Prediction of slaughter age in pigs and assessment of the predictive value of phenotypic and genetic information using random forest

Ahmad Alsahaf,^{*,1} George Azzopardi,^{*} Bart Ducro,[†] Egiel Hanenberg,[‡] Roel F. Veerkamp,[†] and Nicolai Petkov^{*}

^{*}University of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, P.O. Box 407, 9700 AK Groningen, The Netherlands; [†]Wageningen University & Research, P.O. Box 338, 6708 PB Wageningen, The Netherlands; and [‡]Topigs Norsvin Research Center, Beuningen, The Netherlands

ABSTRACT: The weight of a pig and the rate of its growth are key elements in pig production. In particular, predicting future growth is extremely useful, since it can help in determining feed costs, pen space requirements, and the age at which a pig reaches a desired slaughter weight. However, making these predictions is challenging, due to the natural variation in how individual pigs grow, and the different causes of this variation. In this paper, we used machine learning, namely random forest (RF) regression, for predicting the age at which the slaughter weight of 120 kg is reached. Additionally, we used the variable importance score from RF to quantify the importance of different types of input data for that prediction. Data of 32,979 purebred Large White pigs were provided by Topigs Norsvin, consisting

of phenotypic data, estimated breeding values (EBVs), along with pedigree and pedigree-genetic relationships. Moreover, we presented a 2-step data reduction procedure, based on random projections (RPs) and principal component analysis (PCA), to extract features from the pedigree and genetic similarity matrices for use as inputs in the prediction models. Our results showed that relevant phenotypic features were the most effective in predicting the output (age at 120 kg), explaining approximately 62% of its variance (i.e., $R^2 = 0.62$). Estimated breeding value, pedigree, or pedigree-genetic features interchangeably explain 2% of additional variance when added to the phenotypic features, while explaining, respectively, 38%, 39%, and 34% of the variance when used separately.

Key words: breeding, grouping strategies, machine learning, pigs, random forest, regression

© The Author(s) 2018. Published by Oxford University Press on behalf of the American Society of Animal Science. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

J. Anim. Sci. 2018.96:4935–4943

doi: 10.1093/jas/sky359

INTRODUCTION

Variation in body growth speed has a big impact on pig farming, since it directly affects key elements of production costs like feed, logistics, and veterinary medical care (Patience et al., 2004). For instance, if a group of pigs in a finishing pen contains slow growers, then those pigs must be retained in the pen until they reach market weight before the pen can be cleared to receive a new

group. This would incur additional feed cost and labor hours, especially if the farm implements an all-in/all-out management system. (Patience et al., 2004). Therefore, a good estimate of each pig's future growth performance can improve the efficiency at pig farms and breeding facilities, for example, by using those estimates to assign pigs to groups that will be nearly uniform in weight at a target age, or groups that will reach a target weight at a nearly uniform age.

As with other farm animals, pig growth is a complex phenomenon that is influenced by many factors, including sex, age, weight history, feed intake, genetics, health, sow and litter characteristics,

¹Corresponding author: a.m.j.a.alsahaf@rug.nl

Received March 22, 2018.

Accepted September 5, 2018.

and farm conditions (Apichottanakul et al., 2012). Therefore, it is not effective to isolate one, or few of these factors, as predictors of future weight or growth (Gonyou, 1998 and references therein).

But with the rise of modern performance recording systems in pig production, which record large volumes of phenotypic, genetic, and environmental data (Ma et al., 2012; Kim et al., 2014), and the development of computational methods that can utilize these data, more accurate growth predictions can be attained.

In this paper, we used a machine learning approach, namely the random forest (RF) algorithm (Breiman, 2001a), to combine different types of predictors, phenotype, estimated breeding value (EBV), along with pedigree and pedigree-genetic relationship data. Unlike traditional statistical analysis, machine learning emphasizes prediction accuracy of the models rather than the fit of the data to predetermined statistical models or structures (Breiman, 2001b). It allows the inclusion of heterogeneous data types without hypotheses on which underlying distributions generate them.

We aim to demonstrate that a model-free, machine learning approach can be used for the prediction of slaughter age in pigs, and by extension, other related phenotypes. Additionally, we aim to rank different groups of features based on their effectiveness in predicting slaughter age in pigs.

