PigLeg: prediction of swine phenotype using machine learning

Siroj Bakoev[1], Lyubov Getmantseva[1], Maria Kolosova[2], Olga Kostyunina[1], Duane Chartier[3], Tatiana V. Tatarinova[4-7]

1 L. K. Ernst Federal Science Center for Animal Husbandry, Moscow, Russia

2 Don State Agrarian University, Persianovsky, Rostov region, Russia

3 ICAI, Inc. Culver City, CA, 90230, USA

4 Department of Biology, University of La Verne, La Verne, CA, 91750, USA

5 The Institute for Information Transmission Problems, Moscow 127051, Russia

6 Vavilov Institute of General Genetics, Moscow 119333, Russia

7 School of Fundamental Biology and Biotechnology, Siberian Federal University, Krasnoyarsk 660041, Russia

Abstract:

Industrial pig farming is associated with negative technological pressure on the bodies of pigs. Leg weakness and lameness are the sources of significant economic loss in raising pigs. Therefore, it is important to identify predictors of limb condition. This work presents assessments of the state of limbs using indicators of growth and meat characteristics of pigs based on machine learning algorithms. We have evaluated and compared the accuracy of prediction for several ML classification algorithms (Random Forest, K-Nearest Neighbors, Artificial Neural Networks, C50Tree, Support Vector Machines, Naive Bayes, Generalized Linear Models, Boost, and Linear Discriminant Analysis) and have identified the Random Forest and K-Nearest Neighbors as the best performing algorithms for predicting pig leg weakness using a small set of simple measurements that can be taken at an early stage of animal development. Muscle Thickness, Back Fat amount, and Average Daily Gain serve as significant predictors of conformation of pig limbs. Our work demonstrates the utility and relative ease of using machine learning algorithms to assess the state of limbs in pigs based on growth rate and meat characteristics.

Keywords:

Artificial Intelligence, Bioinformatics, Computational Biology, Data Mining & Machine Learning, Evolutionary Studies, Mathematical Biology, Animal Behavior.

Introduction:

One of the main research tasks in animal husbandry is discovery of the biological mechanisms influencing animal productivity and finding efficient ways of increasing it. Pork is the most widely consumed meat in the world. In addition to meat, many valuable products come from pigs: insulin, replacement human heart valves, suede for shoes and clothing, and gelatin for food and industry.

Intensive production of pig products is associated with negative technological pressure on the development of pigs. Breeding for accelerated development and meatiness leads to a rearrangement of the metabolism in the animal's body, resulting in morphological and functional rearrangements of the internal organs, muscle, adipose and bone tissues. Changes associated with the cartilage structure are called osteochondrosis (leg weakness). Under industrial conditions, the term "leg weakness" is used to describe the poor constitution of pig legs or the clinical condition associated with lameness or stiffness of movements. Weakness results from abnormal changes in the cartilage joints and the development of epiphyseal plates, which are responsible for bone enlargement both in length and diameter [1]. Weak epiphyseal plates can break, and the cartilage that covers the joint surface cracks. In the acute phase of the disease, bone fractures may occur near the epiphyseal plate. However, in most cases, the disease takes a chronic form, develops gradually, and manifests itself as incorrect shape and alignment of legs, as well as stiffness of the animal's gait. In this regard, the first step in diagnosing the disease is an exterior assessment of the legs and gait. Typically, pig legs are evaluated visually by specially trained personnel using a point system [2].

Rapid advances in next generation sequencing (NGS) and high-density genotyping technologies allows identification of several quantitative trait loci (QTL) for pig lameness and leg weakness. Leg weakness is partially a heritable trait, with heritability estimates of leg ranging from low (0.07, [3]) to moderate (0.36, [4]). In spite of the agricultural importance of this trait, there has only been a limited number of GWAS for leg weakness. In addition, the trait may be complex and influenced by many factors, such as bone strength, muscle growth, fat accumulation, and body weight gain. Therefore, the task of the present work was to identify these factors using modern statistical approaches.

Rapidly developing data mining approaches are of increasing interest because they provide for acquisition and analysis of information that results in predictive productivity indicators for animals [5–7]. Machine learning (ML) approaches have been successfully used in animal husbandry for early prediction of the growth and quality of adult wool in Australian merino sheep [8], sheep carcass traits from early-life records [9], and skin temperature of piglets [10]. Compared to other statistical approaches, ML is suitable for use even when there are many predictors, missing values, and abnormally distributed data, which is often the case with data obtained from commercial pig production.

