frontiers in Genetics

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fgene.2021.611506

Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning

Miriam Piles[1], Rob Bergsma[2], Daniel Gianola[3,4], Hélène Gilbert[5] and Llibertat Tusell[1,5†]*

[1]Animal Breeding and Genetics Program, Institute of Agriculture and Food Research and Technology (IRTA), Barcelona, Spain, [2]Topigs Norsvin Research Center, Beuningen, Netherlands, [3]Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, United States, [4]Department of Dairy Science, University of Wisconsin-Madison, Madison, WI, United States, [5]GenPhySE, INRAE, Université de Toulouse, Castanet-Tolosan, France

OPEN ACCESS

Edited by:
Luis Varona, University of Zaragoza, Spain

Reviewed by:
Paulino Pérez-Rodríguez, Colegio de Postgraduados (COLPOS), Mexico
Ismo Strandén, Natural Resources Institute Finland (Luke), Finland

***Correspondence:**
Miriam Piles
miriam.piles@irta.es

†Present address:
Llibertat Tusell
Animal Breeding and Genetics Program, Institute of Agriculture and Food Research and Technology (IRTA), Barcelona, Spain

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *29 September 2020*
Accepted: *20 January 2021*
Published: *22 February 2021*

Citation:
Piles M, Bergsma R, Gianola D, Gilbert H and Tusell L (2021) Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. Front. Genet. 12:611506. doi: 10.3389/fgene.2021.611506

Feature selection (FS, i.e., selection of a subset of predictor variables) is essential in high-dimensional datasets to prevent overfitting of prediction/classification models and reduce computation time and resources. In genomics, FS allows identifying relevant markers and designing low-density SNP chips to evaluate selection candidates. In this research, several univariate and multivariate FS algorithms combined with various parametric and non-parametric learners were applied to the prediction of feed efficiency in growing pigs from high-dimensional genomic data. The objective was to find the best combination of feature selector, SNP subset size, and learner leading to accurate and stable (i.e., less sensitive to changes in the training data) prediction models. Genomic best linear unbiased prediction (GBLUP) without SNP pre-selection was the benchmark. Three types of FS methods were implemented: (i) filter methods: univariate (univ.dtree, spearcor) or multivariate (cforest, mrmr), with random selection as benchmark; (ii) embedded methods: elastic net and least absolute shrinkage and selection operator (LASSO) regression; (iii) combination of filter and embedded methods. Ridge regression, support vector machine (SVM), and gradient boosting (GB) were applied after pre-selection performed with the filter methods. Data represented 5,708 individual records of residual feed intake to be predicted from the animal's own genotype. Accuracy (stability of results) was measured as the median (interquartile range) of the Spearman correlation between observed and predicted data in a 10-fold cross-validation. The best prediction in terms of accuracy and stability was obtained with SVM and GB using 500 or more SNPs [0.28 (0.02) and 0.27 (0.04) for SVM and GB with 1,000 SNPs, respectively]. With larger subset sizes (1,000–1,500 SNPs), the filter method had no influence on prediction quality, which was similar to that attained with a random selection. With 50–250 SNPs, the FS method had a huge impact on prediction quality: it was very poor for tree-based methods combined with any learner, but good and similar to what was obtained with larger SNP subsets when spearcor or mrmr were implemented with or without embedded methods. Those filters also led to very stable results, suggesting their potential use for designing low-density SNP chips for genome-based evaluation of feed efficiency.

Keywords: feature selection, stability, machine learning, genomic prediction, SNP, pigs, feed efficiency and growth

INTRODUCTION

Statistical models and methods used for predicting phenotypes or breeding values of selection candidates have an impact on the efficiency of genomic selection (GS). Machine learning (ML) methods are appealing for genomic prediction; they encompass a wide variety of techniques and models to predict outputs or to identify patterns in large datasets. Those methods do not require assumptions about the genetic determinism underlying the trait. ML is increasingly used in situations where the number of parameters is much larger than the number of observations, as it is the case for high-density genetic markers for GS. Thus, in animal and plant breeding, ML models that are non-linear in either features or parameters have been proposed to enhance genome-enabled prediction of complex traits (Gianola et al., 2006, 2011; Gianola and van Kaam, 2008).

Feature selection (i.e., selection of a subset of predictor variables, also known as features, from the input data; FS) reduces computation requirements and prevents over-fitting which occurs with high-dimensional data (Chandrashekar and Sahin, 2014). In addition, when features have a high level of redundancy, different training samples can produce different feature ranks (and therefore different models when a subset of features is selected) with the same prediction accuracy. In genetic studies, the stability of FS methods or "preferential stability" (i.e., the agreement of prediction models produced by an algorithm when trained on different training sets) is important to understand biological processes involved in the trait of interest and to design small low-cost prediction chips for GS or diagnostic (Phuong et al., 2005; Bermingham et al., 2015; Alzubi et al., 2018). Overall, it is wished to achieve a good prediction performance on independent data sets and a stable possible set of predictors, this being understood as those less sensitive to changes in the training set.

A review of FS methods can be found in Saeys et al. (2007), Chandrashekar and Sahin (2014), and Venkatesh and Anuradha (2019). All FS methods take as input a matrix of predictor variables (e.g., SNP genotypes or microarrays) for a set of samples with different output or target (i.e., the phenotype) and return a set of selected features of user-defined or tuned size. FS methods can be classified into three groups of methods: wrapper, embedded, and filter (Guyon and Elissee, 2003). Wrapper methods fit a supervised learning model using different subsets of the whole set of features, which are evaluated by a performance measurement calculated on the resulting model. Examples of wrapper methods are evolutionary FS algorithms and recursive feature elimination methods (Samb et al., 2012). Most wrapper methods are computationally infeasible for high-dimensional data sets (Saeys et al., 2007). Embedded methods perform FS as part of the model construction/fitting procedure. Some examples are least absolute shrinkage and selection operator (LASSO) regression, elastic net (Zou and Hastie, 2005), and tree-based methods (Strobl et al., 2008; Waldmann, 2016). Finally, filter methods compute a score for each feature independently of the learning algorithm and then select a set of a fixed number of them (which can be optimized) with the highest scores or those that exceed a defined threshold. Filter methods can be combined with any kind of predictive method, even methods with embedded FS. Filter methods can be univariate if they do not consider interactions between features or multivariate if they do so. As they do not rely on learning algorithms, filter methods avoid overfitting and are computationally less demanding than wrapper and embedded methods. However, using univariate methods it is possible to select redundant variables and discard features that are informative when combined with others but less informative on their own.

Measures that have been used to quantify FS stability can be classified into similarity-based and frequency-based measures (Nogueira et al., 2018). Similarity-based estimators measure stability over all pairs of feature subsets (e.g., Generalized Kalousis estimator, Kalousis et al., 2007), whereas frequency-based estimators measure stability by the frequencies of selection of each feature over the feature sets (e.g., the relative weighted consistency; Somol and Novovicova, 2010). Recently, Nogueira et al. (2018) established five desirable properties a stability estimator must have. These properties are: (i) to allow variation in the number of features selected; ii) to be a decreasing function of the variable sample variances; (iii) to be upper/lower bounded by constants not dependent on the number of features selected; (iv) to achieve its maximum only when all selected feature sets across training sets are identical and; (v) to be corrected for chance. After concluding that none of the existing stability estimators possesses all these properties, they proposed a novel one meeting these properties and provided confidence intervals and hypothesis tests on stability, which is crucial for proper comparison among FS algorithms.

Several studies from different domains have compared the predictive performance of FS methods combined with classification or prediction methods using experimental or simulated data sets. However, only few used genomic data and none evaluated the stability of the performance of the FS algorithm. For example, Gunavathi et al. (2017) and Bolón-Canedo et al. (2014) compared classification accuracy of different FS methods based on microarrays datasets. Bommert et al. (2020) compared some of the most prominent types of filter methods for FS in terms of accuracy and computing time across 16 high-dimensional classification datasets, including microarray data. The best FS methods differed among datasets, so they recommended testing several ones in each specific analysis.

The objective of our research was to explore the influence of various combinations of FS methods and learners on prediction quality and stability of models for predicting residual feed intake (RFI) from SNP genotypes, in order to find the best strategy for genetic evaluation of growing pigs at reduced genotyping cost.

MATERIALS AND METHODS

The data used was from an existing database made available by Topigs Norsvin (Beuningen, Netherlands). No Animal Care Committee approval was necessary for our purposes.