MATERIALS AND METHODS

The data we used in this study were provided by the company Topigs Norsvin. It consisted of features of 32,979 Large White finisher pigs—24,978 females and 8,001 males—whose ages range from 39 to 168 d at the start of the finishing stage. Since the data were acquired from the databases of Topigs Norsvin, Animal Care and Use Committee approval was not necessary.

Data

The available data were split in 3 groups: 1) *phenotypic data*: this group consists of phenotypic records that domain experts believe to be relevant to growth, including sex, recorded weights (at birth and at the start of the finishing phase), age at 30 kg, birth farm, litter/sow information (e.g., gestation length, parity number, number of born piglets), and performance metrics of similar animals like the average age at 120 kg of farm-mates of the same sex. This group of features forms an input feature matrix of 20 features that we denote by X^{ph} ,

which has a size of $n \times 20$; 2) *estimated breeding values*: this group includes EBV of 7 traits, namely, sow longevity, piglet vitality, back fat thickness, loin depth thickness, total number of born piglets, mothering ability, and daily gain. We also included the inbreeding coefficient as an additional genetic metric. These features form an input feature matrix that we denote by X^{EBV} , which has a size of $n \times 9$; 3) *pedigree and pedigree-genetic pairwise similarities*: the final group of data consists of $2n \times n$ matrices that include the pairwise pedigree and pedigree-genetic similarities between all studied animals. We denote those matrices by $A_{n \times n}$ and $H_{n \times n}$, respectively. Matrix $H_{n \times n}$, the pedigree-genomic similarity matrix, was derived from matrix $A_{n \times n}$, and the genetic similarity matrix G , according to the formula in Eq. 1 (Legarra et al., 2009). The number of genotyped animals was 11,699. The input feature matrices X^{ph} and X^{EBV} are suited for usage in RF without modification, except for the categorical inputs, sex, fostering, and farm of birth in X^{ph} , which need to be encoded as dummy variables. The complete list of features is included in [Supplementary Appendix 1](#). Finally, we denote by Y the predicted output, which is the age at 120 kg, forming an output vector of size $n \times 1$

$$H_{n \times n} = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix} \quad (1)$$

where G is the genomic similarity matrix of the subset of genotyped animals, A_{11} is the pedigree similarity matrix between ungenotyped animals, A_{22} is the pedigree similarity matrix between genotyped animals, A_{21} is the pedigree similarity matrix between the genotyped and ungenotyped animals, and A_{12} is the transpose of A_{21} .

Regression with RF

The RF algorithm is a decision tree-based ensemble learning method. Ensemble learning methods are a subcategory of machine learning algorithms that combine a given number of predictive models (for classification or regression) to obtain a model that predicts better than any of its constituents. Random forest uses sample bagging (Breiman, 1996), and random sampling from the feature space at the nodes of the trees of the ensemble to create a “forest” of diverse decision tree predictors, which leads to a reduction of variance compared to an individual tree, without increasing the bias. A RF model is a nonlinear predictor; therefore, it is able to model nonlinear relations between

regressors and the output, unlike, for instance, multiple linear regression (MLR), which we use in this paper as a performance benchmark.

Random forest is applicable for classification problems, when the output variable is categorical, or for regression problems, when the output variable is continuous. In the regression case, the algorithm works as follows: 1) drawing M bootstrapped subsets (i.e., random subsets with replacement) from the training set to grow M regression trees; 2) sampling p variables from the input matrix at each splitting node in each decision tree, and selecting the best split in each node until each tree is fully grown or a stopping criterion—e.g., the maximum number of levels in a tree—is met; 3) after the RF is fit to the training data, the output prediction for a new unseen sample is given by the average of M predictions; one from each tree. The prediction of each tree is computed by applying the splitting rules learned from the training procedure to the new sample until it reaches a leaf node, and taking the average output of the samples in that leaf node. In this paper, we used the following training parameters: $M = 500$, $p = \sqrt{m}$, where m is the number of independent variables, and a stopping criterion $n_{\min} = 5$, which is the minimum number of samples in a node before splitting is stopped. Random forest models were implemented using the Python package Scikit-learn (Pedregosa et al., 2011).

Multiple Linear Regression

For comparison, we used MLR models in their standard formulation:

$$Y = \beta_0 + X_1\beta_1 + \dots + X_N\beta_N + E \quad (2)$$

where Y is the dependent variables, $X_1, \dots, X_N$ are the independent variables, $\beta_0, \dots, \beta_N$ are the regression coefficients, and E is the error term.

