Article

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; siroj1@yandex.ru
[2] Don State Agrarian University, Persianovsky, Rostov region, Russia
[3] ICAI, Inc. Culver City, CA, 90230, USA; drc@authentica.org
[4] Department of Biology, University of La Verne, La Verne, CA, 91750, USA; ttatarinova@laverne.edu
[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
* Correspondence: siroj1@yandex.ru

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.

1. 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].

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

54 Rapidly developing data mining approaches are of increasing interest because they provide for
55 acquisition and analysis of information that results in predictive productivity indicators for animals
56 [5–7]. Machine learning (ML) approaches have been successfully used in animal husbandry for early
57 prediction of the growth and quality of adult wool in Australian merino sheep [8], sheep carcass
58 traits from early-life records [9], and skin temperature of piglets [10]. Compared to other statistical
59 approaches, ML is suitable for use even when there are many predictors, missing values, and
60 abnormally distributed data, which is often the case with data obtained from commercial pig
61 production.

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

67 **2. Results**

68 ML models were able to predict the state of the fore and hind legs. RF surpassed all other learning
69 algorithms in all respects and scenarios. In some cases, RF did not have significant superiority over
70 KNN (Table 1). Accordingly, KNN was the second most efficient algorithm among all the
71 characteristics and scenarios.

72 **Table 1**. 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

73

74

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

79 A graphical interpretation (Figure 1) of the comparative analysis of the predicted value of all models
80 shows that all models in the training set receive more accurate forecasts than in the test set. At the
81 same time, the RF and KNN models provide high accuracy of prediction relative to other models. To
82 determine the models that achieve the best results in solving the problem after the training
83 procedures and their optimization, a comparative analysis was carried out. Obviously, the indicators
84 obtained by validation are estimates of the ability of the model to predict new observations and
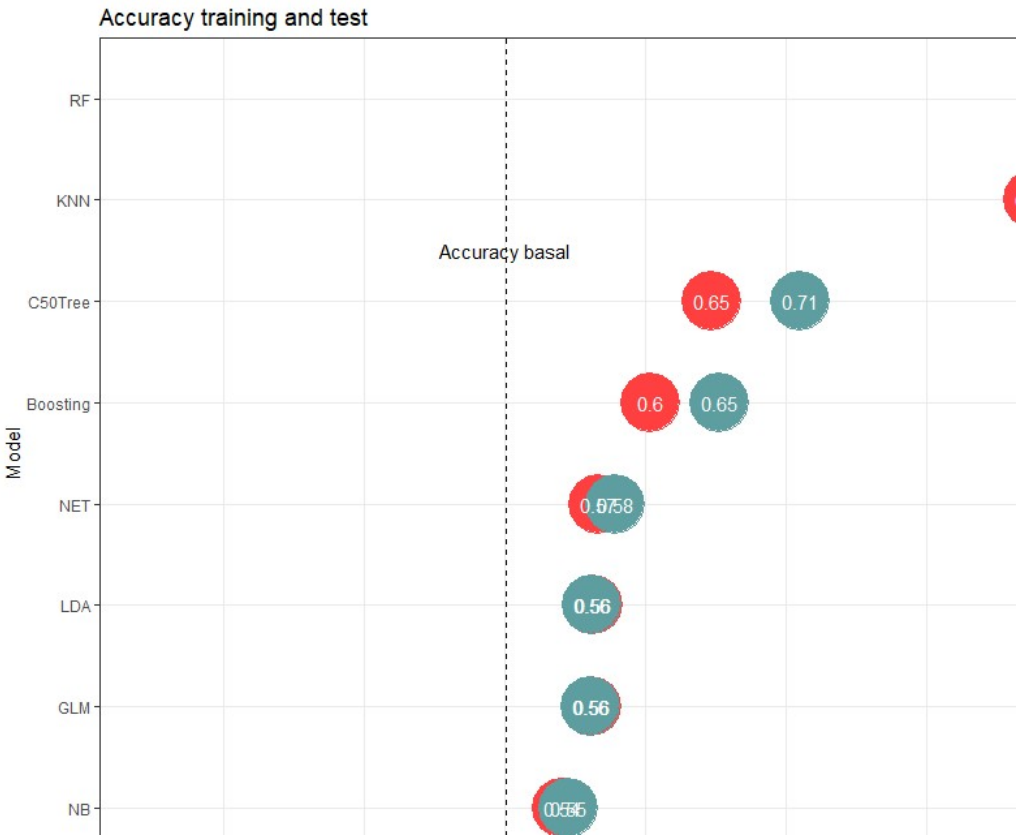85 these estimates have deviations.

86
87 **Figure 1.** Accuracy for training and testing sets for different ML approaches.

88 A comparison was made between all models with the non-parametric Friedman test and a pairwise
89 comparison of all models, the results of which are summarized in Table 2.