In this work we have evaluated the condition of pig legs by application of ML methods to growth and meat characteristics. We have compared a number of ML classification algorithms for predicting the state of the front and hind legs. This led to identification of the most effective algorithm for predicting leg weakness using a small set of cost-effective and easily measurable sets of functions that can be used in the early period of animal rearing.

Results

ML models were able to predict the state of the fore and hind legs. RF surpassed all other learning algorithms in all respects and scenarios. In some cases, RF did not have significant superiority over KNN (Table 2). Accordingly, KNN was the second most efficient algorithm among all the characteristics and scenarios.

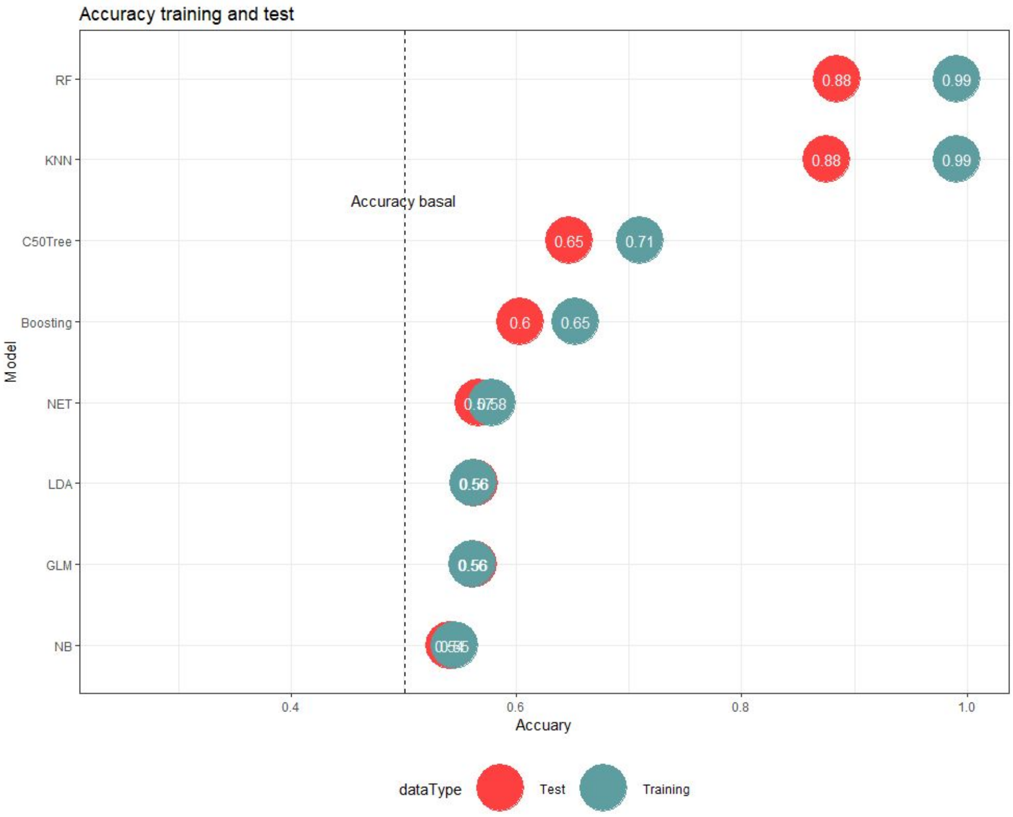
The superiority of RF and KNN and, conversely, the suboptimality of SVM and NB are consistently associated with a lower and higher dispersion of forecasting indicators, respectively. SVM and NB were among the least effective forecasting methods in this study, providing the lowest correlation and the largest forecasting errors.

Table 2. Comparison between the models using the testing dataset.

Model	Accuracy	Kappa	Sensitivity	Specificity
RF	0.8846	0.7693	0.8232	0.9463
KNN	0.8754	0.7509	0.8013	0.9499
C50Tree	0.6469	0.294	0.5746	0.7195
Boost	0.6035	0.207	0.5995	0.6075
NNET	0.5667	0.1335	0.5619	0.5716
LDA	0.563	0.1258	0.5986	0.5272
GLM	0.5624	0.1246	0.5971	0.5275
SVM	0.5603	0.1202	0.653	0.4671
NB	0.5411	0.0816	0.6984	0.3832

A graphical interpretation of the comparative analysis of the predicted value of all models shows that all models in the training set receive more accurate forecasts than in the test set. At the same time, the RF and KNN models provide high accuracy of prediction relative to other models.