Before fitting the linear models, we standardized the input variables by subtracting the mean from each column and dividing by the standard deviation. For consistency, we also did this before fitting the RF models, although RF is not sensitive to variable scaling. We used MATLAB to implement the MLR models, using least squares as the fitting method.

We applied RF and MLR in 10-fold cross-validation, on all input matrices (X^{ph} , X^{EBV} , X^P , and X^G). We then did the same for all possible combination of the input matrices by concatenation. For instance an input matrix denoted by $[X^{ph} X^{EBV}]$ is one that includes all phenotypic features and estimated breeding values. The input matrix containing

all features $[X^{ph} X^{EBV} X^P X^G]$ is denoted by X for brevity.

Feature Importance

Random forest provides an internal score of feature importance, which can be utilized to interpret the resulting models, namely, to understand which features are most relevant to the output. This feature importance score is a natural result of fitting a RF on training data.

In a decision tree, data in each node are split based on a condition on a single feature. A good split is one that decreases the impurity of a subset of objects after splitting. In the case of regression trees, this impurity is based on variance, and thus, the splitting score is called variance reduction (VR), and is defined in Eq. 3 (Geurts et al., 2006). In RF regression, the variable importance score of a variable is the total VR caused by that variable in all regression trees in the ensemble.

$$VR = \frac{\text{var}(y|s) - \frac{s_a}{s} \text{var}(y|s_a) - \frac{s_b}{s} \text{var}(y|s_b)}{\text{var}(y|s)} \quad (3)$$

where s is the given set of objects before splitting; and s_a and s_b are the resulting subsets from applying the split; $\text{var}(y|s_*)$ is the variance of y in the set s_* .

Pedigree and Pedigree-Genetic Similarities

The matrices $A_{n \times n}$ and $H_{n \times n}$, which contain the pairwise pedigree and pedigree-genetic similarities between the animals, can also be used as inputs to an RF model, or other supervised learning models. This can be done by defining the feature vector for each sample—in this case a pig—as its pedigree or pedigree-genetic similarity to all other pigs, i.e., $A_{n \times n}$ and $H_{n \times n}$ are treated as matrices of n dimensional feature vectors for n samples. A feature space of this dimension can, however, create several problems, concerning the interpretation of feature importance scores, as well as increasing the computational requirements of training and testing the RF models. For those reasons, and considering that pedigree and pedigree-genetic similarities among the concerned pigs are highly correlated, we applied a 2-step dimensionality reduction procedure, based on random projection (RP) (Bingham and Mannila, 2001) and principal component analysis (PCA), to both matrices before using them in RF.

Random projections are based on the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss,

1984), which states that any n points in a high m -dimensional Euclidean space can be mapped onto a lower k -dimensional space where $k = \mathcal{O}(\log n / \epsilon^2)$, without distorting the distance between any 2 points by more than a factor of $(1 \pm \epsilon)$. Thus, the lower dimension k depends only on the number of points n and the desired reduction fidelity, but not on the dimension of the original space, m .

For the reduction of matrices $A_{n \times n}$ and $H_{n \times n}$, we used a variant of RPs called very sparse RPs (Li et al., 2006). A data matrix $F \in \mathbb{R}^{n \times m}$ is mapped onto a lower dimensional space $\mathbb{R}^{n \times k}$ by multiplying F with a matrix $R^{n \times k}$, which has entries in

$\{-1, 0, 1\}$ with probabilities $\left\{ \frac{1}{2\sqrt{m}}, 1 - \frac{1}{\sqrt{m}}, \frac{1}{2\sqrt{m}} \right\}$,

resulting in a much smaller matrix $J \in \mathbb{R}^{n \times k}$. This method is more efficient than conventional RP due to the large number of zero entries in the projection matrix compared to a Gaussian projection matrix, with only a small loss of accuracy (Li et al., 2006).

$$J = \frac{1}{\sqrt{k}} FR \in \mathbb{R}^{n \times k}, \quad k \ll \min(n, m) \quad (4)$$

The lower dimension k can be chosen according to the degree of reduction required, while ensuring that the distances between pairs in the lower dimension are not distorted. We chose $k = 500$, which for $n = 32,797$, corresponds approximately to $\epsilon = 0.5$, using the bound given in Eq. 5 (Johnson and Lindenstrauss, 1984). The actual error between the distance after the reduction for different values of k are given in Supplementary Appendix 2.

$$k \geq 4 \frac{1}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \log n \quad (5)$$

We applied very sparse RP to both $A_{n \times n}$ and $H_{n \times n}$, and denoted the resulting projections by $A_{n \times 500}^{RP}$ and $H_{n \times 500}^{RP}$. Then, we applied PCA on those matrices for further reduction, because $k = 500$ is still large compared to the number of inputs in X^{ph} and X^{EBV} . We retained the first 10 principal components from both matrices, and denoted the resulting matrices by X^P and X^G , both having a size of $n \times 10$.

