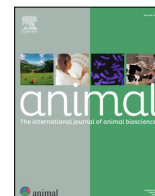




Animal

The international journal of animal biosciences



Classification of light Yorkshire pigs at different production stages using ordinary least squares and machine learning methods



J. Casellas^a, P. Salgado-López^{b,*}, J. Lorente^{c,1}, I. Solar Diaz^d, T. Rathje^d, J. Gasa^b, D. Solà-Oriol^b

^a Department of Animal and Food Science, Autonomous University of Barcelona, Bellaterra 08193, Spain

^b Animal Nutrition and Welfare Service (SNIBA), Department of Animal and Food Science, Autonomous University of Barcelona, Bellaterra 08193, Spain

^c Andrimmer Genética Aplicada, Calvet 30-32, 3^o 2^a, 08021, Barcelona, Spain

^d DNA Genetics LLC, Columbus, NE 68601, USA

ARTICLE INFO

Article history:

Received 23 June 2023

Revised 26 November 2023

Accepted 28 November 2023

Available online 5 December 2023

Keywords:

Area under the curve

Artificial intelligence

Growth retardation

Live weight

Swine

ABSTRACT

Pig homogeneity and growth are major concerns for the pig industry today. Variability in pigs' size has a strong impact on profitability as uniformity plays a key role in the overall economic value of pigs produced. This research focused on statistical methods to identify pigs at risk of growth retardation at different stages of production. Data from 125 083 Yorkshire pigs at weaning (18–28 d), 59 533 pigs at the end of the nursery period (70–82 d) and 48 862 pigs at slaughter (155–170 d) were analyzed under three different cut-points (lowest 10, 20 and 30%) to characterize light animals. Records were randomly split into 2:1 training:testing sets, and each training data set was analyzed through an ordinary least squares approach and four machine learning algorithms (decision tree, random forest, and two alternative boosting approaches). A wide range of weighting functions were applied to give increased relevance to lighter pigs. Each resulting classification norm was used to classify light pigs in the testing data set. Both sensitivity and specificity were retained to construct the receiver operating characteristic curve, and the statistical performance of each analytical approach was evaluated by the area under the curve (AUC). In all production stages and cut-points, the random forest machine learning algorithm provided the highest AUC, closely followed by boosting procedures. For weaning BW (WW), factors related to birth BW and litter size accounted for more than 75% of the important prediction factors for light pigs. BW at the end of the nursery period and slaughter BW analyses revealed a similar pattern where WW and BW at the end of the nursery period accounted for more than 40 and 50% of statistical importance among the prediction factors, respectively. Machine learning algorithms are useful tools to easily evaluate the risk factors affecting the efficiency and homogeneity in swine. Since the BW at birth and weaning are key factors, sow nutrition and feeding management during gestation and lactation, along with piglet management during lactation, are identified as important influences on pig weight variability.

© 2023 The Author(s). Published by Elsevier B.V. on behalf of The Animal Consortium. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Implications

Providing robust tools to identify pigs at risk of growth retardation will help to develop better management strategies for this subset of pigs. This research has tested five different classification algorithms (ordinary least squares regression; decision tree; random forest; two alternative boosting approaches) to determine their ability to classify lightweight pigs at different production stages. The random forest technique had the best performance, followed closely by boosting approaches. The classification ability of

these methods increased when focused on more extreme individuals. Weight-related traits at previous ages played the key role in the classification of light animals.

Introduction

The current swine production systems maintain contemporary pigs in the same batch as it is recommended for optimal animal health (Maes et al., 2004). One of the main important factors affecting the efficiency of pig production in these all-in-all-out systems is the variability of BW at slaughter (Patience et al., 2004). High variability in market weight reduces the value of the carcasses by worsening their quality classification and price quotation, reduces the effectiveness of slaughterhouse equipment and increases the

* Corresponding author.

E-mail address: pau.salgado@uab.cat (P. Salgado-López).

¹ Present address: Master in Swine Health and Production. International Swine technical advisor.

occupation time in finishing farm facilities (López-Vergé et al., 2018), although with relevant economic consequences on earlier stages (Larriestra et al., 2006). These effects represent the main shadow production cost for the swine industry. There is a plethora of scientific literature aiming to identify the environmental (Botermans et al., 2000; Quiniou et al., 2002), behavioral (Georgsson and Svendsen, 2002; Carcò et al., 2018), nutritional and feeding (Brossard et al., 2009; Solà-Oriol and Gasa, 2017; Blavi et al., 2021), and genetic factors (Ibañez-Escriche et al., 2008; Peltoniemi et al., 2021), among others, that contribute to variability in pig weights under commercial conditions.

Pigs with a low BW at the beginning of the grow-finishing barn are expected to reach market BW later than their faster-growing counterparts (Douglas et al., 2014; López-Vergé et al., 2018). Nevertheless, little is known about the usefulness of prior performance information to effectively identify those animals at greatest risk of growth retardation in sufficient time for intervention to improve outcome. Indeed, this implies a change in the analytical paradigm, moving from a goodness of fit of cause-effect parametrizations, to the predictive ability of methodologies aimed at the outcome (Steyerberg et al., 2010).

Advances in both computer performance and classification algorithms have been significant in recent decades, and these advances are providing reliable alternatives to the well-known parametric linear model for the analysis of continuous traits. Machine learning, a branch of the artificial intelligence research field, must be viewed as a paradigmatic example that allow computers to evolve behavior based on empirical data (Mitchell, 1997). There are many prediction methods with potential contributions to the swine industry (Quinlan, 1987; Friedman, 2001; Chen and Guestrin, 2016; Marvin and Ziegler, 2017), although little is known about their ability to identify pigs at risk of growth retardation. Large data sets become mandatory to apply these methods and to determine the best classification tool based on a representative population and avoiding sampling biases.

In order to confidently intervene with effective management strategies for low BW and low growth pigs, the industry would benefit from robust tools to identify these pigs. In that sense, the present study was conducted based on a robust and powerful database available to accurately classify (1) the piglets that would be the main cause of variation within a population and (2) linking and identifying the main production factors that explain low pig BW at each phase within the pig production cycle with accuracy and robustness.

The objective of this research focused on the classification of light pigs at different production stages. More specifically, we tested five different classification algorithms on BW data from more than 100 000 Yorkshire pigs, evaluating three independent weight cut-off points for light pigs and their performance across different production stages. The patterns of sensitivity and specificity were characterized for each method by constructing receiver operating characteristic (ROC) curves. Factors with higher contributions to the final category were identified and discussed.

Material and methods

Animal Care and Use Committee approval was not necessary for this study because analyses were performed on existing field data obtained under standard farm management from Andrimner Genética Aplicada (Barcelona, Spain) and DNA Genetics (Columbus, NE, USA) selection farms. Both productive and pedigree data were collected by trained technicians within the context of the genetic selection program of the Yorkshire population involved in this study.

Selection nuclei pig data

All analyses focused on Line 200 Yorkshire (DNA Genetics) nucleus data set, collected in four farms from both North America and Europe from January 2014 to March 2021. Animals were kept under standard intensive farming conditions, and weighed at: birth (BIW), weaning (WW, 18–28 d), the end of the nursery period (70–82 d), and slaughter (155–170 d). The distribution of WW, BW at the end of the nursery period, and slaughter BW, as well as cut-points (lowest 10, 20 and 30%) are presented in Fig. 1. Age-related differences at each weigh point were removed for WW, BW at the end of the nursery period and slaughter BW by a standard linear adjustment based on total data at 21 d, 75 d and 160 d, respectively (Table 1). For BW at the end of the nursery period and slaughter BW, there was a 3.7-fold increase in data from females compared to males. This discrepancy arose because all females were retained for the grow-finisher period, whereas only a certain number of maternal boars were selected to continue in that phase. Additional data for each piglet (i.e., sex, death cause and date, and cross-fostering date and adoption litter after cross-fostering, if any), sow (i.e., parity number and farm) and litter levels (i.e., birth date, litter size at birth and after cross-fostering, if any) were systematically collected by trained farm staff.