Animals

Animals were 5,708 boars from a terminal sire line originated from 217 boars and 1,120 sows from Topigs Norsvin (Beuningen, Netherlands). All animals were born and raised in two Specific Pathogen Free nucleus farms, located in Netherlands and France, with semen exchange between farms being frequent.

Phenotypes

Nucleus farms were equipped with IVOG feeding stations (INSENTEC, Marknesse, Netherlands) that register individual feed intake of group housed pigs. All pigs had ear tags with unique numbering; individual feed intake records were available for all pigs for each day on the test. The pigs had *ad libitum* access to water and to a commercially available diet until the end of the performance test.

Average daily gain (ADG) was measured between the beginning (median age of 68 days and median weight of 31 Kg) and end of the test (median age of 155 days and median weight of 130 Kg). Only records from boars starting the test period between 50 and 105 days of age and remaining on the test between 60 and 120 days were retained.

Backfat thickness (BFT) was determined ultrasonically on live animals (US-fat in mm) at the end of the test period. Metabolic weight (MW; g) was calculated as: $MW = \left(\frac{W_{start} + W_{end}}{2}\right)^{0.75}$, where W_{start} and W_{end} are the weights at the beginning and end of the test period, respectively.

Multivariate outlier records of ADG, daily feed intake (DFI), BFT, and MW were identified and removed within batch and farm when the squared Mahalanobis distance away from the center of the distribution was >12 (Drumond et al., 2019). Then, RFI was estimated as the residual of a phenotypic linear regression of DFI on ADG, BFT, and MW. Thus, for animal *i*th:

$$\mathrm{DFI_i} = \beta_1 \times \mathrm{ADG_i} + \beta_2 \times \mathrm{BFT_i} + \beta_3 \times \mathrm{MBW_i} + \mathrm{RFI_i}$$

Subsequently, RFI records were pre-adjusted by macro-environmental effects fitting a linear model which included the fixed effects of age at the start of the test (Age, covariate), duration of the performance test (Length, covariate), and the combination of farm and batch (FarmBatch, 46 levels). The FarmBatch effect resulted from the combination of two farms and 2-month period batches. All FarmBatch levels retained for the analyses had at least 10 records. Thus, for animal *i*th and level of FarmBatch *j*th:

$$\mathrm{RFI_{ij}} = \mathrm{FarmBatch}_j + \beta_1 \times \mathrm{Age_i} + \beta_2 \times \mathrm{Length_i} + \mathrm{e_{ij}}$$

The adjusted RFI records were obtained after subtracting the estimates of these systematic environmental effects from the values of the original trait. From here onwards, we refer to adjusted RFI (i.e., e_{ij}) as RFI. Both linear models were fitted using the using lm() function (R Development Core Team, 2020).

Genotypes

Animals were genotyped using the Illumina Porcine SNP60 BeadChip (Illumina Inc., San Diego). Assuming an additive allele substitution effect, genotypes were arbitrarily coded to 0, 1, and 2 for the homozygote for the minor allele, heterozygote, and other homozygote, respectively. SNPs with a call rate lower than 0.90 and a minor allele frequency lower than 0.05 were removed. Boars with a call rate lower than 0.90 and parent-offspring pairs that displayed Mendelian inconsistencies were discarded. After this quality control, 46,610 SNPs were retained to pursue the analyses. Zero and near-zero-variance SNPs were identified and removed with the "nearZeroVar" function which removes predictors that have a unique value or have very few unique values relative to the number of samples, and the ratio of the frequency of the most common value to the frequency of the second most common value is 95/5 (Caret R package, Kuhn, 2008). Subsequently, the "findCorrelation" function (Caret R package, Kuhn, 2008) with a cut-off = 0.8 was used to diminish high pair-wise correlations between features. After this genotype edition, 9,523 SNPs were retained.

Model Fitting and Prediction

Models were fitted using individual genotypes as predictor variables and individual RFI records as output or target variable. For each combination of prediction method (i.e., learner) and SNP subset size, model fitting and (hyper)parameter optimization were conducted with a nested cross-validation (**Figure 1**). Nested cross-validation allows estimating the generalization error of the underlying model and its (hyper)parameter search (Bischl et al., 2016). It consists of several training-validation and testing dataset splits. An outer 10-fold cross-validation using all data was performed using nine equal-size parts of the original data sets for training the model, and the remaining one for testing. Within each outer training set, features (i.e., SNPs) were standardized and FS was performed using several methods and for a varying number of selected features (50, 250, 500, 750, 1,000, and 1,500). Also, within each outer training set, an inner six-fold cross-validation was implemented for tuning the hyper-parameters of the model. Hyper-parameter values were chosen based on a mean square error on the validation set of this inner cross-validation. The model was finally fitted to the whole training set using the optimal hyper-parameters. Same data split (i.e., same data subsets) was used across combinations of learners and datasets to compare prediction performance in the same conditions regarding data structure and composition.

In what follows, a description of the FS methods and learners implemented is provided first. Then, measurements of quality and stability of the predictions and of the selected features are defined.

Feature Selection Methods

Several filter and embedded FS methods (Saeys et al., 2007), and combinations hereof were used to rank and select the most relevant SNPs for predicting the target trait.

Filter Methods

The filter methods implemented were either univariate or multivariate to account for interactions between features. Consider a dataset of N records of p features (i.e., predictor variables,

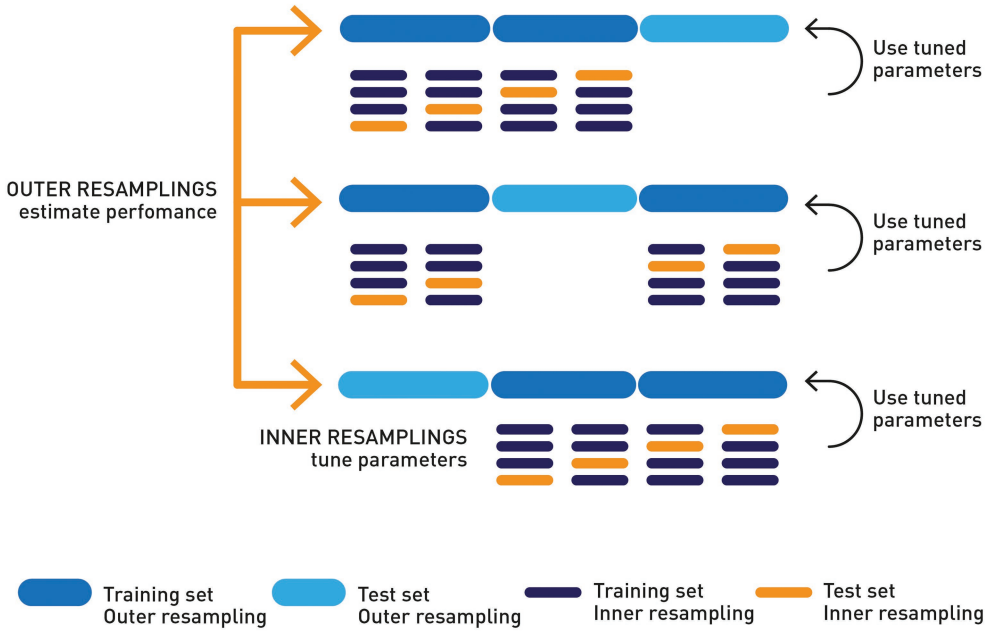

FIGURE 1 | Nested resampling diagram.