Implications on Pen Assignment

Accurate prediction of slaughter age can have direct implications on the logistics of pig farming. Here, we illustrate the potential impact of slaughter age prediction if it were used for grouping pigs before the start of the finishing phase. To that end, we simulated 2 strategies for assigning a thousand

pigs to hundred finishing pens of equal capacity. In the first strategy, we assigned 10 pigs to each pen at random. Then, a pen was cleared after a number of days has elapsed that is equal to the average age at the slaughter weight of the pigs not used in the simulation, minus the average age at the start of finishing of the group in the pen. In the second strategy, we assigned 10 pigs to each pen according to the number of days needed to reach the slaughter weight, using predictions obtained by an RF model trained with a subset of different pigs. Then, each pen was cleared after a number of days has elapsed that is equal to the average slaughter age prediction of the group in the pen, minus their average age at the start of finishing. We evaluated each assignment strategy by counting the number of pigs in each pen that were—at the time of pen clearance—within a week or less of the actual age they reached slaughter weight. Each strategy was simulated a thousand times.

RESULTS

RF Regression Results

In Table 1, we report the regression performance, measured by R^2 and RMSE, and averaged over the 10 test subsets. Since we chose to use 10 principal components to construct X^P and X^G for the main results, we also included the results if other criteria are chosen for the number of used principal components (Table 2). We used the following criteria: 1) the *Eigenvalue-one* criterion (Kaiser, 1960), i.e., keeping the principal components with a corresponding Eigenvalue greater or equal to 1; 2) keeping the leading principal components with an accumulated explained variance (EV) higher than a chosen proportion of the total variance, which we choose here as 90%; 3) the first principal component only, used here to evaluate the accuracy of the most parsimonious option; 4) using the best 10 principal components for the target output, by running RF with the entire set of principal components as input matrix, and finding the top 10 ranking components; and finally, 5) using the input matrices obtained by RP, without further reduction by PCA. The proportions of EV of the first 40 principal components in both matrices are given in Fig. 1.

Feature Importance

The feature importance scores are given in Fig. 2. The scores were derived from training a RF

Table 1. The performance of random forest regression (RF) and multiple linear regression (MLR) for the following input matrices and their combinations: phenotype input matrix (X^{ph}), EBV input matrix (X^{EBV}), pedigree similarity input matrix (X^P), genetic-pedigree similarity input matrix (X^G), and all input features (X)

Input matrix	R^2		RMSE	
	RF	MLR	RF	MLR
X^{ph}	0.625 ± 0.009	0.580 ± 0.009	0.612 ± 0.009	0.648 ± 0.008
X^{EBV}	0.387 ± 0.012	0.124 ± 0.006	0.783 ± 0.009	0.936 ± 0.017
X^P	0.395 ± 0.011	0.218 ± 0.010	0.777 ± 0.010	0.884 ± 0.013
X^G	0.347 ± 0.013	0.206 ± 0.014	0.808 ± 0.010	0.891 ± 0.011
$[X^{ph} X^{EBV}]$	0.641 ± 0.009	0.596 ± 0.010	0.599 ± 0.009	0.635 ± 0.009
$[X^{ph} X^P]$	0.640 ± 0.009	0.589 ± 0.010	0.599 ± 0.009	0.641 ± 0.009
$[X^{ph} X^G]$	0.634 ± 0.009	0.586 ± 0.010	0.604 ± 0.009	0.643 ± 0.009
$[X^{EBV} X^P]$	0.405 ± 0.011	0.253 ± 0.010	0.771 ± 0.009	0.864 ± 0.015
$[X^{EBV} X^G]$	0.398 ± 0.012	0.261 ± 0.013	0.775 ± 0.010	0.860 ± 0.012
$[X^P X^G]$	0.395 ± 0.011	0.238 ± 0.013	0.777 ± 0.010	0.873 ± 0.012
$[X^{ph} X^{EBV} X^P]$	0.646 ± 0.008	0.599 ± 0.010	0.594 ± 0.008	0.633 ± 0.009
$[X^{ph} X^{EBV} X^G]$	0.644 ± 0.008	0.603 ± 0.010	0.597 ± 0.009	0.630 ± 0.009
$[X^{ph} X^P X^G]$	0.642 ± 0.008	0.593 ± 0.010	0.598 ± 0.008	0.638 ± 0.009
$[X^{EBV} X^P X^G]$	0.414 ± 0.011	0.281 ± 0.013	0.765 ± 0.010	0.848 ± 0.013
X	0.646 ± 0.008	0.605 ± 0.010	0.594 ± 0.008	0.628 ± 0.009