Cross-fostering protocol

Cross-fostering is a management technique which could be performed in several ways. The cross-fostering protocol for the present study was based on the following steps: (1) fostering took place within the first 24 hours of birth and after 24 hours any fostering was only because of a lactation failure, (2) loading as many sows as possible to nurse a number of pigs equal to their functional teat count + 1 more pig and (3) when litter sizes exceeded functional teat count + 1, piglets were fostered to smaller litters (this determination was made within the first 24 hours). The biggest piglets were the first ones to be moved, as they seem to be able to withstand easier the changes in the environment. This process is similar to the protocols used in high-health farms; recommending keeping litters intact as much as possible, doing the minimum number of changes and applying an intervention only to balance the number of pigs per litter (Muns et al., 2014). Within this context, 73.4% of litters had no fostered pigs in this Yorkshire line, whereas 9.6, 7.2 and 9.2% of the litters had 1, 2 or more than two fostered pigs, respectively. Protocols for lower health status farms usually make litters including only or mainly small piglets to help and protect those animals during lactation (Alexopoulos et al., 2018). Other strategies such as creep feeding or milk replacers, together with the piglet's good vitality, may help to reduce the risk factors to become a slow-growing pig at weaning and thereafter (Muns and Tummaruk, 2016).

Operational models

Prediction of WW data focused on the following model,

$$WW_{ijklmnopq} = BIW_i + Sex_j + Cross - fostered_k + DB_i + DC_i + VB_i + VC_i + Stillbirths_m + Parity\ number_n + Litter\ size_o + Farm_p + Season_q$$

where BIW_i was the birth BW of the i th piglet, Sex_j was the sex of the i th piglet (male or female), $Cross - fostered_k$ identified if cross-fostered or not, DB_i (DC_i) was the difference between its BIW and the average litter weight at birth (after cross-fostering), VB_i (VC_i) was the within-litter variance of BW at birth (after cross-fostering), $Stillbirths_m$ were the number of stillbirths in the litter (0, 1, 2, 3, 4 or >4), $Parity\ number_n$ was the parity number of the sow (1, 2, 3, 4, 5 or >5), $Litter\ size_o$ was the litter size after cross-

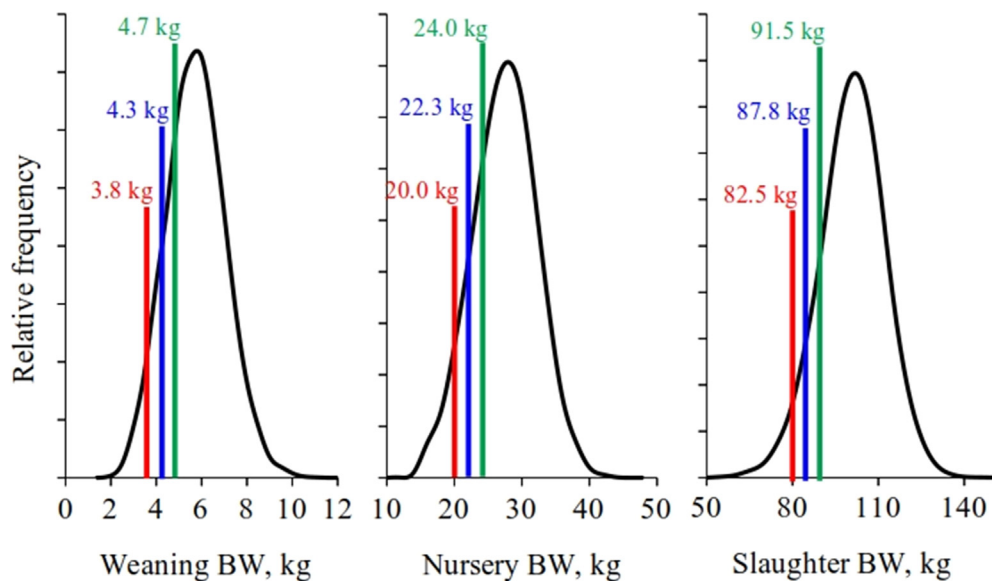


Fig. 1. Distribution of weaning (125 083), end of the nursery period (59 533), and slaughter (48 862) BWs from Yorkshire pigs, as well as cut-points for the lowest 10% (red), 20% (blue) and 30% (green). Abbreviation: Nursery BW = BW at the end of the nursery period.

Table 1
Summary of weaning BW, BW at the end of the nursery period, and slaughter BW in the Yorkshire pig population.

Item	Weaning BW, kg		BW at the end of the nursery period, kg		Slaughter BW, kg	
	n	Mean ± SE	n	Mean ± SE	n	Mean ± SE
Overall	125 083	5.456 ± 0.007	59 533	26.391 ± 0.021	48 862	96.820 ± 0.045
Sex						
Male	63 079	5.606 ± 0.009	12 551	27.259 ± 0.047	10 309	98.334 ± 0.098
Female	62 004	5.304 ± 0.009	46 982	26.159 ± 0.024	38 553	96.416 ± 0.050
Cross-fostered						
No	117 495	5.470 ± 0.007	55 292	26.380 ± 0.022	45 448	96.737 ± 0.046
Yes	7 588	5.239 ± 0.026	4 241	26.523 ± 0.079	3 414	97.930 ± 0.169
Sow parity						
1	64 594	5.255 ± 0.009	31 714	25.762 ± 0.029	25 908	95.745 ± 0.061
2	32 725	5.657 ± 0.013	14 693	27.145 ± 0.043	11 815	98.141 ± 0.091
3	15 171	5.642 ± 0.019	6 926	27.034 ± 0.063	5 883	98.345 ± 0.129
4	7 061	5.750 ± 0.029	3 554	27.149 ± 0.087	3 026	97.793 ± 0.180
5	3 693	5.684 ± 0.039	1 720	26.888 ± 0.125	1 429	96.927 ± 0.260
>5	1 839	5.829 ± 0.056	926	27.303 ± 0.172	801	97.039 ± 0.349
Farm ¹						
A	8 522	5.317 ± 0.025	4 117	25.740 ± 0.079	2 438	94.361 ± 0.197
B	14 951	5.143 ± 0.019	2 647	23.527 ± 0.094	1 418	91.881 ± 0.255
C	17 574	5.434 ± 0.018	11 170	27.207 ± 0.049	9 898	98.428 ± 0.100
D	84 036	5.530 ± 0.008	41 599	26.418 ± 0.025	35 108	96.738 ± 0.053

¹ Farms were anonymized due to privacy concerns and sorted by the number of available weaning BWs.

fostering (<6, 6–7, 8–9, 10–11, 12–13, 14–15, 16–17, 18–19, 20–21, 22–23, or >23), Farm_p was the farm with four levels, and Season_q was the season of birth on a 2 π -radian scale. The same was assumed for the remaining traits, after including WW_i as a predictor in the model for the analysis of BW at the end of the nursery period_{ijklmnopq}, and WW_i and BW at the end of the nursery period_i as predictors in the model for the analysis of slaughter BW_{ijklmnopq}, unless otherwise stated. Weight data were predicted by incorporating the inherent systematic effects associated with each individual pig.

The statistical importance of the prediction factors was assessed by quantifying the improvement in the performance measure associated with each attribute split point, weighted by the number of observations assigned to the respective node. The Gini index was used as a performance measure.

Weighting function

The main target of this research was the classification of light piglets at different growing stages. Given this objective, data from lighter animals is likely more informative than weights from heavier littermates, although there is not a standardized approach to appropriately weigh these records. To avoid bias, the following weighting function was applied,

$$\Psi^{-\delta/100}$$

where $\Psi = \{\text{WW, BW at the end of the nursery period, slaughter BW}\}$, and δ started from $\delta = 0$ and increased by one until the relative weight of average Ψ minus two SDs was 10 times larger than the relative weight of average Ψ (see Fig. 2 as an example plot of the

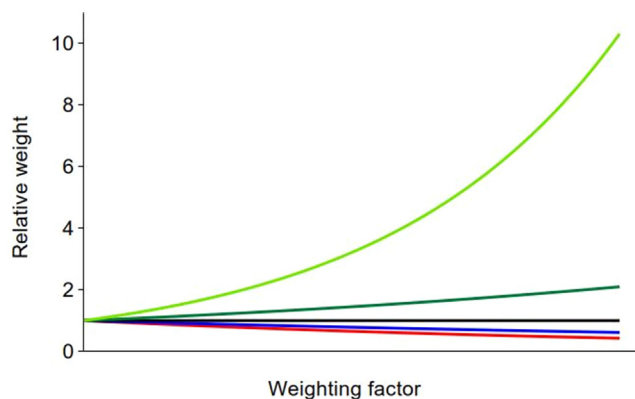


Fig. 2. Plot of the weighting function $\Psi^{-5/100}$ applied to classification models for Yorkshire pigs. Note that the black line was taken as reference ($\Psi = \psi$) whereas red ($\Psi = \psi + 2\kappa$), blue ($\Psi = \psi + \kappa$), dark green ($\Psi = \psi - \kappa$) and fair green ($\Psi = \psi - 2\kappa$) lines were obtained by adding or subtracting a constant amount (κ).

weighting function $\Psi^{-5/100}$). Average Ψ referred to the average of the individual WW of the data set (BW at the end of the nursery period, and slaughter BW in subsequent analyses). Within this context, analyses were repeated 144 times (WW data), 703 times (BW at the end of the nursery period data), and 801 times (slaughter BW data) to create each ROC curve.

