The American journal of Human Genetics*** 81, 3, 559-575.
26. R Development Core Team., 2020 - A language and environmental for statistical computing. R Foundation for Statistical Computing; Vienna, Austria.
27. Srivastava S, Lopez BI, Kumar H, Jang M, Chai H-H, Park W, Park J-E, Lim D., 2021 - Prediction of Hanwoo cattle phenotypes from genotypes using machine learning methods. ***Animals*** 11(7), 2066.
28. VANRADEN P.M., 2008 - Efficient methods to compute genomic predictions. ***Journal of Dairy Science*** 91, 11, 4414-4423.
29. WIGGANS G.R., COLE J.B., HUBBARD S.M., SONSTEGARD T.S., 2017 - Genomic selection in dairy cattle: the USDA experience. ***Annual Review of Animal Biosciences*** 5, 309-327.
30. ZHANG Q., CALUS M.P., GULDBRANDTSEN B., LUND M.S., SAHANA G., 2015 - Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. ***BMC Genetics*** 16, 1, 1-11.
31. ZHAO H., WILLIAMS G.J., HUANG J.Z., 2017 - WSRF: an R package for classification with scalable weighted subspace random forests. ***The Journal of Statistical Software*** 77, 3, 1-30.