To determine the models that achieve the best results in solving the problem after the training procedures and their optimization, a comparative analysis was carried out. Obviously, the indicators obtained by validation are estimates of the ability of the model to predict new observations and these estimates have deviations.



A comparison was made between all models with the non-parametric Friedman test and a pairwise comparison of all models, the results of which are summarized in Table 3.

Table 3. Search results among all the models of non-parametric tests of Friedman and paired comparison of all models.

Model A	Model B	p-value	Model A	Model B	p-value
boosting	arbol	4.37E-02	NB	logistic	2.17E-08
KNN	arbol	2.17E-08	NET	arbol	2.17E-08
KNN	boosting	2.17E-08	NET	boosting	2.17E-08
LDA	arbol	2.17E-08	NET	KNN	2.17E-08

LDA	boosting	2.17E-08	NET	LDA	1.31E-07
LDA	KNN	2.17E-08	NET	logistic	1.31E-07
logistic	arbol	2.17E-08	NET	NB	2.17E-08
logistic	boosting	2.17E-08	rf	arbol	2.17E-08
logistic	KNN	2.17E-08	rf	boosting	2.17E-08
logistic	LDA	9.30E-02	rf	KNN	2.17E-08
NB	arbol	2.17E-08	rf	LDA	2.17E-08
NB	boosting	2.17E-08	rf	logistic	2.17E-08
NB	KNN	2.17E-08	rf	NB	2.17E-08
NB	LDA	2.17E-08	rf	NET	2.17E-08
Friedman rank sum test					
Friedman chi-squared = 286.85, df = 6, p-value < 2.2e-16					

The best predictive capabilities in the dataset were shown by Random Forest models. In addition, it must be noted that such signs as Muscle Thickness, Back Fat, Average Daily Gain can act as predictors of leg weakness. Information on breed and gender were not significant for assessment the status of legs.

Discussion

The increase in the prevalence of leg weakness in pigs in the middle of the 20th century coincided with a surge of targeted breeding work to increase the growth rate of animals. This was mainly due to economic pressure and the need to shorten the period from birth to slaughter. Moreover, in wild boars, which require about two years to reach maturity, osteochondrosis is not observed. A hypothesis was put forward regarding a significant relationship between growth qualities and weakness of the legs. Several large population studies have shown a positive correlation between these traits [1,11,12]. [13] noted that pigs with clinical signs of leg weakness grew faster in the early stages of life than pigs without these signs, but by the time of slaughter, their growth had become slower. He suggested that the unfavorable relationship between fatness and growth rate is balanced by discomfort due to the emerging clinical signs of leg weakness, leading to reduced feed intake. [14] discovered a significant correlation between the length of the carcass and the weight of the ham with the degree of damage to the proximal and distal parts of the femur - osteochondrosis. The relationship between the state of the legs and indicators of meat productivity of pigs was confirmed by a number of studies conducted on pigs of various breeds. A study by [15] showed that Duroc pigs with low foreleg scores had greater muscle length and mass. Draper et al. examined thickness of fat, the length of the body and the yield of meat, but found no significant differences related to the condition of the legs. In another study, the emphasis was placed on studying the relationship between the legs and meat qualities of large white pigs. The results showed that pigs with leg problems were usually heavier and with more back fat compared to healthy pigs.

These observations agree with the results obtained by our machine learning approach. It is clear that that machine learning can be successfully used to evaluate the growth performance and meat characteristics of pigs.

Materials and Methods

Data sources

Data were taken from 24,584 pigs of breeds Landrace and Large White (Table 1). Several factors can affect the conformation of legs: breed, year of birth (BirthDate), sex, Average Daily Gain (ADG), Muscle thickness (MT), and Back Fat thickness (BF). Front and Back legs were visually assessed using a point system from 1 to 5 (from bad to good). Points 1 and 2 were received by animals with obvious leg defects, 3 points — average condition, 4 and 5 — good and excellent, respectively. Preliminary data analysis showed the imbalance of the available data. Imbalanced classes are a common problem in machine learning classification where there are a disproportionate ratio of observations in each class. Since most ML algorithms work best when the number of samples in each class are about equal, a balancing procedure was applied. After the preliminary analysis, the year of birth (BirthDate) and gender (Sex) were excluded as predictors of the least importance.