The performance was measured by R^2 and root mean squared error (RMSE) between the output and the prediction. Both metrics were evaluated on 10 test subsets of 10-fold cross-validation; the mean and standard deviation over the 10 subsets are reported.

Table 2. The performance of random forest regression, measured by R^2 , and averaged over 10 test subsets of 10-fold cross-validation, for different input matrices, and different choices of principal component to construct the pedigree similarity input matrix (X^P) and the genetic-pedigree similarity input matrix (X^G)

Principal components (PC)	Input matrix			
	X^P	X^G	$[X^{ph} X^P]$	$[X^{ph} X^G]$
First PC	-0.0368	-0.2426	0.6280	0.6289
First 10 PC	0.3953	0.3470	0.6409	0.6345
Best 10 PC	0.3978	0.3507	0.6411	0.6368
Eigenvalue-one	0.3953	0.3495	0.6409	0.6353
Accumulative EV > 90	0.4048	0.3751	0.6445	0.6369
$X^P = A_{n \times 500}^{RP}$, $X^G = H_{n \times 500}^{RP}$	0.3896	0.3658	0.6153	0.5980

on one of the training subsets, and they were normalized by the score of the most important feature.

Figure 3 shows the same cross-validated R^2 scores in Table 1, with the addition of the accumulated importance scores of the features in each input matrix. The accumulated score of an input matrix is shown as the proportion of the bar having a color corresponding to that input matrix.

Implications on Pen Assignment

We found that the first strategy, which grouped pigs randomly, resulted in 2.87 pigs per pen (of 10 pigs) on average within a week or less of their actual slaughter age. The second strategy, that is using the predictive model that we propose, resulted in 5.27 pigs per pen on average. This simulation demonstrates a significant improvement that is achieved by the proposed model.

DISCUSSION

Machine learning methods have been used for studying animal growth in the past. Yu et al. (2006) compared traditional statistical regression methods to neural networks in the task of predicting the average growth of groups of shrimp, using age, feed intake, water temperature, and biomass as predictors. Similarly, Roush et al. (2006) found that neural networks outperformed the nonlinear Gompertz growth equation in the prediction of broiler growth, using daily age/weight pairs for training their network. In those studies, the time window of prediction is as short as a week in (Yu et al., 2006), and one day in (Roush et al., 2006). Therefore, few predictors and training samples were needed to achieve good predictive performances, even with traditional regression methods. Conversely, we tried to predict the age of pigs at

Figure 1. The proportion of explained variance (EV) and accumulated EV of the first 40 principal components of $A_{n \times 500}^{RP}$ and $H_{n \times 500}^{RP}$.

Figure 2. The random forest feature importance scores, when X (all input features) is used as an input matrix. The scores are normalized by the score of the most important feature.

120 kg, which has an average value of 183 d; and we did so using predictors obtained before the start of the finishing stage, at which the pigs are 77 d old on average. This large time window of more than 100 d makes the predictive task significantly more difficult, requiring more predictors and a larger number of training samples.

More closely related to our study, Apichottanakul et al. (2012) used neural networks to predict the average weight of a group of pigs at the end of a finishing cycle. They used a number of different predictors, including average age and initial weight,

number of piglets in the group, survival rate, feed intake, and the average number of fattening days. The unit of prediction in that study is a group of pigs, with a range of 200 to 1,099 pigs per group. In this study, we focused instead on the growth prediction of individual pigs. Despite this being a more ambitious task than that of predicting the average weights of groups of pigs, we postulate that using an individual pig's growth as the prediction target serves as the baseline for more elaborate and practical predictive tasks, such as assigning pigs to uniform target weight groups (Alsahaf et al., 2018).