90 **Table 2.** Search results among all the models of non-parametric tests of Friedman and paired
91 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

92

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

97 **3. Discussion**

98 The increase in the prevalence of leg weakness in pigs in the middle of the 20th century coincided
99 with a surge of targeted breeding work to increase the growth rate of animals. This was mainly due
100 to economic pressure and the need to shorten the period from birth to slaughter. Moreover, in wild
101 boars, which require about two years to reach maturity, osteochondrosis is not observed. A
102 hypothesis was put forward regarding a significant relationship between growth qualities and
103 weakness of the legs. Several large population studies have shown a positive correlation between
104 these traits [1,11,12]. Lundeheim [13] noted that pigs with clinical signs of leg weakness grew faster
105 in the early stages of life than pigs without these signs, but by the time of slaughter, their growth had
106 become slower. He suggested that the unfavorable relationship between fatness and growth rate is
107 balanced by discomfort due to the emerging clinical signs of leg weakness, leading to reduced feed
108 intake. Van der Wal et al. [14] discovered a significant correlation between the length of the carcass
109 and the weight of the ham with the degree of damage to the proximal and distal parts of the femur -
110 osteochondrosis. The relationship between the state of the legs and indicators of meat productivity
111 of pigs was confirmed by several studies conducted on pigs of various breeds. A study by Draper et
112 al. [15] showed that Duroc pigs with low foreleg scores had greater muscle length and mass. Draper
113 et al. examined thickness of fat, the length of the body and the yield of meat but found no significant
114 differences related to the condition of the legs. In another study, the emphasis was placed on
115 studying the relationship between the legs and meat qualities of large white pigs. The results
116 showed that pigs with leg problems were usually heavier and with more back fat compared to
117 healthy pigs. These observations agree with the results obtained by our machine learning approach.
118 We have demonstrated that that machine learning can be successfully used to evaluate the growth
119 performance and meat characteristics of pigs.

120 **4. Materials and Methods**

121 *4.1 Data sources*
122 Data were taken from 24,584 pigs of breeds Landrace and Large White (Table 3). Several factors can
123 affect the conformation of legs: breed, year of birth (BirthDate), sex, Average Daily Gain (ADG),
124 Muscle thickness (MT), and Back Fat thickness (BF). Front and Back legs were visually assessed
125 using a point system from 1 to 5 (from bad to good). Points 1 and 2 were received by animals with
126 obvious leg defects, 3 points — average condition, 4 and 5 — good and excellent, respectively.
127 Preliminary data analysis showed the imbalance of the available data. Imbalanced classes are a

128 common problem in machine learning classification where there is a disproportionate ratio of
129 observations in each class. Since most ML algorithms work best when the number of samples in each
130 class are about equal, a balancing procedure was applied. After the preliminary analysis, the year of
131 birth and gender were excluded as predictors of the least importance.

132

133 **Table 3.** Sample description. The dataset contains 21,247 females and 3337 males. 12,195 of Landrace
134 and 12389 of Large White breeds. Predictors: Average Daily Gain, Backfat Thickness, Muscle,
135 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

136

137 *4.2 Methods*
138 Classification models were constructed and analyzed using the following machine learning
139 methods: Random Forest (RF) [16], K-Nearest Neighbors (K-NN) [17,18], artificial neural networks
140 (Neural Networks) [19], C50Tree, Support Vector Machines (SVM) [20], Naive Bayes (NB) [21], GLM
141 [22], Boost [23] and Linear Discriminant Analysis (LDA) [24]. All calculations and simulations were
142 performed in R (version 3.6.1, [25]) using the caret packages [26], DMwR [27]. Leg scores was used as
143 the response variables (Table 1).

144

145 *4.2.1 K-Nearest Neighbors (K-NN)*
146 The K-NN classifier is based on the compactness hypothesis (Zagoruiko, 1999), which assumes that
147 the test object d will have the same class label as the training objects in the local area of its immediate
148 environment. When the value of K is one, the analyzed object is assigned to a certain class
149 depending on information about its single nearest neighbor. In the k-NN variant, each object belongs
150 to the prevailing class of nearest neighbors, where k is the algorithm parameter. Any clustering
151 algorithm can be considered effective if the compact hypothesis is satisfied, i.e. one can find such a
152 partition of objects into groups that the distances between objects from the same group (intra-cluster
153 distances) will be less than a certain value $\varepsilon > 0$, and between objects from different groups
154 (cross-cluster distances) more than ε. [28].