Analytical approaches

All analyses were performed with R v.4.1.2 (R Core Team, 2020). Within a parametric space, a weighted linear model was fitted by ordinary least squares (*lm* function, *stats* package; R Core Team, 2020), and all factors were assumed continuous except for Stillbirths_m, Parity number_n, Litter size_o, Farm_p and Season_q. This last one was categorized by each $\pi/4$ radians.

Four machine learning approaches with varying levels of complexity were also implemented. From lower to higher complexity, we fitted a decision tree (*rpart* package, *anova* method; Quinlan, 1987), random forest with 1 000 trees (*ranger* package; Marvin and Ziegler, 2017) and two boosting approaches, the Generalized Boosted Regression model (*gbm* package; Friedman, 2001) and the Extreme Gradient Boosting (*xgboost* package; Chen and Guestrin, 2016). Both boosting approaches were set with identical parameters for number of trees (1 000), maximum interaction depth (5) and shrinkage (0.01), following Elith et al. (2008). These two boosting methods were employed because assuming identical analytical performance was not feasible and determining the superior one required their application to the data set.

The same parameters were set for all machine learning methods to establish a consistent framework for comparative purposes. Preliminary analyses were conducted on the data sets to ensure their utility and performance. Although this entailed higher computational costs, both the number of trees and interaction depth were fitted in excess to guarantee model fit in all cases. The same principle was applied to the shrinkage parameter, which was undervalued (see below). It is important to note that 1 000 trees exceeded the predicted number of trees with minimum cross-validation error (*i.e.*, which.min function) in all cases (results not shown). Using more trees resulted in a better fit, but the rate of improvement decreased as the number of trees increased, and marginal enhancements became negligible when approaching 1 000 trees. The maximum interaction depth is directly related to the number of relevant factors evaluated by the prediction rule, and typically behaves in a similar way to that of the number of trees (*i.e.*, the fit marginally increases with interaction depth). Within this context, preliminary analyses tested for interaction depths of 3, 5, 7, and 9. However, the area under the curve (AUC) matched the second decimal place from inter-

action depth 5 and above. Four different shrinkage parameters were preliminarily tested in all data sets and boosting approaches (0.1, 0.05, 0.01, and 0.001). The AUC did not differ beyond the fourth decimal place when comparing 0.01 and 0.001, whereas only the extreme gradient boosting revealed differences in the third decimal place for WW when comparing 0.05 and 0.1.

Cross-validation for classification ability

Each model was evaluated by a three-fold cross-validation approach. Each data set was randomly split into three subsets of equal size (33.3%). Two of them merged as training data set to fit the model, and its classification ability was evaluated against the remaining subset (*i.e.*, testing data set). Once the upper threshold for light pigs (Ψ_t) was defined, and the analysis was conducted for each value resulting from this threshold, classification ability was evaluated as the AUC when plotting sensitivity against 1 – specificity, *i.e.*, a receiver operating characteristic curve (Woodward, 1953; Peterson et al., 1954). Note that sensitivity referred to the proportion of light pigs (*i.e.*, $\Psi_i \leq \Psi_t$) classified as such by the classification rule, whereas specificity was the proportion of non-light pigs (*i.e.*, $\Psi_i > \Psi_t$) correctly classified as non-light pigs. Both sensitivity and specificity are standard parameters. The AUC characterizes the performance of the discrimination rule and fluctuates between 0.5 (*i.e.*, successful classifications are occurred by chance) and 1 (*i.e.*, perfect sensitivity and specificity). In general, the classification performance of any rule measured by the AUC can be categorized as worthless (<0.7), acceptable (0.7–0.8), excellent (0.8–0.9) and outstanding (>0.9) (Hosmer and Lemeshow, 2000).

Results

Classification of light BW pigs at weaning

The parametric linear model reached an intermediate classification ability when compared with machine learning methodologies, regardless of the cut-point assumed for WW. The single decision tree provided remarkably lower AUC estimates, whereas both boosting and random forest approaches increased AUC at the second decimal place (Table 2).

Classifications from random forest reached the highest AUC in all scenarios and increased AUC between 1.08% (cut-point, 4.7 kg) and 2.53% (cut-point, 3.7 kg) when compared with the second highest AUC (*i.e.*, extreme gradient boosting; Table 2). It is important to note that the advantage of random forest extended to all specificity values, and ROC curves only converged under complete or null specificity (Fig. 3). Moreover, AUC increased when reducing the cut-point for WW, although changes were smaller for methods with higher AUC estimates (Table 2).

When examining the importance of various values, only BIW_i, DB_i, DC_i, VB_i, VC_i, and Litter size_o accounted for 5% or more as shown in Fig. 4. Indeed, importance changed along the specificity space, although these six factors always captured more than 80%. Five of them were birthweight-related factors that, when combined, explained more than 75% of the specificity space for high specificity values, or even more than 80% (moderate to low specificity values). Dependence plots for all these factors were provided in Fig. 5.

Classification of light BW pigs at the end of the nursery period and slaughter

The random forest technique maintained its superiority when classifying light pigs at older stages, with an advantage between

Table 2
Area under the curve for five classification models used as discrimination tools to identify light pigs at weaning (i.e., live weight standardized to 21 d of age) under three different weight cut-points.

Item	Testing data set			Mean ± SE
	1	2	3	
Cut-point < 3.7 kg (10%)				
Linear model	0.729	0.732	0.732	0.731 ± 0.001
Decision tree	0.608	0.609	0.614	0.610 ± 0.002
Random forest	0.766	0.771	0.772	0.770 ± 0.002
Generalized boosted regression	0.748	0.746	0.748	0.747 ± 0.001
Extreme gradient boosting	0.752	0.752	0.750	0.751 ± 0.001
Cut-point < 4.3 kg (20%)				
Linear model	0.725	0.721	0.717	0.721 ± 0.002
Decision tree	0.579	0.582	0.584	0.582 ± 0.002
Random forest	0.762	0.759	0.761	0.761 ± 0.001
Generalized boosted regression	0.742	0.745	0.742	0.743 ± 0.001
Extreme gradient boosting	0.749	0.745	0.744	0.746 ± 0.002
Cut-point < 4.7 kg (30%)				
Linear model	0.721	0.722	0.727	0.723 ± 0.002
Decision tree	0.530	0.535	0.535	0.533 ± 0.002
Random forest	0.748	0.745	0.747	0.747 ± 0.001
Generalized boosted regression	0.735	0.736	0.739	0.736 ± 0.001
Extreme gradient boosting	0.738	0.742	0.738	0.739 ± 0.001

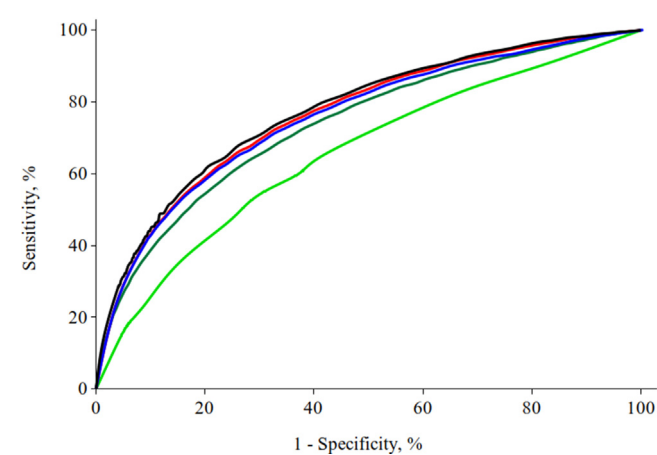


Fig. 3. Receiver operating characteristic curve for the classification of light pigs at weaning (<4.3 kg) calculated by a parametric linear model (dark green) and four machine learning procedures (single decision tree, fair green line; random forest, black line; generalized boosted regression, blue line; extreme gradient boosting, red line).