SNPs in this study which are considered to be continuous variables) in a set S of features and a target variable of interest Y. Five filter methods were implemented to rank the available SNPs according to their relevance for prediction of the target trait (Bommert et al., 2020): (i) Sort features with the Spearman's rank correlation (spearcor) between each feature X and Y. (ii) Univariate decision tree (univ.dtree) resamples a decision tree for each feature individually. The resampling performance is used as a filter score to rank features. (iii) Maximal relevance minimal redundancy filter (mrmr) is based on the concept of mutual information of two variables, defined as $I(Y;X) = H(Y) - H(Y|X)$ where $H(Y) = -\int f(Y)\log f(Y)dy$ is the differential entropy and $H(Y|X) = -\int\int f(Y,X)\log f(Y|X)dy\,dx$ is the conditional differential entropy (Ding and Peng, 2005). The entropy measures the uncertainty of the variable. The mutual information of two variables can be interpreted as the decrease in uncertainty about Y conditional on knowing X or as the amount of information shared by both variables since $I(Y;X) = I(X;Y)$. Filter mrmr uses the score $I(Y;X_k) - \frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j)$ where term $I(Y;X_k)$ measures the relevance of the kth feature *via* the information this feature has about Y, while the term $\frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j)$ measures its redundancy by the mean information that the feature shares with other jth features in the set S of size $|S|$. Therefore, the variables with the highest values of the score are those that have the maximum relevance and minimum redundancy with the other variables. (iv) Conditional permutation importance for correlated predictors (Strobl et al., 2008) of fitted random forest (cforest) uses a randomly permuted feature X_k to predict the response and to evaluate the difference in prediction accuracy before and after permuting that predictor X_k. If the original X_k variable is associated with the response, this permutation will lead to a decrease in prediction accuracy. Its advantage over univariate screening methods is that it covers the impact of each predictor variable individually, as well as in multivariate interactions with the other predictor variables. Conditional permutation importance (Strobl et al., 2008) was chosen to rank the markers because of the existence of correlation patterns among them. In this method, the permutation is performed within groups of observations that are defined by the values of the remaining predictor variables. (v) Random selection of SNPs (random), used as benchmark.

Embedded Methods

In the embedded methods the search for an optimal subset of features is done within the prediction model. Like wrapper methods, they are specific to a learning algorithm but less computationally demanding. Embedded methods used here were LASSO regression (LR, Park and Casella, 2008) and elastic net (ENET, Zou and Hastie, 2005), as explained in the section below.

Learners

Ridge regression (RR), LR, ENET, support vector machine for regression (SVM), and gradient boosting (GB) were used for predicting RFI records. Genomic best linear unbiased prediction (GBLUP, VanRaden, 2008) was used as benchmark.

Elastic net (Zou and Hastie, 2005) is originally a regression method that combines *L1* ($\lambda_1 \times \left[\sum_{j=1}^{p} \beta_j^2\right]$) and *L2* ($\lambda_2 \times \left[\sum_{j=1}^{p} |\beta_j|\right]$) penalties of ridge and LASSO in a mixture of the two. Parameters λ_1 and λ_2 control the strength of the *L1* and *L2* penalties and β_j is the regression coefficient on SNP *j*th. The ENET penalty is: $\lambda \times \left((1-\alpha) \times \left[\sum_{j=1}^{p} |\beta_j|\right] + \alpha \times \left[\sum_{j=1}^{p} \beta_j^2\right]\right)$, where $\alpha = (\frac{\lambda_2}{\lambda_1 + \lambda_2})$.

Thus,

$$\hat{\beta} = \arg min_{\beta} \left(\left\| \mathbf{y} - \mathbf{X}\beta \right\|^2 \right) + \lambda \times \left((1-\alpha) \times \left[\sum_{j=1}^{p} |\beta_j| \right] + \alpha \times \left[\sum_{j=1}^{p} \beta_j^2 \right] \right),$$

where $\mathbf{y}$ is the vector of adjusted phenotypes of dimension $N \times 1$, $\mathbf{X}$ is the matrix of standardized genotypes of dimension $N \times p$ and $\beta = \{\beta_j\}$ is the vector of regression coefficients of dimension $p \times 1$. The genotypes were standardized to have a mean of 0 and a standard deviation of 1.

Elastic net is an embedded method of FS because it allows selecting a subset of predictor variables out of *p* candidates. When $N \ll p$, ENET can select more than *N* predictor variables. This learner was implemented with the various SNP subset sizes, and with the full SNP set as well (9,523 SNPs). The function "cv.glmnet" from the "glmnet" R package (Friedman et al., 2010) was used to fit ENET. The value of α was tuned by testing values from 0 (i.e., a LR model) to 1 (i.e., a RR model) in increments of 0.1; the optimal λ parameter was found by cross-validation. Variable importance in the different fitted models was measured as the regression coefficients of each standardized predictor variable.

Support vector machine aims at identifying a function that has a maximum deviation ε from the observed values (*Y*) and has a maximum margin for the set of prediction variables (*S*). A review of this method can be found in Smola and Schölkopf (2004). The power of the SVM resides in a particular component known as kernel. One of the most used kernel is the Gaussian Radial Basis Function (RBF) because almost every surface can be obtained with it (Christianini and Shawe-Taylor, 2000). One of the main parameters in a SVM is the "cost parameter" (*C*), which is a trade-off between the prediction error and the simplicity of the model. The other hyper-parameter (γ) of SVM enters into the Gaussian function inside the RBF kernel. Performance of SVM is very sensitive to changes in γ parameter. Tested values for *C* were 0.001, 0.1, 1, 5, and 10, and for γ 0.005, 0.05, 0.5, and 5. The "e1071" R package was used for the analyses (Meyer et al., 2019).

Gradient boosting or GB (Mason et al., 1999) is an ensemble method because it uses several fast and easy computation learning algorithms to get a better predictive performance than the one that could be obtained by using the algorithms individually. Predictors are combined sequentially by applying some shrinkage to each. Details on GB can be found in Hastie et al. (2009). For implementing GB, three hyper-parameters were tuned: depth of tree (10, 12, 15, 17), a learning rate that controls the size of the steps in the gradient descent process (0.01 and 0.02), and the number of trees (500, 1,000, and 3,000 trees). The "gbm" R package was used (Greenwell et al., 2019).

A Bayesian GBLUP was used as a reference employing the same outer training and testing datasets partitions as those used with the other learners. In Bayesian GBLUP, the model was $\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \boldsymbol{\varepsilon}$ the vector of genomic breeding values, $u = \{u_i\}$, are assumed to be normally distributed as $p\left(\mathbf{u}|\sigma_u^2\right) \sim N\left(0, \mathbf{G}\sigma_u^2\right)$. The σ_u^2 is the additive genomic variance and the genomic relationship matrix (*G*) was computed from the 46,610 SNPs as $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{E})(\mathbf{M}-\mathbf{E})'}{2\sum_{j=1}^{p} q_j(1-q_j)}$ (VanRaden, 2008). Marker genotypes in M were previously centered by subtracting the average allele frequencies at each locus (i.e., $2q_j$ from each element on column j of $\mathbf{E}$ where q_j is the allelic frequency of the major allele of the *j*th SNP. Since allele frequencies at each locus from the base population were not available, they were computed directly from the available data. The distribution for random residuals was assumed to be $p\left(\boldsymbol{\varepsilon}|\sigma_\varepsilon^2\right) \sim N\left(0, \mathbf{I}\sigma_\varepsilon^2\right)$. Priors assumed for σ_u^2 and σ_ε^2 were scaled inverse $-\chi^2$ distributions with degrees of freedom d_l and scale factor S_l: $p\left(\sigma_l^2|d_l, S_l\right) \sim \chi^{-2}\left(\sigma_l^2|d_l, S_l\right)$ for $l \varepsilon \{u, \varepsilon\}$.

The Gibbs2f90 software (Misztal, 1999) was used to implement this method. Flat priors were assumed for σ_u^2 and the residual variance. Single chains of 250,000 iterations were run by discarding the first 25,000. The number of discarded samples was, in all folds, larger than the required burn-in that was determined by visual inspection of the chains. Samples of parameters of interest were saved every 10 iterations and their posterior means were retained for each training/testing partition of the dataset for later comparison. Effective sample size was larger than 700 for all the parameters of the model.

Quality of Prediction and Stability of Feature Selectors

The objective was to find the best combination of FS method and learner to obtain the smallest and most stable SNP subset that leads to the most accurate prediction.

The quality of trait prediction was evaluated for accuracy, as the median of the Spearman correlation (SC) between observed (i.e., adjusted phenotypes) and predicted trait across the 10 outer testing sets, and for stability/generality of results, as the interquartile range (IQR) of those values.