Figure 3. The performance of random forest regression for different input matrices, measured by the R^2 , and averaged over 10 test subsets of 10-fold cross-validation. The proportion of the colors within each bar represents the relative accumulated importance of the input matrix that the color represents: X^{ph} (light gray), X^{EBV} (dark gray), X^P (dashed lines), X^G (cross-pattern).

Besides growth prediction, machine learning methods have been used to solve other problems in animal science. The RF algorithm in particular has been applied in identifying additive and epistatic genes associated with residual feed intake in dairy cattle (Yao et al., 2013), identifying geographic patterns of different pig production systems (Thanapongtharm et al., 2016), and predicting the insemination outcome of dairy cattle (Shahinfar et al., 2014).

The prediction performance metrics in Table 1 show an advantage of RF regression over MLR. The advantage is clearer with input matrices other than X^{ph} or combinations that do not contain it, and less so otherwise. The most extreme example of this being X^{EBV} , which achieves $R^2 = 0.387$ with RF and $R^2 = 0.124$ with MLR. A possible explanation for this is that X^{ph} contains input features that are linearly correlated with the regression output, age at 120 kg, like the age at 30 kg of the animal or the average age at 120 kg of farm-mates (Supplementary Fig. A.4), thereby enabling MLR to achieve a similar performance to RF, whereas the remaining input matrices, X^{EBV} , X^P , and X^G are likely to contain more complex nonlinear dependencies that MLR is not able to exploit. In the literature, RF was found to have better predictability

than MLR in different applications, such as predicting fire occurrences (Oliveira et al., 2012), estimating biomass using from satellite images (Mutanga et al., 2012), and the prediction of protein-ligand binding affinity in biochemistry (Li et al., 2014).

In the concerned application, we use multiple features that are correlated. This apparent redundancy is an intrinsic feature of many machine learning methods. While traditional statistical methods achieve better fitting by increasing the complexity of the model (e.g., polynomial instead of linear regression), “model-free” (also called “nonparametric”) machine learning methods achieve better fitting by using multiple features that may be mutually dependent.¹ In both cases this may lead to overfitting and inflated performance on the *training data* and a validation or test set must be used to decide if the fitting is “just right” or overdone.

The advantages of using RF in this application extend to the interpretability of the model, through the use of the internal variable importance score. The ranking of input features given by RF (Fig. 2) makes intuitive sense, especially if we examine the top 5 ranking features, which are all phenotypes related to growth. Figure 3 shows the accumulative score of different feature matrices when used in combination with each other. An important observation to make here is that X^{EBV} , X^P , and X^G seem to explain a similar amount of the output’s variance, and contribute a similar increase in performance when combined with X^{ph} ; suggesting that they contain redundant information. This implies that the 2-step reduction procedure we propose in this paper, based on RP and PCA, extracts useful information from pedigree and pedigree-genetic matrices, $A_{n \times n}$, $H_{n \times n}$, with minimal computation requirements. With the reduction of $A_{n \times n}$, X^P explains more variance than with the reduction of $H_{n \times n}$, X^G . This is likely due to the fact that only a subset of the animals were genotyped in the construction of $H_{n \times n}$, causing it to be less informative than $A_{n \times n}$, which is based on pedigree only. It is worth investigating in future work if $H_{n \times n}$ constructed from n genotyped animals would be more predictive than $A_{n \times n}$.

There are 2 caveats to using feature importance scores derived from RF. First, the score can be biased to categorical features that have too many categories (Strobl et al., 2007). This is not an issue with the data in this study, since the categorical features—sex, fostering, and farm of birth—have 2 or 3 categories. The second caveat concerns highly correlated features. Due to the random sampling

¹This strategy is fundamental for kernel methods.

of features at each node, it should be expected that if a group of highly correlated features exist in the data set, and one of them is randomly selected at a node, the impurity that this feature removes will not be removed by features correlated to it in the same tree, hence making it “more important” than those other features, according to this importance score definition. This problem is partially mitigated by using a sufficiently large number of trees, and interpreting the importance scores of multiple RFs, fitted on different subsamples of the data.