1 and 3% of AUC when compared with the next highest methodology based on the AUC (i.e., extreme gradient boosting; Tables 3 and 4). Higher AUC was observed for smaller cut-points for BW at the end of the nursery period and slaughter BW, and older production stages (i.e., slaughter BW vs. BW at the end of the nursery period).

The inclusion of WW as a prediction factor for the analysis of BW at the end of the nursery period had a substantial impact on the final performance of the discrimination rule. This is demonstrated by the importance plot shown in Fig. 6. When excluded, the AUC reduced between 0.05 (cut-point, 24.0 kg) and 0.07 (cut-point, 22.0 kg), and the importance of birthweight-related factors substantially increased (results not shown). A similar pattern was observed for the analysis of slaughter BW with regard to WW and BW at the end of the nursery period as prediction factors. Given the highest contribution of BW at the end of the nursery period (Fig. 7), its arbitrary exclusion from the model generated a remarkable reduction of the AUC, between 0.11 (cut-point, 91.5 kg) and 0.17 (cut-point, 82.5 kg). Dependence plots for factors accounting for more than 5% importance when classifying light pigs at the end of the nursery period and slaughter were provided in Figs. 8 and 9. When BW at the end of the nursery period and WW

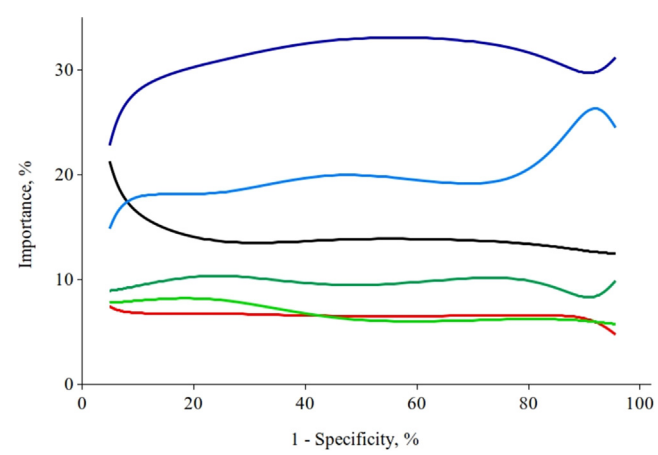


Fig. 4. Importance evolution for factors accounting for more than 5% when classifying light pigs at weaning (<4.3 kg) with random forest. Green lines were for within-litter birth BW variance, and blue lines referred to the difference between birth BW and average litter weight, both at birth (fair lines) and after cross-fostering (dark lines). The black and the red lines were for birth BW and litter size after cross-fostering, respectively.

were excluded together, these reductions increased up to 0.18 and 0.24, respectively (results not shown).

Thus, the impact on the ROC curves when BIW is only considered as main factor on the prediction for the slaughter BW is reported in Fig. 10, showing a reduction of the AUC when the BWs for the intermediate phases are not considered (WW and BW at the end of the nursery period).

Discussion

Classification ability trade-off for pig weight models

The ROC curve was initially developed for operators of military radar receivers during the decade of the 1940s (Woodward, 1953). This curve illustrates the relationship between sensitivity (Y-axis) and 1 – specificity (X-axis) of a binary classifier system as its discrimination threshold varies (i.e., light vs. non-light pig). The overall performance of the classification rule is typically evaluated by the AUC, thus summarizing all possible scenarios combining sensitivity and specificity. Nevertheless, the performance of AUC as an

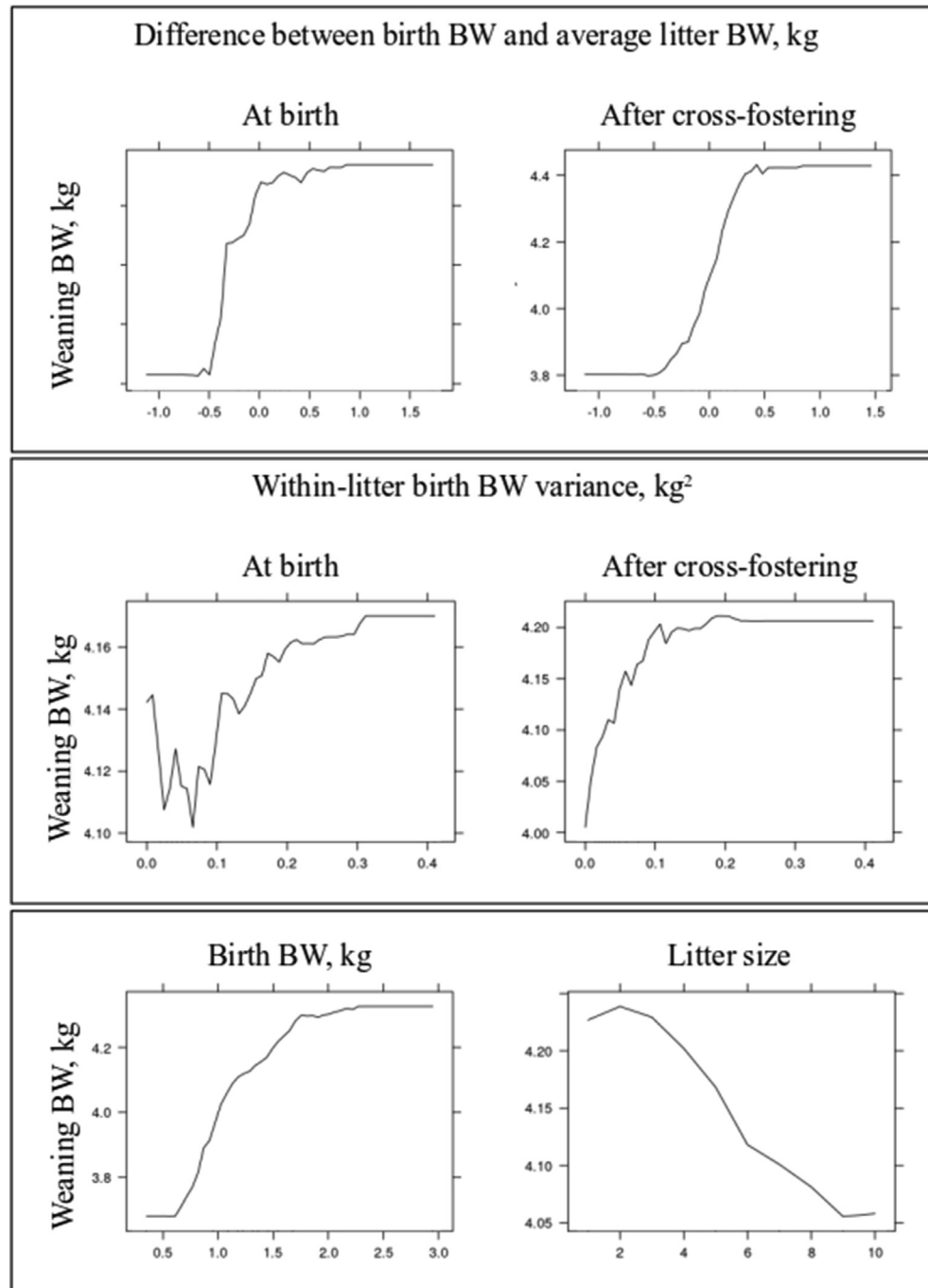


Fig. 5. Partial dependence plots for factors accounting for more than 5% importance when classifying light pigs at weaning (<4.3 kg) with random forest (sensitivity, 82%; specificity, 60%). Abbreviation: Nursery BW = BW at the end of the nursery period.

evaluation criterion for classification ability would suffer from statistical inconsistency if two or more ROC curves intersect at a given point. The ROC curve with higher sensitivity at a given interval of 1 – specificity values becomes preferable in that region, even if its overall AUC is smaller than those for competing ROC curves (Fawcett, 2006). It is important to highlight that this behavior has not been observed in our analyses and all ROC curves from different analytical methodologies did not intersect when calculated on the same data set. The classification advantage (or disadvantage) regarding each analytical methodology persisted along the curve and could be effectively captured by the AUC in our data sets.

Maximum AUC for all analyses showed acceptable discrimination performance (Hosmer and Lemeshow, 2000) with values ranging between 0.7 and 0.8, and even excellent (0.8–0.9) in some specific cases, although the trade-off between sensitivity and specificity was remarkable. Far from being a penalty for this research, this must be viewed as a flexible framework where the optimum combination of sensitivity and specificity can be defined according to management practices for light pigs during different production stages. A higher rate of false positives (*i.e.*, lower specificity) could be assumed for cheap management practices such as cross-fostering (Calderón-Díaz et al., 2018), where (almost) all

Table 3

Area under the curve for five classification models used as discrimination tools to identify light pigs at the end of the nursery period (i.e., live weight standardized to 75 d of age) under three different weight cut-points.