Stability of FS algorithms measures how variation in the training sample produces a change in the selected feature subset (Kalousis et al., 2007). If FS is performed setting a threshold for a number of the most important features for prediction based on a weight, score, or rank assigned to each feature, preferential stability can be measured as the mean Pearson's correlation (PC) or as the SC between all pairs of

weighting-scoring and ranking values, respectively, obtained in different training sets. When FS is performed by a procedure that does not involve any weight or rank (i.e., using embedded and wrapper methods), preferential stability can be measured as the amount of overlap between two sets of an arbitrary size (Generalized Kalousis, Kalousis et al., 2007). All those measures require the equal size of the feature subsets. To compare the stability of subsets of varying sizes, as obtained with embedded methods, Somol and Novovicova (2010) introduced the relative weighted consistency measure. Relative Weighted Consistency and Generalized Kalousis do not meet the properties (iv) and (v), respectively, for a proper stability estimator established by Nøgueira et al. (2018) who proposed a new stability estimator $\Phi(S)$ (NOG) defined as:

$$\widehat{\Phi}(S) = 1 - \frac{\frac{1}{p}\sum_{i=1}^{p}\sigma^2_{f_i}}{E\left[\frac{1}{p}\sum_{i=1}^{p}\sigma^2_{f_i} \mid H_0\right]} = 1 - \frac{\frac{1}{p}\sum_{i=1}^{p}\sigma^2_{f_i}}{\frac{\bar{d}}{p}\left(1-\frac{\bar{d}}{p}\right)}$$

For $T = \{f_1, f_2, \ldots, f_p\}$ being the whole set of features of size p (i.e., 9,523 in our study) and $S = \{S_1, S_2, \ldots, S_n\}$ being a system of $n > 1$ subsets $S_j = \{f_i | i = 1, \ldots, d_j, f_i \in Y, d_j \in \{1,2, \ldots, p\}\}$ with $j = 1, 2, \ldots, n$ obtained from n runs of the FS algorithm ($n = 10$ in our study), $\sigma^2_{f_i} = \frac{n}{n-1}\hat{p}_{f_i}\left(1-\hat{p}_{f_i}\right)$ is the unbiased sample variance of the ith SNP, F_{f_i} is the frequency of the feature f_i, $\hat{p}_{f_i} = \frac{1}{n}F_{f_i}$ is the mean of the former, and $\bar{d}$ is the average number of features selected over the n feature sets. To correct for similarity between feature sets due to chance, the estimator is rescaled by its expected value under the Null model of random FS (H_0). NOG ranges from 0 to 1. Confidence intervals for the population stability $\Phi(S)$ were approximated with a 0.05 significance level. Refer to Nogueira et al. (2018) for further details. In this study, PC, SC, and NOG were used as stability measurements.

RESULTS

Results (Singleton, 2001) refer to prediction of yet-to-be observed individual RFI based on pigs' genotypes. Prediction performances correspond to the 10 best configurations obtained in the 10-outer cross-validation folds. Notice that FS was done by cross-validation in each of the 10-outer training folds. Therefore, for each FS method and learner, there are 10 subsets of selected features and 10 prediction performances obtained with those subsets, allowing a measurement of the stability of the FS method, as well as a measurement of the dispersion of prediction accuracy.

Prediction Quality: Accuracy and Stability

Figures 2, 3 show boxplots for SC obtained in the 10-outer testing sets with GBLUP, ENET, LR, and RR without pre-selection of SNPs (**Figure 2**), and with SVM, GB, ENET, LR, and RR with pre-selection of SNPs performed using the various filter methods (**Figure 3**). No results were obtained with SVM and GB without pre-selection of SNPs due to numerical problems.

When the prediction was performed with GBLUP the median (IQR) of SC was 0.22 (0.01; **Figure 2**), not significantly different from the obtained with RR (median of SC = 0.22) and ENET (median of SC = 0.19) with 9,523 SNPs. However, it was more stable for GBLUP since IQR was 0.04 and 0.03 for RR and ENET, respectively (**Figure 2**). LASSO regression had the poorest performance in terms of accuracy and stability of results [median of SC = 0.19 (0.03)].

Elastic net, RR, and LR combined with different filter methods had the same pattern of prediction performance (**Figure 3**). For subset sizes smaller than 500, prediction accuracy was smaller than with GBLUP when univ.dtree and random selection of SNPs were used as filter methods. However, the same accuracy as GBLUP was obtained when spearcor and mrmr filters were used, even with only 50 SNPs, while cforest filter led to an intermediate predictive performance. The effect of the filter on SC decreased with an increasing number of SNPs up to 1,500, for which SC was not statistically different across filters. Random selection of SNPs led to very poor performances with subset sizes of 50 and 250 SNPs. When more than 250 SNP were used as predictor variables, with SVM and GB used as learner SC of the random filter was the same as the one attained with other filter methods, whereas it remained lower than the other filters when ENET, RR, and LR were used.

Globally, SC attained with SVM and GB increased as the number of SNPs increased up to 500 SNPs, and then remained at about the same level. The stability of results measured as IQR of SC followed the same pattern. However, when spearcor or mrmr were used to perform pre-selection of SNPs, the SC attained with 50 SNPs was close to that attained with 500 or more SNPs: 0.18 (0.04) and 0.18 (0.06) with spearcor and mrmr combined with SVM, respectively; 0.18 (0.04) and 0.20 (0.07) with spearcor and mrmr combined with GB, respectively. The highest median SC was obtained with SVM [0.28 (0.02)] and GB [0.27 (0.04)] using a subset with the 1,000 best-ranked SNPs according to cforest (**Figure 3**). Performance obtained with SVM and GB with just 750 SNPs was in all cases equal or superior to the median SC obtained using GBLUP (0.22), although results were slightly more stable with GBLUP (IQR: 0.02 for the best model with SVM or GB compared to 0.01 for GBLUP).

Feature Selection Stability

Stability of FS methods for prediction of RFI with ENET and LR implemented with SNP subsets obtained with or without pre-selection with various filter methods are presented in **Tables 1** and **2**, respectively. Boxplots for NOG values from ENET and LR by filter method in the 10 subsets are shown in the left and middle panels of **Figure 4**, whereas boxplots for NOG values from just filter methods are in the right panel of **Figure 3**. According to the scale defined by Nogueira et al. (2018), without pre-selection of SNPs the

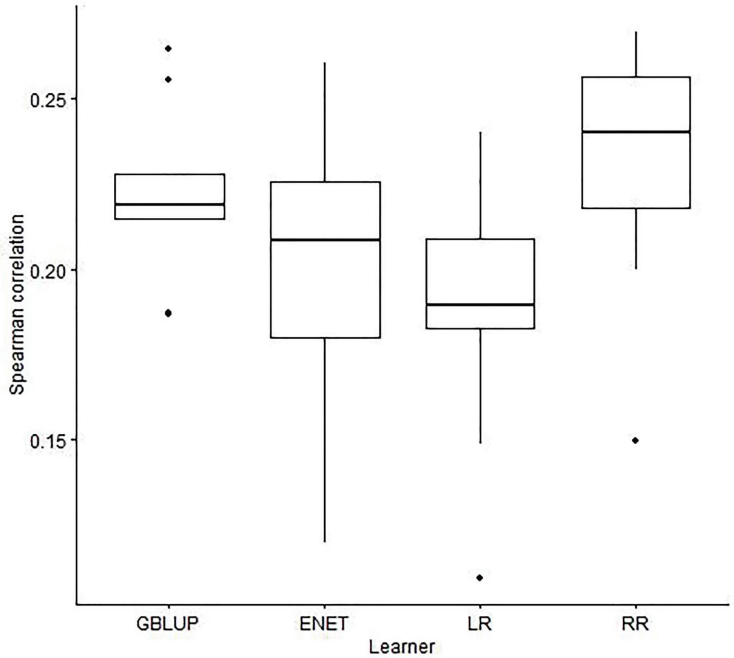

FIGURE 2 | Boxplots for the Spearman correlation obtained in 10-outer testing sets with a genomic best linear unbiased predictor (GBLUP), elastic net (ENET), least absolute shrinkage and selection operator regression (LR), and ridge regression (RR) with no feature selection method used to select SNPs.

stability estimator for ENET and LR was intermediate to good (0.57 and 0.56 for ENET and LR, respectively; **Tables 1** and **2**). In this case, the median (IQR) number of selected SNPs for ENET was 717 (11) out of 9,523 SNPs (**Table 1**), while LR performed a stronger but less stable selection, retaining only 269 (111) SNPs (**Table 2**). With pre-selection of the most important features, ENET and LR removed a considerable part of them and showed differences in FS stability among filter methods and subset sizes (**Tables 1** and **2, Figure 4**). As expected, the number of SNPs selected in regularized regression (ENET and LR) was smaller for random SNP pre-selection than for other filters, ranging from 42% for subset size of 50 to 2.2% for subset sizes of 1,500. However, it ranged from 100% for subset sizes of 50 to 40% for subset sizes of 1,500 SNPs when pre-selection was performed with mrmr. Univariate decision tree performed like random selection whereas in cforest and spearcor the selected percentage was between random and mrmr filters. Elastic net and LR produced similar patterns of FS stability with increasing subset size. The most stable subsets were obtained with spearcor, with excellent values when the subset size of pre-selected SNPs was 50 (0.70 and 0.69 for ENET and LR, respectively, **Figure 4, Tables 1** and **2**). For other subset sizes, no differences in stability were found between spearcor and mrmr, or across subset sizes within filter, with NOG values ranging from 0.52 to 0.55. All other filter methods (i.e., cforest, univ.dtree, or random) gave very poor FS stabilities. Univariate decision tree for SNP pre-selection combined with ENET and LR had null stabilities, with magnitudes that were similar to a random selection. Cforest combined with either ENET or LR slightly improved FS stability up to 0.14 with 1,500 SNPs, which was significantly different from random selection, but very unstable FS and far from the most stable methods.