Table 1. Sample description. The dataset contains 21,247 females and 3337 males. 12,195 of Landrace and 12389 of Large White breeds. Predictors: Average Daily Gain, Backfat Thickness, Muscle, Thickness, Birth Date, Breed, Sex. Dependent variables: scores for front and back legs.

Variable	**min**	**1st Qu.**	**median**	**mean**	**3rd Qu.**	**max**
Average Daily Gain	0.33	0.72	0.79	0.79	0.85	1.61
Backfat Thickness	4.30	10.90	12.90	13.25	15.20	35.60
Muscle Thickness	32.12	56.04	59.70	59.68	63.40	96.00
Birth Date	2012	2014	2015	2015	2016	2016
Scores						
Front Legs	1.00	3.00	3.00	3.11	3.00	5.00
Back Legs	1.00	3.00	3.00	2.99	3.00	5.00

Methods

Classification models were constructed and analyzed using the following machine learning methods: Random Forest (RF) [16], K-Nearest Neighbors (K-NN) [17,18], artificial neural networks (Neural Networks) [19], C50Tree, Support Vector Machines (SVM) [20], Naive Bayes (NB) [21], GLM [22], Boost [23] and Linear Discriminant Analysis (LDA) [24]. All calculations and simulations were performed in R (version 3.6.1, [25]) using the caret packages [26], DMwR [27]. Leg scores was used as the response variables (Table 1).

K-Nearest Neighbors (K-NN). The K-NN classifier is based on the compactness hypothesis (Zagoruiko, 1999), which assumes that the test object d will have the same class label as the training objects in the local area of its immediate environment. When the value of K is one, the analyzed object is assigned to a certain class depending on information about its single nearest neighbor. In the k-NN variant, each object belongs to the prevailing class of nearest neighbors, where k is the algorithm parameter. Any clustering algorithm can be considered effective if the compact hypothesis is satisfied, i.e. one can find such a partition of objects into groups that the distances between objects from the same group (intra-cluster distances) will be less than a certain value ε> 0, and between objects from different groups (cross-cluster distances) more than ε. [28].

Linear Discriminant Analysis (LDA). LDA is a multidimensional analysis section that allows one to evaluate differences between two or more groups of objects by several variables at the same time. It is a generalization of Fisher's linear discriminant, a method used in machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination can be used as a linear classifier or, more often, to reduce the dimension before subsequent classification. LDA is closely related to the analysis of variance (ANOVA) procedure. The LDA implements two closely related statistical procedures:

1. Interpretation of group differences, needing to answer the question: how a well-used set of variables is able to form a dividing surface for objects of the training sample and which of these variables are the most informative.
2. Classification, i.e. prediction of the value of the grouping factor for the examined group of observations.

The support vector machines (SVM), previously called the “generalized portrait” algorithm, was developed by Soviet mathematicians Vapnik and Chervonenkis [29] and has since gained widespread popularity. The main idea of the classifier on support vectors is to build a separating surface using only a small subset of points lying in the zone critical for separation, while the rest of the correctly classified observations of the training sample outside of this zone are ignored (more precisely, they are a “reservoir” for an optimization algorithm). If there are two classes of observations and a linear form of the boundary between the classes is assumed, then two cases are possible. The first of them is connected with the possibility of perfect data separation with the help of some hyperplane. Since there can be many such hyperplanes, the dividing surface is optimal, which is as far as possible from the training points, i.e. having a maximum gap M (margin).

Naive Bayes classifier (NB). Naive Bayes classifiers are a family of simple probabilistic ML classifiers based on the application of Bayes theorem. Making the “naive” assumption that all the signs describing the classified objects are completely equal and are not related to each other, then the probability of an object to belong to a given class given its observed features, P(class|features), is calculated using the Bayes formula from known distributions P(features|class). The NB assigns the objects refer to the class that has the greatest probability.

Neural Networks. Neural network models that were born in the process of developing the concept of artificial intelligence have two completely transparent analogies - the biological neural system of the brain and the computer network. Their main paradigm is that the solution in the network is formed by many simple neuron-like elements that form a graph with weighted synaptic (informational) connections that work together and purposefully to obtain a common result.