Item	Testing data set			Mean ± SE
	1	2	3	
Cut-point < 20.0 kg (10%)				
Linear model	0.771	0.767	0.767	0.768 ± 0.001
Decision tree	0.672	0.675	0.675	0.674 ± 0.001
Random forest	0.823	0.822	0.817	0.821 ± 0.002
Generalized boosted regression	0.794	0.792	0.795	0.794 ± 0.001
Extreme gradient boosting	0.803	0.805	0.806	0.805 ± 0.001
Cut-point < 22.3 kg (20%)				
Linear model	0.759	0.760	0.762	0.760 ± 0.001
Decision tree	0.639	0.644	0.643	0.642 ± 0.002
Random forest	0.812	0.814	0.816	0.814 ± 0.001
Generalized boosted regression	0.781	0.776	0.777	0.778 ± 0.001
Extreme gradient boosting	0.796	0.796	0.791	0.794 ± 0.002
Cut-point < 24.0 kg (30%)				
Linear model	0.751	0.753	0.749	0.751 ± 0.001
Decision tree	0.600	0.601	0.596	0.599 ± 0.001
Random forest	0.781	0.780	0.776	0.779 ± 0.001
Generalized boosted regression	0.766	0.768	0.766	0.767 ± 0.001
Extreme gradient boosting	0.769	0.773	0.774	0.772 ± 0.002

Table 4

Area under the curve for five classification models used as discrimination tools to identify light pigs at slaughter (i.e., live weight standardized to 160 d of age) under three different weight cut-points.

Item	Testing data set			Mean ± SE
	1	2	3	
Cut-point < 82.5 kg (10%)				
Linear model	0.871	0.874	0.877	0.874 ± 0.002
Decision tree	0.781	0.782	0.780	0.781 ± 0.001
Random forest	0.898	0.897	0.902	0.899 ± 0.001
Generalized boosted regression	0.870	0.873	0.878	0.874 ± 0.002
Extreme gradient boosting	0.875	0.873	0.869	0.872 ± 0.002
Cut-point < 87.8 kg (20%)				
Linear model	0.851	0.849	0.847	0.849 ± 0.001
Decision tree	0.723	0.720	0.717	0.720 ± 0.002
Random forest	0.874	0.876	0.873	0.874 ± 0.001
Generalized boosted regression	0.859	0.855	0.858	0.858 ± 0.001
Extreme gradient boosting	0.860	0.863	0.866	0.863 ± 0.002
Cut-point < 91.5 kg (30%)				
Linear model	0.839	0.835	0.838	0.837 ± 0.001
Decision tree	0.661	0.659	0.661	0.660 ± 0.001
Random forest	0.861	0.860	0.860	0.860 ± 0.001
Generalized boosted regression	0.848	0.844	0.845	0.846 ± 0.001
Extreme gradient boosting	0.852	0.848	0.852	0.851 ± 0.001

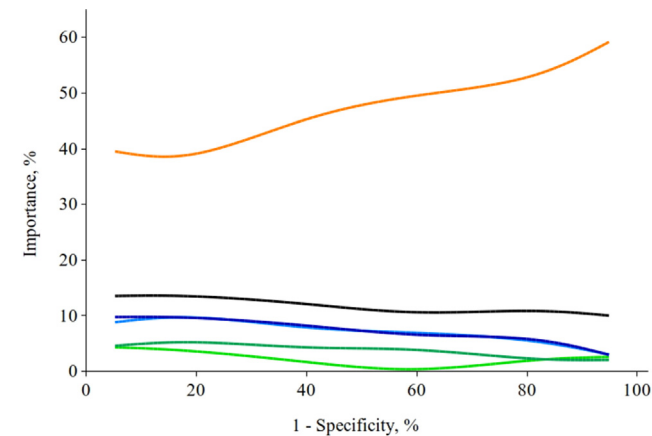


Fig. 6. Importance evolution for factors accounting for more than 5% when classifying light pigs at 75 d old (<22.3 kg) with random forest. Green lines were for within-litter birth BW variance, and blue lines referred to the difference between birth BW and average litter weight, both at birth (fair lines) and after cross-fostering (dark lines). Remaining lines referred to birth (black) and weaning BW (orange).

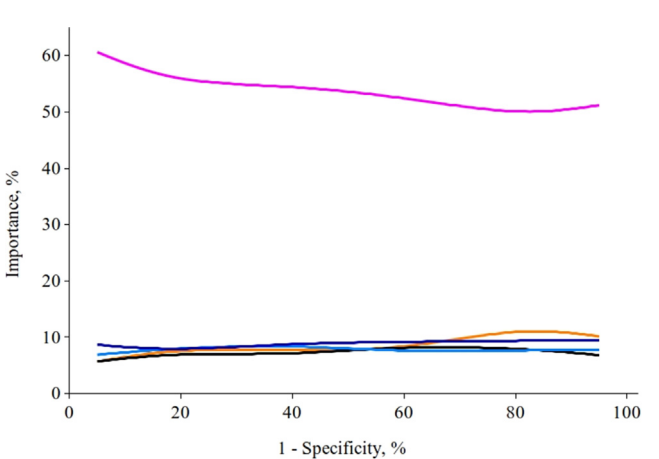


Fig. 7. Importance evolution for factors accounting for more than 5% when classifying light pigs at 160 d old (<87.8 kg) with random forest. Blue lines referred to the difference between birth BW and average litter weight, both at birth (fair line) and after cross-fostering (dark line). Remaining lines referred to birth (black), weaning (orange) and end of the nursery period BWs (pink).