Stability measurements were slightly smaller when filter methods were combined with embedded methods than when only filter methods were implemented, for all subset sizes (**Tables 1** and **2** vs. **Table 3**). According to the scale defined by Nogueira et al. (2018), when only filter methods were used, the best FS stability was obtained with spearcor for any subset size ranging from 0.73 to 0.69 when 50 and 1,000–1,500 SNPs were pre-selected, respectively (**Figure 4**, right panel). Maximum relevance minimum redundancy also had a good FS stability. Unlike for spearcor, FS stability of mrmr increased with an increasing number of SNPs, from 0.53 to 0.66 with 50 and 1,500 SNPs, respectively. Univariate decision tree and cforest showed null stability, as random selection, with a maximum of 0.04 for cforest with 1,500 SNPs, and marginal improvement with an increase of subset sizes.

Table 4 shows the mean of the PC and SCs over all pairs of feature scores in the 10 outer training sets obtained with the different filter methods for FS. The highest PC and SCs between the scores obtained for each SNP across training sets were obtained with mrmr and spearcor, which indicates the highest stability in the selection of relevant features for prediction across outer training sets. Tree-based methods (univariate or multivariate) exhibited low stability in the selection of predictor variables.

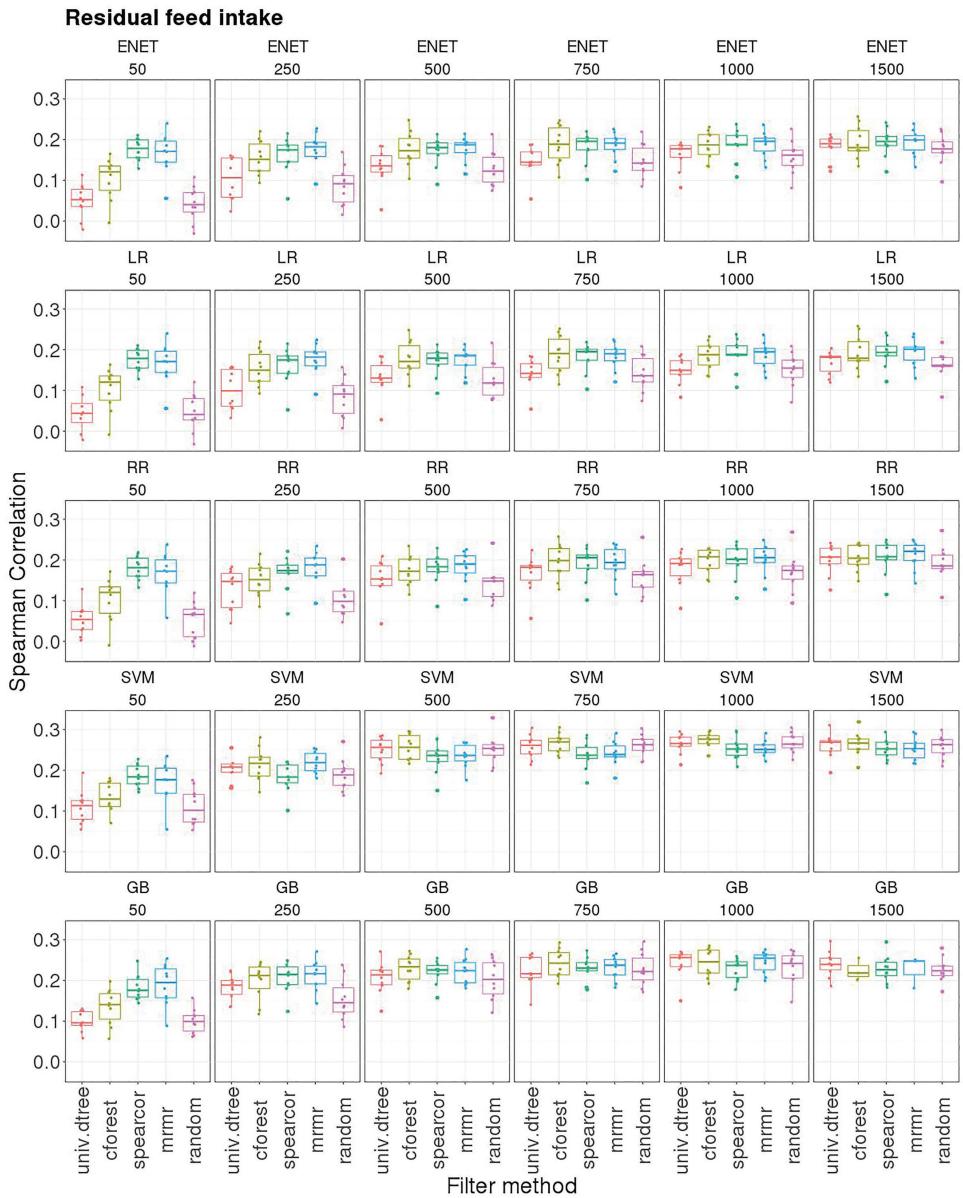

FIGURE 3 | Boxplots for the Spearman correlation obtained in 10-outer testing sets with support vector machine for regression (SVM), gradient boosting (GB), elastic net (ENET), least absolute shrinkage and selection operator regression (LR), and ridge regression (RR) with 50, 250, 500, 750, 1,000, and 1,500 SNP subsets selected with different filter methods. Filter methods: Maximum relevance minimum redundancy (mrmr), random forest (cforest), Spearman's correlation (spearcor), univariate decision tree (univ.dtree), and random selection (random).

The same analyses were performed for ADG with similar results. The corresponding tables and figures showing the results obtained for this trait are provided in **Supplementary Material**.

DISCUSSION

It is well known that classification or prediction with high dimensional data is computationally demanding and may produce overfitted models with poor prediction quality that are difficult to interpret (Huang, 2014). Therefore, the search of effective FS algorithms is important, and it is an active area of research, despite overfitting can be avoided *via* regularization in some models. Such algorithms are required to develop prediction models that are accurate and insensitive to small changes in the training data. Many authors have addressed the question of the sensitivity of FS methods with respect to small changes in the training data in different domains of research. Kalousis et al. (2007) were the first to consider the stability of FS procedures. Their research was

TABLE 1 | Stability of feature selection methods for prediction of residual feed intake with elastic net.