To train artificial neural networks in the R environment, the *nnet* package [30] was used; it provides flexible functionality for constructing classification models based on a multilayer perceptron.

GLM Logistic regression is commonly used as a binary classifier for alternate response samples. However, this method can also be generalized to the case with several classes. Nominal or ordinal variables can be used as the simulated response Y, and in both cases a multidimensional binomial distribution is assumed. Simply put, linear regression should be used to predict a quantitative (i.e., numerical) response variable, and logical regression should be used to predict a qualitative (i.e., categorical) response variable. Both linear regression and logistic regression are types of generalized linear models (GLM).

Gradient Boosting. One of the methods for improving predictions is boosting; an iterative process of sequentially constructing private models. Each new model is trained using information on errors made at the previous stage, and the resulting function is a linear combination of all, taking into account the minimization of any penalty function. Like bagging, boosting is a general approach that can be applied to many statistical classification methods. The idea of increasing the gradient arose as a result of Leo Braiman's observations that increasing the gradient can be interpreted as an optimization algorithm on an appropriate cost function. Several algorithms for increasing the gradient of direct regression were developed [31][32]. [32] approach optimizes the cost function with respect to the functional space by iteratively choosing a function, indicating the direction of the negative gradient.

The C 50Tree method is based on the application of a strategy of dividing data into smaller and smaller parts to identify patterns that can ultimately be used for forecasting. The model itself includes a large number of logical decisions, with decision nodes. They are divided into branches that indicate the choice of solution. The tree ends with leaf nodes (also called terminal nodes), which indicates the result of a combination of decisions. The data to be classified begins at the root node, where the ripple is transmitted to them, and various decisions in the tree, in accordance with the values of the predictors, depending on their influence on the response variable.

Random Forest is a controlled learning method in which the target class is *a priori* known, and a model is built (classification or regression) to predict future responses. Several hundred decision trees are built for training bootstrap samples. However, at each iteration of the tree construction, randomly selected *m* is from *p* predictors to be considered, and the partition can be performed on only one of these *m* variables. The meaning of this procedure, which turned out to be very effective for improving the quality of the obtained solutions, is that with the probability *(p - m) / p* some potentially dominant predictor which seeks to enter every tree is blocked. By blocking dominants, other predictors will get their chance, and tree variation will increase.

Data Preprocessing

The number of observations for training models allows one to achieve high predictive effectiveness. The data includes both continuous and high-quality variables, which allows facile problem solving. The response variable (target variable) was a leg score, which varies from 1 to 5. For practical reasons, the values were adjusted and divided into two bins: scores [1:2] - animals with "bad" legs (Q1) and scores [3:5] - animals with “good legs” (Q2). As a metric, Accuracy was calculated as the proportion of correct answers of the algorithm, precision, recall, and the F1 score (harmonic mean of precision and recall). The data points were assigned to the bins (2708 (4930) for Q1; 21876 (19654) for Q2), corresponding to 11% (20%) and 89% (80%) of measurements for the two breeds. The imbalance of the data classes (a large difference between the numbers of samples in different bins) can negatively affect learning and prediction phases of the approach. If the unbalance ratio is high, the decision function favours the “majority” class, where the largest number of samples is located. Therefore, the leg scores data for the bins were balanced using the ROSE package. Although most machine learning (ML) classification algorithms are generalized to several classes, their interpretation is simpler if there are only two.

Data Analysis

Before choosing the most important predictors and training the prognostic model, a descriptive study of variables was conducted. This process allows for a better understanding of what information each variable contains, as well as to identify possible errors. Since, in practice, it is not always possible to obtain data in its entirety, missing values were found in our data. Using the *preProcess* function from the caret package in R (short for Classification And REgression Training, http://topepo.github.io/caret/index.html), the problem of missing data values was addressed.

Studying the distribution of the response variable relative to quantitative (Muscle Thickness, Back Fat, Average Daily Gain) and qualitative (Breed) variables is an equally important procedure. Analysis of quantitative variables showed a pronounced asymmetric distribution of some predictors (Back Fat). The calculation of correlations between continuous predictors indicates that they do not contain redundant information (Fig. 1).