Besides feature selection, the architecture of decision trees—the building blocks of RF—is simple. In fact, it is possible to visualize a decision tree trained on the available data to infer the rules that partition the samples, leading to further understanding of the prediction problem. In [Supplementary Appendix 2](#), we show 2 such visualizations of decision trees, one trained with all available samples and features, and another trained with X^{EBV} . These elements of RF, namely, feature ranking and interpretable architecture, are in contrast to the often unfair characterization of machine learning models as uninterpretable black boxes. Accordingly, this makes RF models suited for prediction problems in livestock science, as they can provide the end users—farmers or breeders—with insights about the data ([Ribeiro et al., 2016](#); [Doshi-Velez and Kim, 2017](#)). For more on the interpretability of RF and tree ensembles, we refer to [Petkovic et al. \(2018\)](#), [Pereira et al. \(2018\)](#), [Hara and Hayashi \(2016\)](#).

The reduction of $A_{n \times n}$ and $H_{n \times n}$ was achieved using RPs and PCA. Given a data set with a large number of variables, PCA finds through an orthogonal transformation, a lower number of variables that explain the largest proportion of variance possible from the original data set. For very high dimensional data sets, however, PCA is computationally expensive. An effective and computationally efficient alternative to PCA is RP, which works by projecting the high dimensional data set onto a lower dimensional subspace with a random matrix with unit-length columns. It has been shown that this method, despite its computational simplicity, is effective for many variable reduction applications on different types of data sets, e.g., text and image data ([Bingham and Mannila, 2001](#)), speeding-up k -means clustering ([Boutsidis et al., 2015](#)), clustering microarray data ([Avogadri and Valentini, 2009](#)), and speeding-up nearest neighbor classification ([Deegalla and Bostrom, 2006](#)). For more on the usage of PCA with RP, we refer to [Qi and Hughes \(2012\)](#) and [Anaraki and Hughes \(2014\)](#).

CONCLUSIONS

In conclusion, we showed that RF regression, a nonparametric machine learning algorithm, is effective in the prediction of slaughter age of pigs at the start of the finishing period. The methodology we described, namely the prediction of a future phenotype with machine learning through the use of phenotypic and genotypic data, could be applied to other phenotype prediction problems in pigs or other species. Using the feature importance scores of RF, we showed that phenotypes related to slaughter age are the most predictive group of features, compared to EBVs, and features extracted from pedigree and pedigree-genetic similarity matrices. Further exploration is still needed—using nonlinear feature selection—to address feature redundancy and find parsimonious subsets of input features for this prediction problem.

SUPPLEMENTARY DATA

Supplementary data are available at *Journal of Animal Science* online.

ACKNOWLEDGMENTS

This work is part of the Breed4Food research programme, project “Smart animal breeding with advanced machine learning” with project number 14295, which is financed by the Netherlands Organisation for Scientific Research (NWO), the Dutch Ministry of Economic Affairs (TKI Agri & Food project 12018) and the Breed4Food Partners Cobb Europe CRV, Hendrix Genetics, and Topigs Norsvin. The data for this study were provided by Topigs Norsvin (Beuningen, The Netherlands).

LITERATURE CITED

- Anaraki, F. P., and S. Hughes. 2014. Memory and computation efficient PCA via very 371 sparse random projections. In *International Conference on Machine Learning 2014* Jan 27 372 (pp. 1341–1349).
- Alsahaf, A., G., Azzopardi, B., Ducro, R. F., Veerkamp, and N. Petkov. 2018. Assigning pigs to uniform target weight groups using machine learning. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, vol. Species - Porcine 1:112.
- Apichottanakul, A., S. Pathumnakul, and K. Piewthongngam. 2012. The role of pig size prediction in supply chain planning. *Biosyst. Eng.* 113:298–307. doi: 10.1016/j.biosystemseng.2012.07.008
- Avogadri, R., and G. Valentini. 2009. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. *Artif. Intell. Med.* 45:173–183. doi:10.1016/j.artmed.2008.07.014