Subset size[1]	Filter[2]	NSNPs[3]	PDF[4]	medianSel[5]	IQRSel[6]	NOG[7]
	mrmr	499	0.30	50	0	0.53
	cforest	458	0.89	47	3	0.03
50	spearcor	445	0.20	45	2	0.70
	univ.dtree	219	0.98	22	6	0.00
	random	209	1.00	21	4	0.00
	mrmr	2,163	0.28	217	8	0.53
	cforest	1,641	0.79	167	12	0.05
250	spearcor	1,541	0.28	155	8	0.55
	univ.dtree	900	0.89	93	8	0.02
	random	854	0.91	87	15	0.01
	mrmr	3,294	0.28	330	20	0.53
	cforest	2,175	0.71	223	35	0.08
500	spearcor	2,597	0.27	263	11	0.53
	univ.dtree	1,501	0.84	154	14	0.02
	random	1,547	0.85	148	19	0.02
	mrmr	4,258	0.27	431	19	0.54
	cforest	2,493	0.65	250	16	0.10
750	spearcor	3,426	0.27	344	21	0.53
	univ.dtree	2,105	0.79	213	13	0.04
	random	2,039	0.80	202	18	0.03
	mrmr	5,059	0.26	511	13	0.55
	cforest	2,879	0.62	292	18	0.11
1,000	spearcor	4,074	0.27	412	20	0.53
	univ.dtree	2,546	0.73	253	11	0.06
	random	2,455	0.75	244	7	0.05
	mrmr	5,929	0.26	593	61	0.55
	cforest	3,557	0.56	355	7	0.14
1,500	spearcor	4,973	0.27	497	23	0.55
	univ.dtree	3,241	0.64	326	16	0.09
	random	3,217	0.66	324	18	0.07
9,523	none	7,193	0.25	717	11	0.57

[1]*Subset size = number of selected features.*

[2]*Filter method = Maximum relevance minimum redundancy (mrmr); Random forest (cforest); Spearman's correlation (spearcor); Univariate decision tree (univ.dtree); Random selection (random).*

[3]*NSNPs = Total number of SNPs pre-selected in the 10 subsets.*

[4]*PDF = Proportion of distinct features in the 10 subsets.*

[5]*medianSel =Mean number of selected SNPS in the 10 subsets.*

[6]*IQRSel = Interquartile range of the number of selected SNPS in the 10 subsets.*

[7]*NOG = Nogueira et al. (2018) stability estimator.*

followed by several publications in application areas where stability is critical, such as microarray classification or molecular profiling and by studies addressing how to quantify stability (Davis et al., 2006; Kuncheva, 2007; Jurman et al., 2008; Zucknick et al., 2008; Zhang et al., 2009; Somol and Novovicova, 2010; Goh and Wong, 2016). The measurement of stability is important because it indicates how much the output of an algorithm can be trusted by capturing the underlying mechanism. This is very important in many biological and biomedical domains, and genetics applied to livestock production is not an exception. Here, an objective could be, for example, to design low-density SNP chips for GS, or to assign further resources to the search of genes with a major effect on important production traits.

In this research, several univariate and multivariate algorithms combined with parametric and non-parametric learners were applied to the prediction of RFI of growing pigs from high-dimensional genomic data (60 K SNP chip). The objective was to find the best combination of feature selector, subset size, and learner leading to as high as possible accurate and stable predictions. GBLUP with no SNP selection beyond the standard quality control was the benchmark. Three types of FS methods were implemented: (i) filter methods: univariate (univ.dtree, spearcor) or multivariate (cforest, mrmr) and a random selection filter as benchmark; (ii) embedded methods: ENET and LR; (iii) the combination of filter and embedded methods. Regularized regression using RR, which does not perform FS, was also used as an intermediate option between no and strong FS. In addition, SVM and GB, considered to be among the most efficient ML methods, were implemented, but only with the SNP pre-selection performed with filter methods. These two methods have been successfully used in various fields (James et al., 2013; Attewell et al., 2015) including livestock and plant breeding (Moser et al., 2009; Long et al., 2011; González-Recio et al., 2014; Montesinos-Lopez et al., 2019).

The best prediction quality in terms of accuracy and stability of results was obtained with SVM and GB for subset sizes equal or larger than 500 SNPs. These two non-parametric methods outperformed GBLUP and regularized methods (RR,

TABLE 2 | Stability of feature selection methods for prediction of residual feed intake with LASSO regression.

Subset size[1]	Filter[2]	NSNPs[3]	PDF[4]	medianSel[5]	IQRSel[6]	NOG[7]
	mrmr	499	0.30	50	0	0.53
	cforest	454	0.89	46	5	0.03
50	spearcor	441	0.20	44	3	0.69
	univ.dtree	151	0.99	24	10	0.00
	random	189	0.99	23	12	0.00
	mrmr	2,096	0.28	209	8	0.53
	cforest	1,693	0.79	172	13	0.05
250	spearcor	1,575	0.28	158	12	0.54
	univ.dtree	804	0.90	83	18	0.02
	random	740	0.92	80	48	0.01
	mrmr	3,420	0.29	347	23	0.52
	cforest	2,287	0.72	239	39	0.07
500	spearcor	2,560	0.28	260	9	0.52
	univ.dtree	1,292	0.86	141	40	0.02
	random	1,332	0.86	151	63	0.02
	mrmr	4,333	0.28	437	24	0.53
	cforest	2,646	0.67	275	35	0.09
750	spearcor	3,339	0.27	338	21	0.53
	univ.dtree	1,667	0.79	165	40	0.04
	random	1,745	0.80	160	57	0.03
	mrmr	5,125	0.27	515	23	0.54
	cforest	3,008	0.63	321	53	0.10
1,000	spearcor	3,978	0.27	394	15	0.53
	univ.dtree	1,750	0.74	193	21	0.06
	random	1,879	0.76	192	62	0.05
	mrmr	5,978	0.26	595	21	0.55
	cforest	3,478	0.56	341	8	0.14
1,500	spearcor	5,096	0.27	506	12	0.54
	univ.dtree	2,270	0.66	235	19	0.09
	random	2,200	0.68	235	80	0.08
9,523	none	2,963	0.27	269	111	0.56

[1]*Subset size = number of selected features.*
[2]*Filter method = Maximum relevance minimum redundancy (mrmr); Random forest (cforest); Spearman's correlation (spearcor); Univariate decision tree (univ.dtree); Random selection (random).*
[3]*NSNPs = Total number of SNPs pre-selected in the 10 subsets.*
[4]*PDF = Proportion of distinct features in the 10 subsets.*
[5]*medianSel = Mean number of selected SNPS in the 10 subsets.*
[6]*IQRSel = Interquartile range of the number of selected SNPS in the 10 subsets.*
[7]*NOG = Nogueira et al. (2018) stability estimator.*

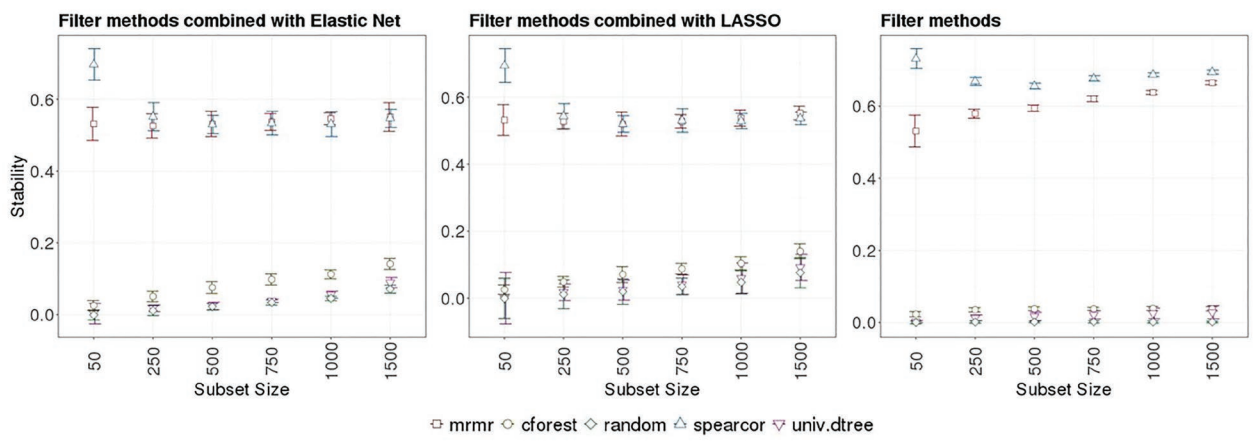

FIGURE 4 | Left and middle panels: Boxplots for Nogueira et al. (2018) stability estimator obtained from embedded methods implemented with different SNP subsets sizes and several filter methods in 10 outer-training folds. Right panel: Boxplots for Nogueira et al. (2018) stability estimator of different SNP subset sizes obtained with different filter methods in the 10-outer training folds. Subsets sizes: 50, 250, 500, 750, 1,000, and 1,500 SNPs. Filter methods: Maximum relevance minimum redundancy (mrmr), random forest (cforest), Spearman's correlation (spearcor), univariate decision tree (univ.dtree), and random selection (random).

TABLE 3 | Stability of filter methods for prediction of residual feed intake.