In order for the predictive model to be useful, it must have a success rate higher than expected by chance or at a certain base level. In classification problems, the base level is the level obtained if all observations are assigned to the majority class (mod). In our case, given that 89% (80%) of the animals have healthy legs, if Q2 = Yes, it is always predicted, then the success rate will be about 89% (80%) for unbalanced and 50% for balanced data participating in the training set, respectively. This is the minimum percentage that must be overcome with the help of predictive models (strictly speaking, this percentage should be counted only with the help of a training set).

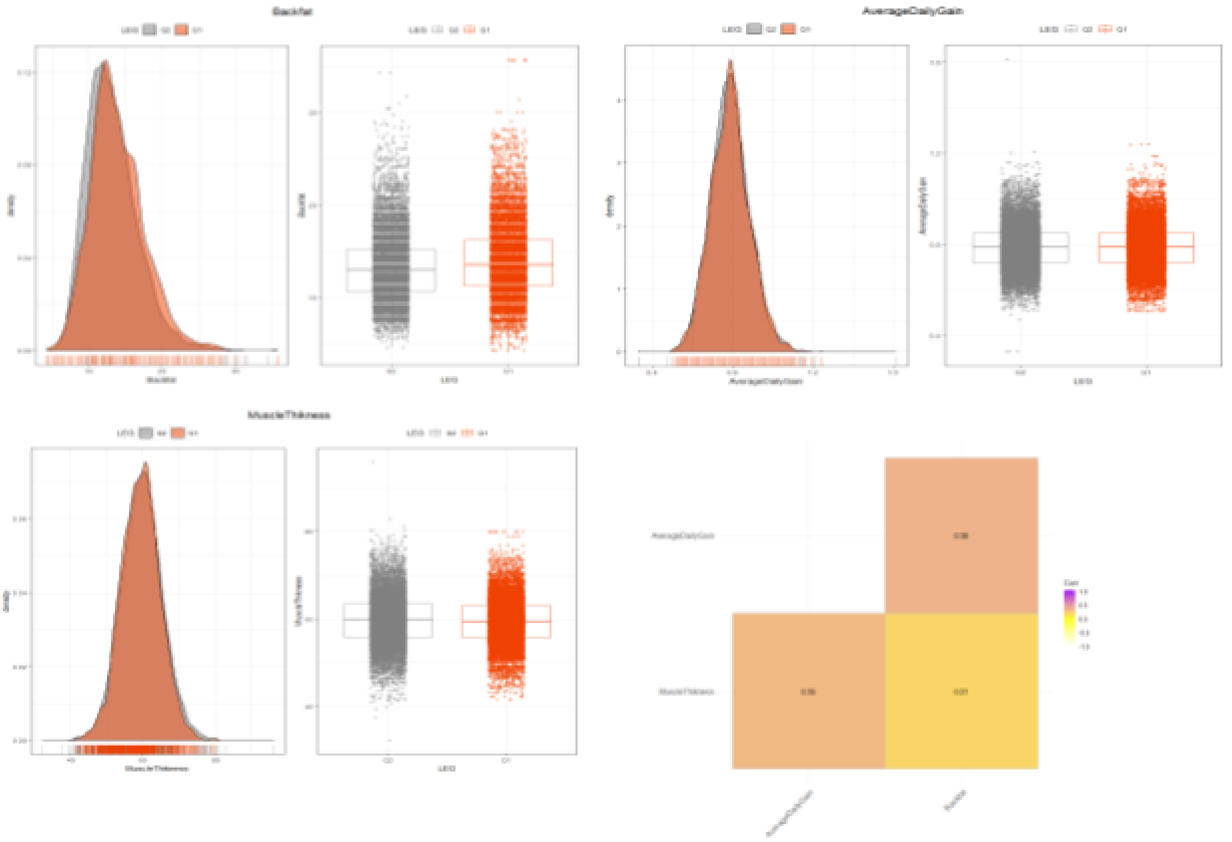


Figure 1. Correlation between continuous predictors

Since the aim of the study is to assess the state (conformation) of legs by means of selected predictors (growth and meat quality), each variable is analyzed with respect to the variable Q2 = “good”. By analyzing the data in this way, one can begin to extract ideas about which variables are most associated with “good” legs. Alternatively, to study the importance of predictors, we use the Random Forest package. Different algorithms identified that that the most important predictors are Muscle Thickness, Back Fat, Average Daily Gain, while the predictor Breed is not significant (Fig. 2).