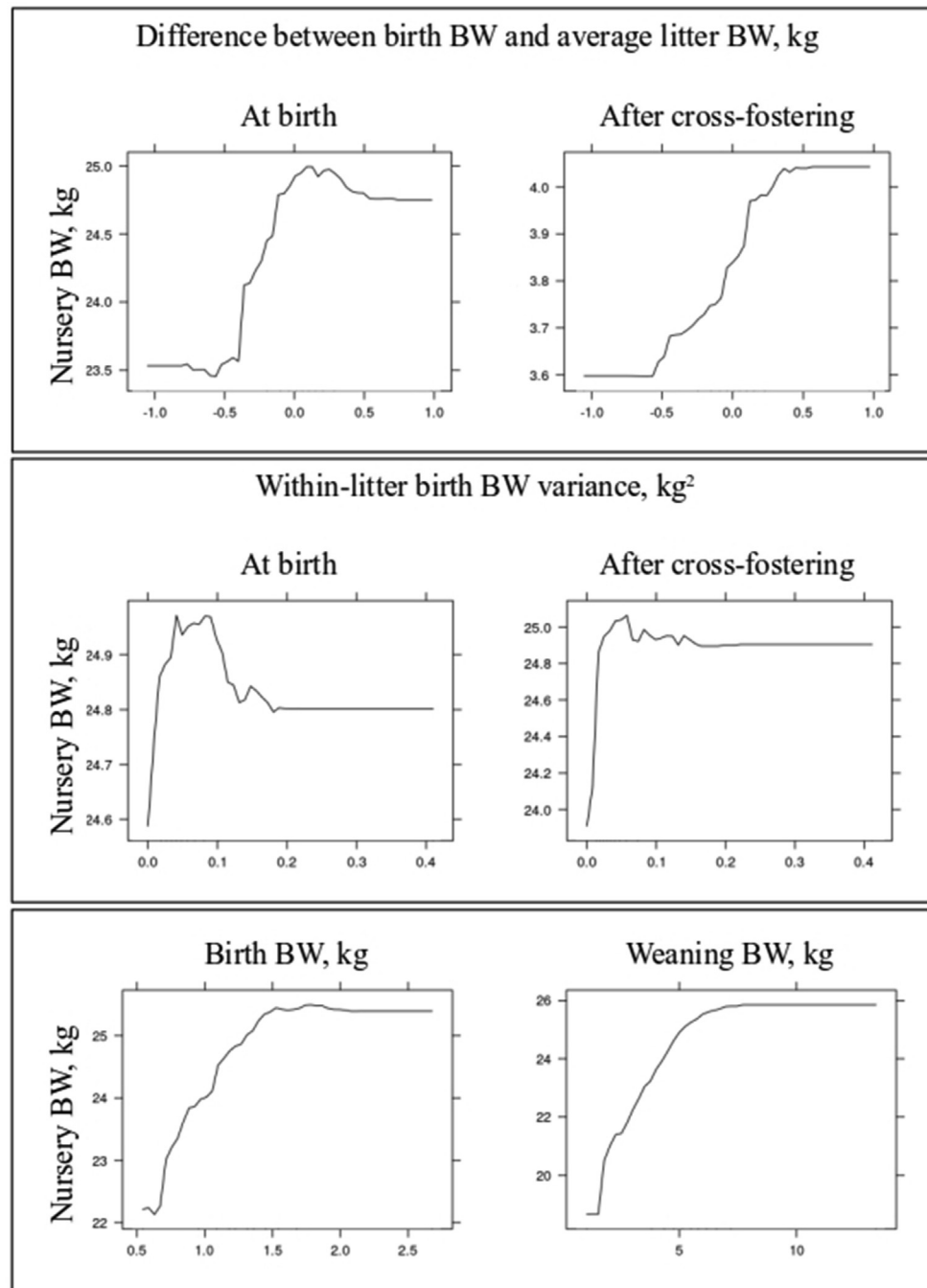


Fig. 8. Partial dependence plots for factors accounting for more than 5% importance when classifying light pigs at 75 d old (<22.3 kg) with random forest (sensitivity, 84%; specificity, 60%). Abbreviation: Nursery BW = BW at the end of the nursery period.

potentially light animals would be managed according to appropriate management practices. On the other hand, time-consuming individual treatments, or expensive supplements (Casellas et al., 2005; Threadgold et al., 2021) must focus on those pigs who were truly light, avoiding false positives (i.e., high specificity) and accepting a higher rate of false negatives (i.e., low sensitivity), even if they suffered a commercial penalty.

Area under the curve across cut-points and production stages

The AUC increased when reducing the weight cut-point for light pigs and indicated a superior classification ability when focusing on more extreme individuals. This could be anticipated because

of the increase in the average difference between both groups of animals, as previously reported by Steyerberg et al. (2010). Nevertheless, differences were remarkably small for classification methods with higher AUC, which suggested negligible benefit for more restrictive weight cut-off points. This stability of the AUC for methods with higher estimates is of special interest for the implementation of these methodologies in the swine industry, providing a wide range of cut-point values without relevant departures from the expected power of discrimination.

The increase of AUC along production stages is directly related to the inclusion of additional sources of information in the analytical model. It must be noted that WW contributed strongly to the prediction of BW at the end of the nursery period and slaughter

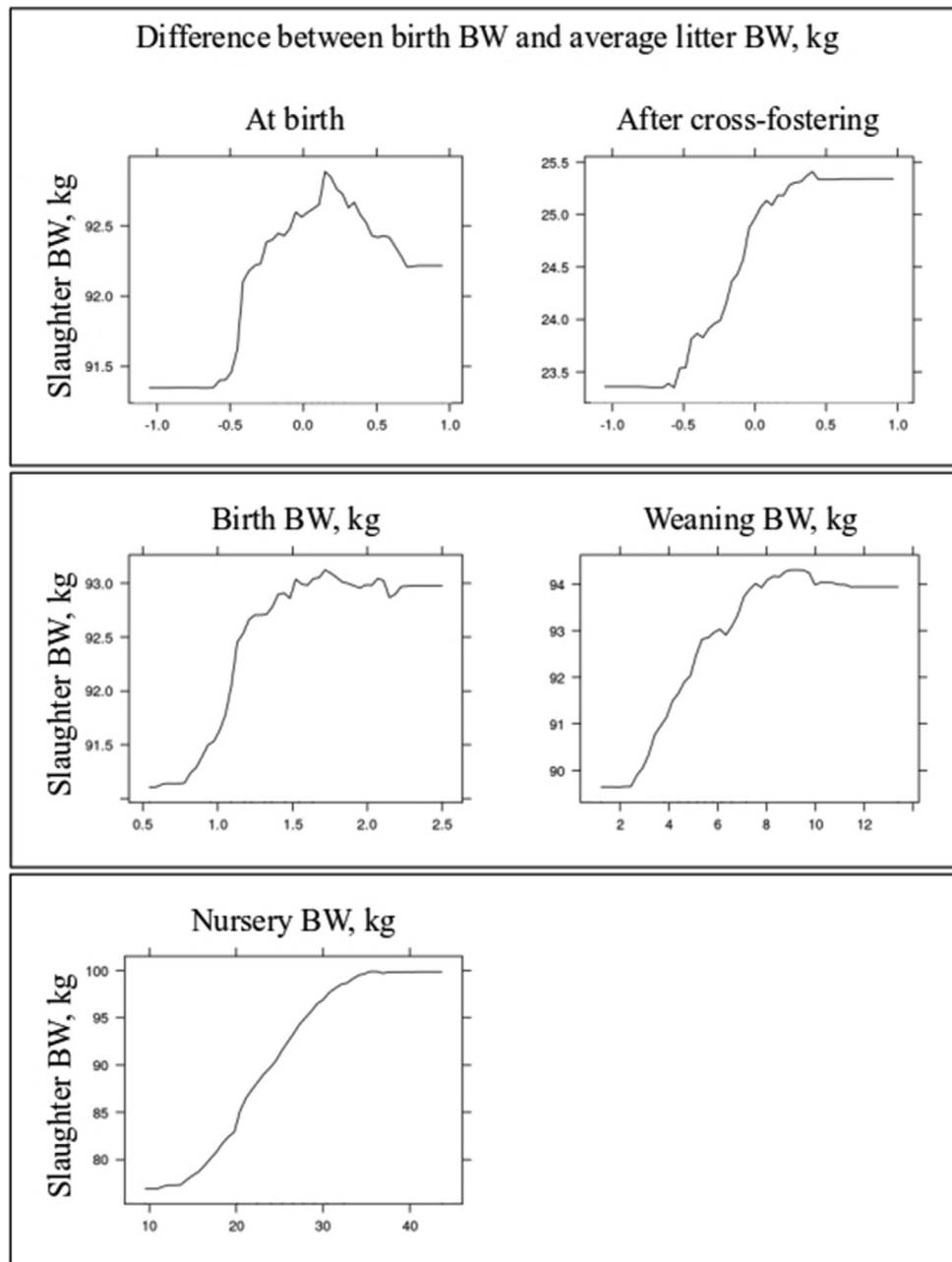


Fig. 9. Partial dependence plots for factors accounting for more than 5% importance when classifying light pigs at 160 d old (<87.8 kg) with random forest (sensitivity, 87%; specificity, 60%). Abbreviation: Nursery BW = BW at the end of the nursery period.

BW, and BW at the end of the nursery period was a strongly determinant factor for the prediction of slaughter BW. Indeed, the AUC fell below values calculated at weaning for any classification model and weight cut-point used if WW and BW at the end of the nursery period were arbitrarily removed from the model. This is directly linked to the time elapsed between birth and BW at the end of the nursery period and slaughter BW recording, where birth-related factors must be less informative than weight records collected closer in time to the event being predicted. Thus, the influence of a prediction factor increases the closer it is to the weight record. Although weighing animals at later phases of production could be a time-consuming practice, our results suggest that this source of information is the most useful for optimizing a classification tool for lightweight animals.

Pig weight classification techniques

The random forest methodology was the approach that obtained the best performance among the five models evaluated in this research. This technique systematically outperformed all other methodologies across all testing data sets, production stages and cut-points for light pigs. This demonstrated the stability of random forest for light pig classification at different production stages and revealed a useful tool for the swine industry. The superior statistical performance of random forest was previously reported in pigs, although with other data sources (Bakoev et al., 2020), and has also been seen in other livestock species such as sheep (Sant'Ana et al., 2021). This indicates the existence of methods, specifically those based on machine learning, which can out-

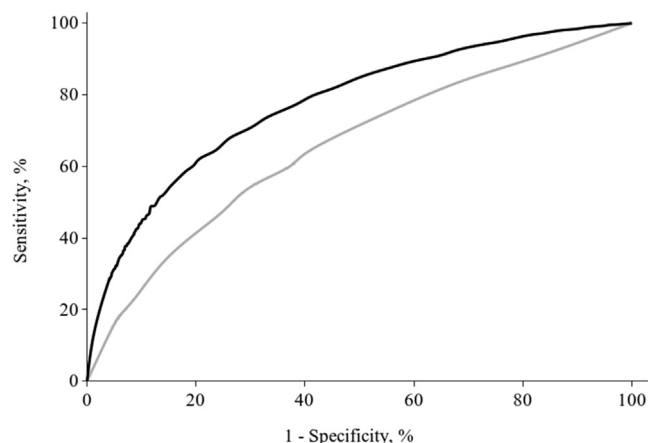


Fig. 10. Receiver operating characteristic curve for predicting the marketing BW of Yorkshire pigs based on BW at birth, weaning BW, and BW at the end of the nursery period (black line), compared to considering only the BW at birth (gray line).

perform traditional ordinary least squares regression as predictive tools for weight traits in pigs.