Subset size[1]	Filter method[2]	NSNPs[3]	PDF[4]	NOG[5]
	mrmr	500	0.30	0.53
	cforest	500	0.90	0.02
50	spearcor	500	0.19	0.73
	univ.dtree	500	0.96	0.00
	random	500	0.98	0.00
	mrmr	2,500	0.26	0.58
	cforest	2,500	0.80	0.04
250	spearcor	2,500	0.23	0.67
	univ.dtree	2,500	0.85	0.01
	random	2,500	0.88	0.00
	mrmr	5,000	0.25	0.59
	cforest	5,000	0.72	0.04
500	spearcor	5,000	0.21	0.66
	univ.dtree	5,000	0.75	0.02
	random	5,000	0.79	0.00
	mrmr	7,500	0.22	0.62
	cforest	7,500	0.65	0.04
750	spearcor	7,500	0.20	0.68
	univ.dtree	7,500	0.67	0.02
	random	7,500	0.71	0.00
	mrmr	10,000	0.21	0.64
	cforest	10,000	0.59	0.04
1,000	spearcor	10,000	0.19	0.69
	univ.dtree	10,000	0.60	0.03
	random	10,000	0.64	0.00
	mrmr	15,000	0.19	0.66
	cforest	15,000	0.49	0.04
1,500	spearcor	15,000	0.18	0.69
	univ.dtree	15,000	0.50	0.03
	random	15,000	0.52	0.00

[1]*Subset size = number of selected features.*
[2]*Filter method = Maximum relevance minimum redundancy (mrmr); Random forest (cforest); Spearman's correlation (spearcor); Univariate decision tree (univ.dtree); Random selection (random).*
[3]*NSNPs = Total number of SNPs selected in the 10 subsets.*
[4]*PDF = Proportion of distinct features in the 10 subsets.*
[5]*NOG = Nogueira et al. (2018) stability estimator.*

TABLE 4 | Mean of the Pearson and Spearman correlation over all pairs of feature scores in the 10-outer training sets obtained with different filter methods for feature selection for prediction of residual feed intake.

Filter method[1]	Pearson	Spearman
mrmr	0.875	0.893
cforest	0.025	0.015
spearcor	0.863	0.802
univ.dtree	0.062	0.059
random	0.001	0.001

[1]*Filter method = Maximum relevance minimum redundancy (mrmr); Random forest (cforest); Spearman's correlation (spearcor); Univariate decision tree (univ.dtree); Random selection (random).*

LR, and ENET). This could be due to the ability of non-parametric models to capture interactions among predictor variables and non-linear relationships with the target variable without explicitly modeling these interactions or functional forms (Gianola et al., 2006; Gianola and van Kaam, 2008). They are potentially able to capture complex signals from the data and deliver a better predictive accuracy, even if the trait is under additive gene action (Perez-Rodriguez et al., 2012, 2013). With large subset sizes (i.e., 1,000–1,500 SNPs), the filter method had no appreciable influence on prediction quality, which was almost the same as with random SNP selection. However, FS was essential for the implementation of both methods, which had numerical problems when all 9,523 SNPs were included in the analysis. With small subset sizes (i.e., equal or smaller than 250 SNPs), the FS algorithm used had a huge impact. In fact, prediction quality was poor when tree-based methods or random selection were used for FS with any learner, but it was comparable to the one attained with larger SNP subsets when spearcor or mrmr were implemented for SNP pre-selection combined or not with embedded methods. Those filter methods also led to more stable results (i.e., smaller IQR of SC).

Regularized methods (ENET, LR, and RR) with or without embedded FS followed the same pattern with respect to subset size (**Figure 3**). When no SNP pre-selection was performed or subset size was equal or larger than 750 SNPs, prediction accuracy was not different from the one attained with GBLUP, although results were more stable with the latter (**Figure 2**). In this situation (i.e., with medium-large subset sizes), FS would reduce computation time and resources, but would not improve over the predictions from regularized regression. Like for SVM and GB, when subset size was smaller than 500 SNPs, the filter method had a marked influence on prediction quality, with spearcor and mrmr being the only methods that produced a prediction quality comparable to the one obtained when all 9,523 SNPs were used.

Results regarding the stability of FS (**Tables 1**–**4**) were quite clear. Stability generally increased with the subset size. Stability of tree-based methods for FS (univ.dtree and cforest) was considered null. The FS methods that produced the best prediction quality (spearcor and mrmr) even with small subset sizes (i.e., 50 or 250 SNPs) were also the ones that showed stable compositions of SNP subsets, insensitive to changes in datasets. Univariate FS was computationally fast, but it does not account for the potential correlation between features. This could lead to similar ranking scores of correlated features that would potentially be selected together leading to increased redundancy in the feature subset retained. The similar performance of spearcor and mrmr methods, one univariate and the other multivariate, could be due to the fact that most correlated and less variable SNPs were excluded in the previous step of SNP quality control, before the FS process. This could have masked the potential differences in stability resulting from accounting for correlations between features or not. However, when combined with regularized methods (ENET and LR), the SNP subsets obtained with spearcor were more trimmed by the regularization than those obtained with mrmr, suggesting that some redundancies remained. With this two-step approach, spearcor seems preferable over mrmr, because of its much lower requirements in computation time and resources.

Elastic net and LR had the same stability values within subset size, despite different underlying SNP selection strategies. According to Nogueira et al. (2018), when features are correlated, LR would tend to select different features in different subsets.

Elastic net might increase stability. However, as stated above, we used 9,523 out of 45,610 SNPs, with a PC smaller than 0.8 in order to reduce computation requirements of the ML algorithms. This pre-selection may have reduced redundancy, which produces instability (Gulgezen et al., 2009), reducing the differences between both regressors.

In conclusion, both accuracy and stability of results and of feature selector should be accounted for when constructing a model for prediction. In the case of prediction of RFI in growing pigs, different FS algorithms performing similarly well for prediction had wide differences in terms of stability. This issue can have important consequences on the interpretability and reproducibility of results and should be considered as an additional criterion to consider when evaluating FS methods. Elastic net, LR, and RR did not outperform GBLUP when the 9,523 SNPs were used for prediction or when they were pre-selected according to some criteria. With these learners, when the subset size was small (50–250 SNPs), only SNPs pre-selected with mrmr or spearcor produced prediction accuracies comparable to that of GBLUP with all the SNPs, and good FS stabilities. Thus, a strong SNP pre-selection could be performed to reduce computation requirements for regularized regression. Nevertheless, the best prediction quality in terms of accuracy and stability was obtained with the ML approaches (SVM and GB) using 500 or more SNPs pre-selected with mrmr or spearcor as predictor variables. Thus, the use of low-density SNP chips for GS seems feasible. Finally, when SNP quality control includes removing highly correlated SNPs, SC is recommended for FS over mrmr because of its simplicity and small requirements in computation time and resources.

Results and conclusions for ADG were consistent with the ones obtained for RFI. They are provided in **Supplementary Material**.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The datasets presented in this article are not readily available because of containing information that could be used by commercial competitors. Topigs Norsvin (Beuningen, Netherlands) fully agrees with transparency in science and welcomes alternative analyses of the data with a gatekeeper, a trusted person, who considers the relevance and the motives of the people interested in the data. Requests to access these datasets should be directed to Dr. Rob Bergsma (Rob.Bergsma@topigsnorsvin.com).

AUTHOR CONTRIBUTIONS

MP and LT conceived and designed the study, carried out the analyses, and wrote the original draft. RB prepared and provided the raw data. DG and HG provided critical insights and gave methodological suggestions. All authors discussed the results, reviewed, and approved the final manuscript.

FUNDING

This study was supported by the European Unions' Horizon 2020 Research & Innovation programme under grant agreement N° 633531 – FEED-A-GENE, the INRAE SelGen Metaprogram project (OptiMAGicS) and by Spanish Ministry of Economy, Industry and Competitiveness (MINECO; RTI2018-097610-R-I00).

ACKNOWLEDGMENTS

We are grateful to Topigs Norsvin and their staff for collecting and providing the data and to Noemí Piles for **Figure 1** layout. We are also grateful to the GenoToul bioinformatics platform Toulouse Midi-Pyrenees for providing computing and storage resources, to R GNU project contributors for making the R environment freely available, to the developers of the R packages used in this research and to Ignacy Misztal and coworkers for the blupf90 suite of programs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.611506/full#supplementary-material

Supplementary Table 1 | Stability of feature selection methods for prediction of average daily gain with Elastic Net.

Supplementary Table 2 | Stability of feature selection methods for prediction of average daily gain with LASSO regression.

Supplementary Table 3 | Stability of filter methods for prediction of average daily gain.