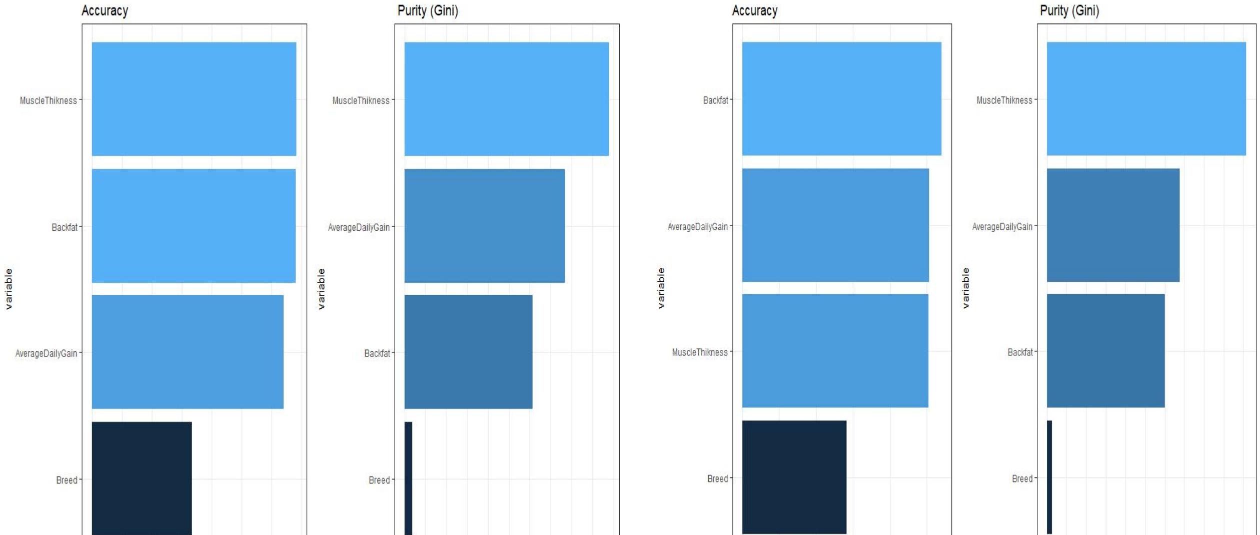


Figure 2. Determining the relative importance of predictors

Model training

Normalization (standardization) of the data was carried out on the basis that all the predictors were approximately on the same scale by using the formula:

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

By dividing each predictor by its standard deviation after centering, the data obeys the normal distribution law.

Machine learning algorithms were trained and tested based on the following structure for all three features of interest in this study. A random 10% of the data was excluded from the complete data set for the final assessment, which in our study, will be called a set of independent trials. The remaining 90% were randomly divided into 70% for training and 30% for testing 100 times. In every 100 training iterations, hyperparameters were set up using a search within the 10-fold cross-validation structure on a random 70% of the training set. The most effective hyperparameters at this stage were used to train each ML model on a training set and were tested on a test set in each iteration. All processes were implemented in R. The performance of the final model, has been evaluated on the test set.

Conclusions

Leg weakness is a source of significant economic loss in pig production, therefore, the search for predictors of leg condition is of great interest and potential value. Our comparison of various machine learning algorithms proved that growth rate and meat parameters were effective predictors of the condition of pig legs. This provides a powerful tool to assess the health of the animals. The best predictive performance was achieved by the Random Forest approach.

Funding

SB, LG, MK, and OK were supported by the Russian Found[1]ation for Basic Research grant 19-016-00068 A.