- Bingham, E., and H. Mannila. 2001. Random projection in dimensionality reduction. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 01. doi:10.1145/502512.502546
- Boutsidis, C., A. Zouzias, M. W. Mahoney, and P. Drineas. 2015. Randomized dimensionality reduction for k-means clustering. *IEEE Trans. Inf. Theory* 61:1045–1062. doi:10.1109/tit.2014.2375327
- Breiman, L. 1996. Bagging predictors. *Mach. Learn.* 24:123–140. doi:10.1007/bf00058655
- Breiman, L. 2001a. Random forests. *Mach. Learn.* 45:5–32. doi:10.1023/A:101093340
- Breiman, L. 2001b. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16:199–231. doi:10.1214/ss/1009213726
- Deegalla, S., and H. Bostrom. 2006. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In 2006 5th International Conference on Machine Learning and Applications (ICMLA06). doi:10.1109/icmla.2006.43
- Doshi-Velez, F., and B. Kim. 2017. Towards a rigorous science of interpretable machine learning. eprint arXiv:q-alg/9503002
- Geurts, P., D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Mach. Learn.* 63:3–42. doi:10.1007/s10994-006-6226-1
- Gonyou, H. 1998. Sorting and mixing of grower/finisher pigs. Proceedings of the Allen D. Leman Swine Conference. p. 126–128.
- Hara, S., and K. Hayashi. 2016. Making tree ensembles interpretable. eprint arXiv:1606.05390
- Johnson, W. B., and J. Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. Conference on Modern Analysis and Probability Contemporary Mathematics. p. 189–206. doi:10.1090/conm/026/737400
- Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20:141–151. doi:10.1177/001316446002000116
- Kim, S. H., S. H. Park, and Y. C. Choe. 2014. Effects using information system in pig production. *Adv. Sci. Technol. Lett.* 49:206–212. doi:10.14257/astl.2014.49.40
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy. Sci.* 92:4656–4663. doi:10.3168/jds.2009-2061
- Li, P., T. J. Hastie, and K. W. Church. 2006. Very sparse random projections. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 06. doi:10.1145/1150402.1150436
- Li, H., K. S. Leung, M. H. Wong, and P. J. Ballester. 2014. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. *BMC Bioinform.* 15:291. doi:10.1186/1471-2105-15-291
- Ma, C., Y. Li, G. Yin, and J. Ji. 2012. The monitoring and information management system of pig breeding process based on internet of things. 2012 Fifth International Conference on Information and Computing Science. doi:10.1109/ici.2012.61
- Mutanga, O., E. Adam, and M. A. Cho. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth. Obs. Geoinf.* 18:399–406. doi:10.1016/j.jag.2012.03.012
- Oliveira, S., F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J. M. Pereira. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecol. Manag.* 275:117–129. doi:10.1016/j.foreco.2012.03.003
- Patience, J. F., K. Engele, A. D. Beaulieu, H. W. Gonyou, and R. T. Zijlstra. 2004. Variation: costs and consequences. *Adv. Pork Prod.* 15:257–66.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12:2825–30.
- Pereira, S., R. Meier, R. McKinley, R. Wiest, V. Alves, C. A. Silva, and M. Reyes. 2018. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. *Med. Image Anal.* 44:228–244. doi:10.1016/j.media.2017.12.009
- Petkovic, D., R. Altman, M. Wong, and A. Vigil. 2018. Improving the explainability of random forest classifier—user centered approach. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, vol. 23, pp. 204–215. doi:10.1142/9789813235533_0019
- Qi, H., and S. M. Hughes. 2012. Invariance of principal components under low-dimensional random projection of the data. 2012 19th IEEE International Conference on Image Processing. doi:10.1109/icip.2012.6467015
- Ribeiro, M., S. Singh, and C. Guestrin. 2016. “Why should i trust you?”: explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. doi:10.18653/v1/n16-3020
- Roush, W. B., W. A. Dozier, and S. L. Branton. 2006. Comparison of gompertz and neural network models of broiler growth. *Poult. Sci.* 85:794–797. doi:10.1093/ps/85.4.794
- Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. Weigel. 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *J. Dairy. Sci.* 97:731–742. doi:10.3168/jds.2013-6693
- Strobl, C., A. L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinform.* 8:25. doi:10.1186/1471-2105-8-25
- Thanapongtharm, W., C. Linard, P. Chinson, S. Kasemsuwan, M. Visser, A. E. Gaughan, M. Epprech, T. P. Robinson, and M. Gilbert. 2016. Spatial analysis and characteristics of pig farming in Thailand. *BMC Vet. Res.* 12:218. doi:10.1186/s12917-016-0849-7
- Yao, C., D. M. Spurlock, L. E. Armentano, C. D. Page, M. J. VandeHaar, D. M. Bickhart, and K. A. Weigel. 2013. Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *J. Dairy Sci.* 96:6716–6729. doi:10.3168/jds.2012-6237
- Yu, R., P. Leung, and P. Bienfang. 2006. Predicting shrimp growth: artificial neural network versus nonlinear regression models. *Aquac. Eng.* 34:26–32. doi:10.1016/j.aquaeng.2005.03.003