Supplementary Table 4 | Mean of the Pearson and Spearman correlation over all pairs of feature scores in the 10-outer training sets obtained with different filter methods for feature selection for prediction of average daily gain.

Supplementary Figure 1 | Boxplots for the Spearman correlation for average daily gain obtained in 10-outer testing sets with a Genomic Best linear Unbiased Predictor (GBLUP), Elastic Net (ENET), Least Absolute Shrinkage and Selection Operator Regression (LR), and Ridge Regression (RR) with no feature selection method used to select the SNPs.

Supplementary Figure 2 | Boxplots for the Spearman correlation for average daily gain obtained in 10-outer testing sets with Support Vector Machine for regression (SVM), Gradient Boosting (GB), Elastic Net (ENET), Least Absolute Shrinkage and Selection Operator Regression (LR), and Ridge Regression (RR) with 50, 250, 500, 750, 1,000 and 1,500 SNP subsets selected with different filter methods. Filter methods: Maximum relevance minimum redundancy (mrmr), Random forest (cforest), Spearman's correlation (spearcor), Univariate decision tree (univ.dtree) and random selection (random)

Supplementary Figure 3 | Left and middle panels: Boxplots for Nogueira et al. (2018) stability estimator obtained for average daily gain from embedded methods implemented with different SNP subsets sizes and several filter methods in 10 outer-training folds. Right panel: Boxplots for Nogueira et al. (2018) stability estimator of different SNP subset sizes obtained with different filter methods in the 10-outer training folds. Subsets sizes: 50, 250, 500, 750, 1,000 and 1,500 SNPs. Filter methods: Maximum relevance minimum redundancy (mrmr), Random forest (cforest), Spearman's correlation (spearcor), Univariate decision tree (univ.dtree) and random selection (random).

REFERENCES

Alzubi, R., Ramzan, N., Alzoubi, H., and Amira, A. (2018). A hybrid feature selection method for complex diseases SNPs. *IEEE Access* 6, 1292–1301. doi: 10.1109/ACCESS.2017.2778268

Attewell, P., Monaghan, D. B., and Kwong, D. (2015). *Data mining for the social sciences: An introduction*. Oakland, CA: University of California Press.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci. Rep.* 5:10312. doi: 10.1038/srep10312

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J. S. E., Casalicchio, G., et al. (2016). mlr: machine learning in R. *J. Mach. Learn. Res.* 17, 1–5.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Inf. Sci.* 282, 111–135. doi: 10.1016/j.ins.2014.05.042

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., and Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* 143:106839. doi: 10.1016/j.csda.2019.106839

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. Electr. Eng.* 40, 16–28. doi: 10.1016/j.compeleceng.2013.11.024

Cristianini, N., and Shawe-Taylor, L. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

Davis, C. A., Gerick, F., Hintermair, V., Friedel, C. C., Fundel, K., Küffner, R., et al. (2006). Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics* 22, 2356–2363. doi: 10.1093/bioinformatics/btl400

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinforma. Comput. Biol.* 3, 185–205. doi: 10.1142/s0219720005001004

Drumond, D. A., Rolo, R. M., and Costa, J. F. C. L. (2019). Using Mahalanobis distance to detect and remove outliers in experimental covariograms. *Nat. Resour. Res.* 28, 145–152. doi: 10.1007/s11053-018-9399-y

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gianola, D., Hayrettin, O., Weigel, K. A., and Rosa, G. J. M. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12:87. doi: 10.1186/1471-2156-12-87

Gianola, D., and van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Goh, W. W., and Wong, L. (2016). Evaluating feature-selection stability in next-generation proteomics. *J. Bioinforma. Comput. Biol.* 14:1650029. doi: 10.1142/s0219720016500293

González-Recio, O., Rosa, G. J. M., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166, 217–231. doi: 10.1016/j.livsci.2014.05.036

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G.. 2019. gbm: Generalized Boosted Regression Models. R package version. 2.1.5. Available at: https://CRAN.R-project.org/package=gbm (Accessed June 10, 2020).

Gulgezen, G., Cataltepe, Z., and Yu, L. (2009). *Stable and accurate feature selection*. Heidelberg: Berlin, 455–468.

Gunavathi, C., Premalatha, K., and Sivasubramanian, K. (2017). A survey on feature selection methods in microarray gene expression data for cancer classification. *Res. J. Pharm. Technol.* 10, 1395–1401. doi: 10.5958/0974-360X.2017.00249.9

Guyon, I., and Elissee, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Huang, J. Z. (2014). An introduction to statistical learning: with applications in R by Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten. *J. Agric. Biol. Environ. Stat.* 19, 556–557. doi: 10.1007/s13253-014-0179-9

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). "Linear model selection and regularization" in *An introduction to statistical learning. Springer texts in statistics. Vol. 103*. eds. G. Casella, S. Fienberg and I. Olkin (New York, NY: Springer).

Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., and Furlanello, C. (2008). Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* 24, 258–264. doi: 10.1093/bioinformatics/btm550

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.* 12, 95–116. doi: 10.1007/s10115-006-0040-8

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28:26. doi: 10.18637/jss.v028.i05.

Kuncheva, L. I. (2007). "A stability index for feature selection." in *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*. ACTA Press, Innsbruck, Austria, 390–395.

Long, N., Gianola, D., Rosa, G. J., and Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123, 1065–1074. doi: 10.1007/s00122-011-1648-y

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. (1999). "Boosting algorithms as gradient descent" in *Advances in neural information processing systems 12*. eds. S. A. Solla, T. K. Leen and K. Müller (MIT Press), 512–518.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R. package version 1.7-2. Available at: https://CRAN.R-project.org/package=e1071 (Accessed June 10, 2020).

Misztal, I. (1999). Complex models, more data: simpler programming. *Interbull Bull. Proc. Inter. Workshop Comput. Cattle Breed. Tuusala, Finland* 20, 33–42.

Montesinos-Lopez, O. A., Martin-Vallejo, J., Crossa, J., Gianola, D., Hernandez-Suarez, C. M., Montesinos-Lopez, 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* 9, 601–618. doi: 10.1534/g3.118.200998

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., and Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41:56. doi: 10.1186/1297-9686-41-56

Nogueira, S., Sechidis, K., and Brown, G. (2018). On the stability of feature selection algorithms. *J. Mach. Learn. Res.* 18, 6345–6398.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337

Perez-Rodriguez, P., Gianola, D., Gonzalez-Camacho, J. M., Crossa, J., Manes, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3* 2, 1595–1605. doi: 10.1534/g3.112.003665

Perez-Rodriguez, P., Gianola, D., Weigel, K. A., Rosa, G. J., and Crossa, J. (2013). Technical note: an R package for fitting Bayesian regularized neural networks with applications in animal breeding. *J. Anim. Sci.* 91, 3522–3531. doi: 10.2527/jas.2012-6162

Phuong, T. M., Lin, Z., and Altman, R. B. (2005). "Choosing SNPs using feature selection." *2005 IEEE Computational Systems Bioinformatics Conference (CSB'05), 2005*; Stanford, CA, USA, 301–309.

R Development Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Samb, M. L., Camara, F., Ndiaye, S., Slimani, Y., and Esseghir, M. A. (2012). A novel RFE-SVM-based feature selection approach for classification. *Int. J. Adv. Sci. Technol.* 43, 27–36.

Singleton, W. L. (2001). State of the art in artificial insemination of pigs in the United States. *Theriogenology* 56, 1305–1310. doi: 10.1016/s0093-691x(01)00631-8

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Somol, P., and Novovicova, J. (2010). Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1921–1939. doi: 10.1109/tpami.2010.34

Strobl, C., Boulesteix, A. -L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics* 9:307. doi: 10.1186/1471-2105-9-307

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

Venkatesh, B., and Anuradha, J. (2019). A review of feature selection and its methods. *Cybern. Inf. Technol.* 19, 3–26. doi: 10.2478/cait-2019-0001

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

Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., et al. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics* 25, 1662–1668. doi: 10.1093/bioinformatics/btp295

Zou, H., and Hastie, T. (2005). Regularization and variable selection *via* the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 301–320.

Zucknick, M., Richardson, S., and Stronach, E. A. (2008). Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Stat. Appl. Genet. Mol. Biol.* 7. doi: 10.2202/1544-6115.1307

Conflict of Interest: RB is a member of Topigs Norsvin's (Beuningen, Netherlands) staff.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.