References

1. Ekman, S.; Carlson, C.S. The pathophysiology of osteochondrosis. *Vet. Clin. North Am. Small Anim. Pract.* **1998**, *28*, 17–32.
2. Le, T.H.; Christensen, O.F.; Nielsen, B.; Sahana, G. Genome-wide association study for conformation traits in three Danish pig breeds. *Genet. Sel. Evol.* **2017**, *49*, 12.
3. Aasmundstad, T.; Olsen, D.; Sehested, E.; Vangen, O. The genetic relationships between conformation assessment of gilts and sow production and longevity. *Livest. Sci.* **2014**, *167*, 33–40.
4. Knauer, M.T.; Cassady, J.P.; Newcom, D.W.; See, M.T. Phenotypic and genetic correlations between gilt estrus, puberty, growth, composition, and structural conformation traits with first-litter reproductive measures. *J. Anim. Sci.* **2011**, *89*, 935–942.
5. Morota, G.; Ventura, R.V.; Silva, F.F.; Koyama, M.; Fernando, S.C. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. *Journal of Animal Science* 2018, *96*, 1540–1550.
6. Putz, A.M.; Harding, J.C.S.; Dyck, M.K.; Fortin, F.; Plastow, G.S.; Dekkers, J.C.M.; PigGen Canada Novel Resilience Phenotypes Using Feed Intake Data From a Natural Disease Challenge Model in Wean-to-Finish Pigs. *Front. Genet.* **2018**, *9*, 660.
7. Howard, J.T. The use of "Big Data"in a modern swine breeding program now and in the future.
8. Shahinfar, S.; Kahn, L. Machine learning approaches for early prediction of adult wool growth and quality in Australian Merino sheep. *Comput. Electron. Agric.* **2018**, *148*, 72–81.
9. Shahinfar, S.; Kelman, K.; Kahn, L. Prediction of sheep carcass traits from early-life records using machine learning. *Comput. Electron. Agric.* **2019**, *156*, 159–177.
10. Gorczyca, M.T.; Milan, H.F.M.; Maia, A.S.C.; Gebremedhin, K.G. Machine learning algorithms to predict core, skin, and hair-coat temperatures of piglets. *Comput. Electron. Agric.* **2018**, *151*, 286–294.
11. Jørgensen, B.; Andersen, S. Genetic parameters for osteochondrosis in Danish Landrace and Yorkshire boars and correlations with leg weakness and production traits. *Anim. Sci.* **2000**, *71*, 427–434.
12. Nakano, T.; Brennan, J.J.; Aherne, F.X. LEG WEAKNESS AND OSTEOCHONDROSIS IN SWINE: A REVIEW. *Can. J. Anim. Sci.* **1987**, *67*, 883–901.
13. Lundeheim, N. Genetic Analysis of Osteochondrosis and Leg Weakness in the Swedish Pig Progeny Testing Scheme. *Acta Agriculturae Scandinavica* 1987, *37*, 159–173.
14. Van der Wal, P.G.; Van der Valk, P.C.; Goedegebuure, S.A.; Van Essen, G. Osteochondrosis in six breeds of slaughter pigs: II. Data concerning carcass characteristics

[1]

in relation to osteochondrosis. *Vet. Q.* **1980**, *2*, 42–47.

15. Draper, D.D.; Rothschild, M.F.; Christian, L.L. Effects of divergent selection for leg weakness on muscle and bone characteristics in Duroc swine. *Genet. Sel. Evol.* **1992**, *24*, 363.
16. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
17. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 1967, *13*, 21–27.
18. Lantz, B. *Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems*; Packt Publishing Ltd, 2015; ISBN 9781784394523.
19. Ripley, B.D.; Hjort, N.L. *Pattern Recognition and Neural Networks*; Cambridge University Press, 1996; ISBN 9780521460866.
20. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
21. Rennie, J.D.; Shih, L.; Teevan, J.; Karger, D.R. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the Proceedings of the 20th international conference on machine learning (ICML-03); 2003; pp. 616–623.
22. Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179.
23. Breiman, L. Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics* 1998, *26*, 801–849.
24. Fisher, R.A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics* 1936, *7*, 179–188.
25. Team, R.C.; Others R: A language and environment for statistical computing. **2013**.
26. Kuhn, M.; Others Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26.
27. Torgo, L. *Data mining with R: learning with case studies*; Chapman and Hall/CRC, 2011;.
28. Shitikov, V.K.; Mastitsky, S.E. Classification, regression, Data Mining algorithms using R 2017.
29. Vapnik, V.; Chervonenkis, A. Theory of pattern recognition 1974.
30. Ripley, B.; Venables, W. nnet: Feed-forward neural networks and multinomial log-linear models. *R package version* **2011**, *7*.
31. Friedman, J.H. Greedy function approximation: a gradient boosting machine. Department of Statistics. *University of Stanford: Stanfors, CA, USA* **1999**.
32. Mason, L.; Baxter, J.; Bartlett, P.L.; Frean, M.R. Boosting Algorithms as Gradient Descent. In *Advances in Neural Information Processing Systems 12*; Solla, S.A., Leen, T.K., Müller, K., Eds.; MIT Press, 2000; pp. 512–518.