One must be cautious with the extrapolation of these results to other pig populations or breeds because of the slight advantage reported for random forest in relation with other machine learning algorithms such as boosting. Indeed, the objective of this research did not focus on the identification of a specific analytical method or any “golden rule” based on the optimum combination of factors, sensitivity, and specificity, but the evaluation of several approaches under different levels of stringency (*i.e.*, weight cut-points), characterizing the behavior of each classification method along the analytical space.

Factors influencing light pig classification

Under commercial conditions, the results obtained from the random forest procedure agree with those reported by López-Vergé et al. (2018) based on an epidemiological approach. López-Vergé et al. (2018) indicated that the probability for a light piglet to be sent to the slaughterhouse late increases with age, and increases closer to the end of the cycle, probably due to the short time available to implement preventive measures. The relative importance of factors that impact the classification of low BW pigs at weaning (difference between BIW and average litter weight, both at birth and after cross-fostering, accounting for >20%), decreased at the entrance of the growing-finishing barn only six weeks later. The contribution of BIW also decreases from weaning to 160 days old pigs and most of the evaluated factors (sex, parity, farm, or season) become residual as the marketing weight approaches. By contrast, the importance of the pig BW at each previous step remains the most important factor and increases with age, accounting for 30, 40, and 50% of relative importance for weaning, end of the nursery (72 d) or slaughterhouse (160 d), respectively. However, a significant reduction of the AUC in the ROC curve is observed when only BIW is considered to predict the final marketing weight. In that sense, lower sensitivity and specificity are obtained independently of the classification model used.

This indicates that the main effect on nursery and grow-finish is accumulated and best described by BW at the beginning of each period (weaning and end of the nursery period, respectively) and it follows that the possibility of low BW pigs to recover BW is reduced with age or approaching the end of the cycle. Factors not considered, for example, sex (Douglas et al., 2013; Paredes et al., 2012), genetic line (Damgaard et al., 2003), birth season

(Douglas et al., 2013; Paredes et al., 2012) or sanitary status of the farm (van der Meer et al., 2016; 2020) may influence the risk of being a slow-growing pig. Nevertheless, the influence of these factors in the analysis becomes marginal, as the inclusion of BW in the model reduces the significance of the other factors, given that it already incorporates a portion of their impact.

As in the present study, WW has been shown to have a high impact on postweaning growth and the time to reach the desirable slaughter BW (Collins et al., 2017; Wolter and Ellis, 2001). Larriestra et al. (2006), by fitting a logistic mixed model, reported a threshold of 3.6 kg of BW at weaning (17 ± 2 d), which maximizes sensitivity and specificity to correctly predict the likelihood of dying or of being lightweight when exiting the nursery stage. Moreover, He et al. (2016) stated that 80% of pigs weaned at 28 days of age at <6.4 kg of BW became slow-growing pigs at slaughter age. Similarly, López-Vergé et al. (2019) reported that 6.9 kg of BW at a minimum of 26 days of age at weaning was a threshold that predicted better growth performance in the grow-finish period. With lower lactation length (ranging between 21 and 28 d), the cut-point of 5.0 kg at weaning may account for >25% of a weaning batch, showing that piglet BW at weaning is essential. Based on our observed response, and extrapolating to the weaning BWs previously reported by He et al. (2016) and López-Vergé et al. (2019), it may be assumed that these pigs will differ by >10–15 kg of BW at marketing as compared to the pigs with the $WW \leq 3.6$ kg reported by Larriestra et al. (2006).

In contrast with the end of the nursery or grow-finisher periods, the probability of being small at weaning is less related to the BW at the beginning of the period (birth or after cross-fostering BW), indicating that other factors are involved, and that action may be taken at those early stages on these factors. This study shows the relative importance of cross-fostering for the classification of the low BW piglets at weaning, ranging from 40 to 30% with specificities of 40 and 80%. One of the main predictive factor is the within-litter weight variance after cross-fostering, showing the importance of performing a proper adoption management in order to reduce WW variability.

Conclusion

Classification of light pigs at different production stages can be effectively carried out by both ordinary least squares and machine learning algorithms, although with a balance between realized sensitivity and specificity. The random forest machine learning algorithm reached the highest statistical performance at weaning, end of the nursery period, and slaughter, with weight-related traits at previous ages playing the key role in classifying light animals. In terms of farm decision making, the machine learning models could become a key tool for improving efficiency. We have demonstrated that the probability to stop being a low BW pig decreases with age and becomes progressively worse as the pig approaches the final BW. The BW and variation of the newborn and weaned piglets, as well as on-farm handling techniques (*i.e.*, cross-fostering), are key factors to reduce weight gain variability and to achieve maximum return.

Ethics approval

Not applicable.

Data and model availability statement

None of the data were deposited in an official repository. The data that support the study findings are available from the authors upon request.

Declaration of Generative AI and AI-assisted technologies in the writing process.

The authors did not use any artificial intelligence-assisted technologies in the writing process.

Author ORCIDs

Joaquim Casellas: <https://orcid.org/0000-0002-4982-3556>.
Pau Salgado-López: <https://orcid.org/0000-0002-1791-009X>.
Javier Lorente: <https://orcid.org/0009-0002-3590-5005>.
Iara Solar Diaz: <https://orcid.org/0000-0002-7285-0384>.
Tom Rathje: <https://orcid.org/0009-0008-1413-2721>.
Josep Gasa: <https://orcid.org/0000-0002-5828-7142>.
David Solà-Oriol: <https://orcid.org/0000-0001-8365-340X>.

Declaration of interest

The authors declare no conflicts of interest.

CRediT authorship contribution statement

J. Casellas: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **P. Salgado-López:** Writing – review & editing. **J. Lorente:** Resources. **I. Solar Diaz:** Writing – review & editing, Resources. **T. Rathje:** Writing – review & editing, Resources. **J. Gasa:** Writing – review & editing, Conceptualization. **D. Solà-Oriol:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Acknowledgements

The authors gratefully acknowledge Andrimner and DNA genetics for providing their collaboration in the present study.

Financial support statement

This work was funded by the Ministerio de Ciencia, Innovación y Universidades, Gobierno de España (PID2019-103915GB-I00) and an FI grant (2022 FI_B00080) from the Departament de Recerca i Universitats, Generalitat de Catalunya to Pau Salgado-López. Moreover, the support of the UAB-Banco de Santander Talent Program is also thanked for the provision of research scholarship for David Solà-Oriol.

References

- Alexopoulos, J.G., Lines, D.S., Hallett, S., Plush, K.J., 2018. A review of success factors for piglet fostering in lactation. *Animals* 8, 38. <https://doi.org/10.3390/ani8030038>.
- Bakoev, S., Getmantseva, L., Kolosova, M., Kostyunina, O., Chartier, D.R., Tatarinova, T.V., 2020. PigLeg: prediction of swine phenotype using machine learning. *PeerJ* 8, e8764. <https://doi.org/10.7717/peerj.8764>.
- Blavi, L., Solà-Oriol, D., Llonch, P., López-Vergé, S., Martín-Orúe, S.M., Pérez, J.F., 2021. Management and feeding strategies in early life to increase piglet performance and welfare around weaning: a review. *Animals* 11, 302. <https://doi.org/10.3390/ani11020302>.
- Botermans, A.M., Georgsson, L., Weström, B.R., Olson, A., Svendsen, J., 2000. Effect of feeding environment on performance, injuries, plasma cortisol and behavior in growing-finishing pigs: studies on individual pigs housed in groups. *Acta Agriculturae Scandinavica Section A* 50, 250–262. <https://doi.org/10.1080/090647000750069449>.
- Brossard, L., Dourmad, J.-Y., Rivest, J., Van Milgen, J., 2009. Modelling the variation in performance of a population of growing pig as affected by lysine supply and feeding strategy. *Animal* 3, 1114–1123. <https://doi.org/10.1017/S1751731109004546>.
- Calderón-Díaz, J.A., García-Manzanilla, E., Diana, A., Boyle, L.A., 2018. Cross-fostering implications for pig mortality, welfare and performance. *Frontiers in Veterinary Science* 5, 123. <https://doi.org/10.3389/fvets.2018.00123>.
- Carcó, G., Gallo, L., Bona, M.D., Latorre, M.A., Fondevila, M., Schiavon, S., 2018. The influence of feeding behaviour on growth performance, carcass and meat characteristics of growing pigs. *PLoS One* 13, e0205572.
- Casellas, J., Casas, X., Piedrafitra, J., Manteca, X., 2005. Effect of medium- and long-chain triglyceride supplementation on small newborn-pig survival. *Preventive Veterinary Medicine* 67, 213–221. <https://doi.org/10.1016/j.pvetmed.2004.10.007>.
- Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A., Aggarwal, C., Shen, D., Rastogi, R. (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, Vol. 42. ACM, New York, NY, USA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Collins, C.L., Pluske, J.R., Morrison, R.S., McDonald, T.N., Smits, R.J., Henman, D.J., Stensland, I., Dunshea, F.R., 2017. Post-weaning and whole-of-life performance of pigs is determined by live weight at weaning and the complexity of the diet fed after weaning. *Animal Nutrition* 3, 372–379. <https://doi.org/10.1016/j.aninu.2017.01.001>.
- Damgaard, L.H., Rydhmer, L., Løvendahl, P., Grandinson, K., 2003. Genetic parameters for within-litter variation in piglet birth weight and change in within-litter variation during suckling. *Journal of Animal Science* 81, 604–610. <https://doi.org/10.2527/2003.813604x>.
- Douglas, S.L., Edwards, S.A., Sutcliffe, E., Knap, P.W., Kyriazakis, I., 2013. Identification of risk factors associated with poor lifetime growth performance in pigs. *Journal of Animal Science* 91, 4123–4132. <https://doi.org/10.2527/jas.2012-5915>.
- Douglas, S.L., Edwards, S.A., Kyriazakis, I., 2014. Management strategies to improve the performance of low birth weight pigs to weaning and their long-term consequences. *Journal of Animal Science* 92, 2280–2288. <https://doi.org/10.2527/jas.2013-7388>.
- Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77, 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>.
- Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 29, 1189–1232.
- Georgsson, L., Svendsen, J., 2002. Degree of competition at feeding differentially affects behavior and performance of group-housed growing-finishing pigs of different relative weights. *Journal of Animal Science* 80, 376–383. <https://doi.org/10.2527/2002.802376x>.
- He, Y., Deen, J., Shurson, G.C., Wang, L., Chen, C., Keisler, D.H., Li, Y.Z., 2016. Identifying factors contributing to slow growth in pigs. *Journal of Animal Science* 94, 2103–2116. <https://doi.org/10.2527/jas.2015-0005>.
- Hosmer, D.W., Lemeshow, S., 2000. *Applied Logistic Regression*, 2nd edition. John Wiley and Sons, New York, NY, USA. <https://doi.org/10.1002/0471722146>.
- Ibañez-Escriche, N., Varona, L., Sorensen, D., Noguera, J.L., 2008. A study of heterogeneity of environmental variance for slaughter weight in pigs. *Animal* 2, 19–26. <https://doi.org/10.1017/S1751731107001000>.
- Larriestra, A.J., Wattanaphansak, S., Neumann, E.J., Bradford, J., Morrison, R.B., Deen, J., 2006. Pig characteristics associated with mortality and light exit weight for the nursery phase. *Canadian Veterinary Journal* 47, 560–566.
- López-Vergé, S., Gasa, J., Farré, M., Coma, J., Bonet, J., Solà-Oriol, D., 2018. Potential risk factors related to pig body weight variability from birth to slaughter in commercial conditions. *Translational Animal Science* 2, 383–395. <https://doi.org/10.1093/tas/txy082>.
- López-Vergé, S., Gasa, J., Coma, J., Bonet, J., Solà-Oriol, D., 2019. Effect of lactation length caused by the management production system on piglet performance until slaughter. *Livestock Science* 224, 26–30. <https://doi.org/10.1016/j.livsci.2019.04.003>.
- Maes, D.G.D., Duchateau, L., Larriestra, A., Deen, J., Morrison, R.B., De Kruif, A., 2004. Risk Factors for Mortality in Grow-finishing Pigs in Belgium. *Journal of Veterinary Medicine B* 51, 321–326.
- Marvin, N., Ziegler, A., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77, 1–17. <https://doi.org/10.18637/jss.v077.i01>.
- Mitchell, T., 1997. *Machine Learning*. McGraw Hill, New York, NY, USA.
- Muns, R., Silva, C., Manteca, X., Gasa, J., 2014. Effect of cross-fostering and oral supplementation with colostrums on performance of newborn piglets. *Journal of Animal Science* 92, 1193–1199. <https://doi.org/10.2527/jas.2013-6858>.
- Muns, R., Tummaruk, P., 2016. Management strategies in farrowing house to improve piglet pre-weaning survival and growth. *Thai Journal of Veterinary Medicine* 46, 347–354.
- Paredes, S.P., Jansman, A.J., Verstegen, M.W., Awati, A., Buist, W., Den Hartog, L.A., Van Hees, H.M., Quiniou, N., Hendriks, W.H., Gerrits, W.J., 2012. Analysis of factors to predict piglet body weight at the end of the nursery phase. *Journal of Animal Science* 90, 3243–3251. <https://doi.org/10.2527/jas.2011-4574>.
- Patience, J.F., Engele, K., Beaulieu, A.D., Gonyou, H.W., Zijlstra, R.T., 2004. Variation: Costs and consequences. *Advanced Pork Production* 15, 257–266.
- Peltoniemi, O., Yun, J., Björkman, S., Han, T., 2021. Coping with large litters: the management of neonatal piglets and sow reproduction. *Journal of Animal Science and Technology* 63, 1–15. <https://doi.org/10.5187/jast.2021.e3>.
- Peterson, W., Birdsall, T., Fox, W., 1954. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* 4, 171–212.
- Quiniou, N., Dagorn, J., Gaudre, D., 2002. Variation of piglets' birth weight and consequences on subsequent performance. *Livestock Production Science* 78, 63–70. [https://doi.org/10.1016/S0301-6226\(02\)00181-1](https://doi.org/10.1016/S0301-6226(02)00181-1).

- Quinlan, J.R., 1987. Simplifying decision trees. *International Journal of Man-Machine Studies* 27, 221–234. [https://doi.org/10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6).
- R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Sant'Ana, D.A., Pache, M.C.B., Martins, J., Soares, W.P., de Melo, S.L.N., Garcia, V., Webber, V.A.M., Heimbach, N.S., Mateus, R.G., Pistori, H., 2021. Weighting live sheep using computer vision techniques and regression machine learning. *Machine Learning with Applications* 5, <https://doi.org/10.1016/j.mlwa.2021.100076> 100076.
- Solà-Oriol, D., Gasa, J., 2017. Feeding strategies in pig production: Sows and their piglets. *Animal Feed Science and Technology* 233, 34–52. <https://doi.org/10.1016/j.anifeedsci.2016.07.018>.
- Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 21, 128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>.
- Threadgold, T., Greenwood, E.C., Van Wettere, W., 2021. Identifying suitable supplements to improve piglet survival during farrowing and lactation. *Animals* 11, 2912. <https://doi.org/10.3390/ani11102912>.
- Van der Meer, Y., Lammers, A., Jansman, A.J.M., Rijnen, M.M.J.A., Hendriks, W.H., Gerrits, W.J.J., 2016. Performance of pigs kept under different sanitary conditions affected by protein intake and amino acid supplementation. *Journal of Animal Science* 94, 4704–4719. <https://doi.org/10.2527/jas.2016-0787>.
- Van Der Meer, Y., Jansman, A.J.M., Gerrits, W.J.J., 2020. Low sanitary conditions increase energy expenditure for maintenance and decrease incremental protein efficiency in growing pigs. *Animal* 14, 1811–1920. <https://doi.org/10.1017/S1751731120000403>.
- Wolter, B.F., Ellis, M., 2001. The effects of weaning weight and rate of growth immediately after weaning on subsequent pig growth performance and carcass characteristics. *Canadian Journal of Animal Science* 81, 363–369. <https://doi.org/10.4141/A00-100>.
- Woodward, P.M., 1953. *Probability and Information Theory with Applications to Radar*. Pergamon Press, London, United Kingdom.