155

156 *4.2.2 Linear Discriminant Analysis (LDA)*
157 LDA is a multidimensional analysis section that allows one to evaluate differences between two or
158 more groups of objects by several variables at the same time. It is a generalization of Fisher's linear
159 discriminant, a method used in machine learning to find a linear combination of features which
160 characterizes or separates two or more classes of objects or events. The resulting combination can be
161 used as a linear classifier or, more often, to reduce the dimension before subsequent classification.
162 LDA is closely related to the analysis of variance (ANOVA) procedure. The LDA implements two
163 closely related statistical procedures:

164   1.   Interpretation of group differences, needing to answer the question: how a well-used set of
165   variables can form a dividing surface for objects of the training sample and which of these
166   variables are the most informative.
167   2.   Classification, i.e. prediction of the value of the grouping factor for the examined group of
168   observations.
169

170 *4.2.3 The support vector machines (SVM)*
171 SVM was previously called the "generalized portrait" algorithm, was developed by Soviet
172 mathematicians Vapnik and Chervonenkis [29] and has since gained widespread popularity. The
173 main idea of the classifier on support vectors is to build a separating surface using only a small
174 subset of points lying in the zone critical for separation, while the rest of the correctly classified
175 observations of the training sample outside of this zone are ignored (more precisely, they are a
176 "reservoir" for an optimization algorithm). If there are two classes of observations and a linear form
177 of the boundary between the classes is assumed, then two cases are possible. The first of them relates
178 to the possibility of perfect data separation with the help of some hyperplane. Since there can be
179 many such hyperplanes, the dividing surface is optimal, which is as far as possible from the training
180 points, i.e. having a maximum gap M (margin).
181

182 *4.2.4 Naive Bayes classifier (NB)*
183 Naive Bayes classifiers are a family of simple probabilistic ML classifiers based on the application of
184 Bayes theorem. Making the "naive" assumption that all the signs describing the classified objects are
185 completely equal and are not related to each other, then the probability of an object to belong to a
186 given class given its observed features, *P(class|features)*, is calculated using the Bayes formula from
187 known distributions *P(features|class).* The NB assigns the objects refer to the class that has the
188 greatest probability.
189

190 *4.2.5 Neural Networks*
191 Neural network models that were born in the process of developing the concept of artificial
192 intelligence have two completely transparent analogies - the biological neural system of the brain
193 and the computer network. Their main paradigm is that the solution in the network is formed by
194 many simple neuron-like elements that form a graph with weighted synaptic (informational)
195 connections that work together and purposefully to obtain a common result. To train artificial neural
196 networks in the R environment, the *nnet* package [30] was used; it provides flexible functionality for
197 constructing classification models based on a multilayer perceptron.
198

199 *4.2.6 GLM*
200 Logistic regression is commonly used as a binary classifier for alternate response samples. However,
201 this method can also be generalized to the case with several classes. Nominal or ordinal variables can
202 be used as the simulated response Y, and in both cases a multidimensional binomial distribution is
203 assumed. Simply put, linear regression should be used to predict a quantitative (i.e., numerical)
204 response variable, and logical regression should be used to predict a qualitative (i.e., categorical)
205 response variable. Both linear regression and logistic regression are types of generalized linear
206 models (GLM).

207

208   *4.2.7 Gradient Boosting*

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

219

220   *4.2.8 The C 50Tree method*

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

228

229   *4.2.9 Random Forest*

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

238

239   *4.3 Data Preparation*

240   The number of observations for training models allows one to achieve high predictive effectiveness.
241   The data includes both continuous and high-quality variables, which allows facile problem solving.
242   The response variable (target variable) was a leg score, which varies from 1 to 5. For practical
243   reasons, the values were adjusted and divided into two bins: scores [1:2] - animals with "bad" legs
244   (Q1) and scores [3:5] - animals with "good legs" (Q2). As a metric, Accuracy was calculated as the
245   proportion of correct answers of the algorithm, precision, recall, and the F1 score (a harmonic mean
246   of precision and recall). The data points were assigned to the bins (2708 (4930) for Q1; 21876 (19654)
247   for Q2), corresponding to 11% (20%) and 89% (80%) of measurements for the two breeds. The
248   imbalance of the data classes (a large difference between the numbers of samples in different bins)
249   can negatively affect learning and prediction phases of the approach. If the unbalance ratio is high,

250   the decision function favors the "majority" class, where the largest number of samples is located.
251   Therefore, the leg scores data for the bins were balanced using the ROSE package. Although most
252   machine learning (ML) classification algorithms are generalized to several classes, their
253   interpretation is simpler if there are only two.
254

255   *4.4 Data Analysis*
256   Before choosing the most important predictors and training the prognostic model, a descriptive
257   study of variables was conducted. This process allows for a better understanding of what
258   information each variable contains, as well as to identify possible errors. Since, in practice, it is not
259   always possible to obtain data in its entirety, missing values were found in our data. Using the
260   *preProcess* function from the caret package in R (short for Classification And REgression Training,
261   http://topepo.github.io/caret/index.html), the problem of missing data values was addressed.
262   Studying the distribution of the response variable relative to quantitative (Muscle Thickness,
263   Back Fat, Average Daily Gain) and qualitative (Breed) variables is an equally important procedure.
264   Analysis of quantitative variables showed a pronounced asymmetric distribution of some predictors
265   (Back Fat). The calculation of correlations between continuous predictors indicates that they do not
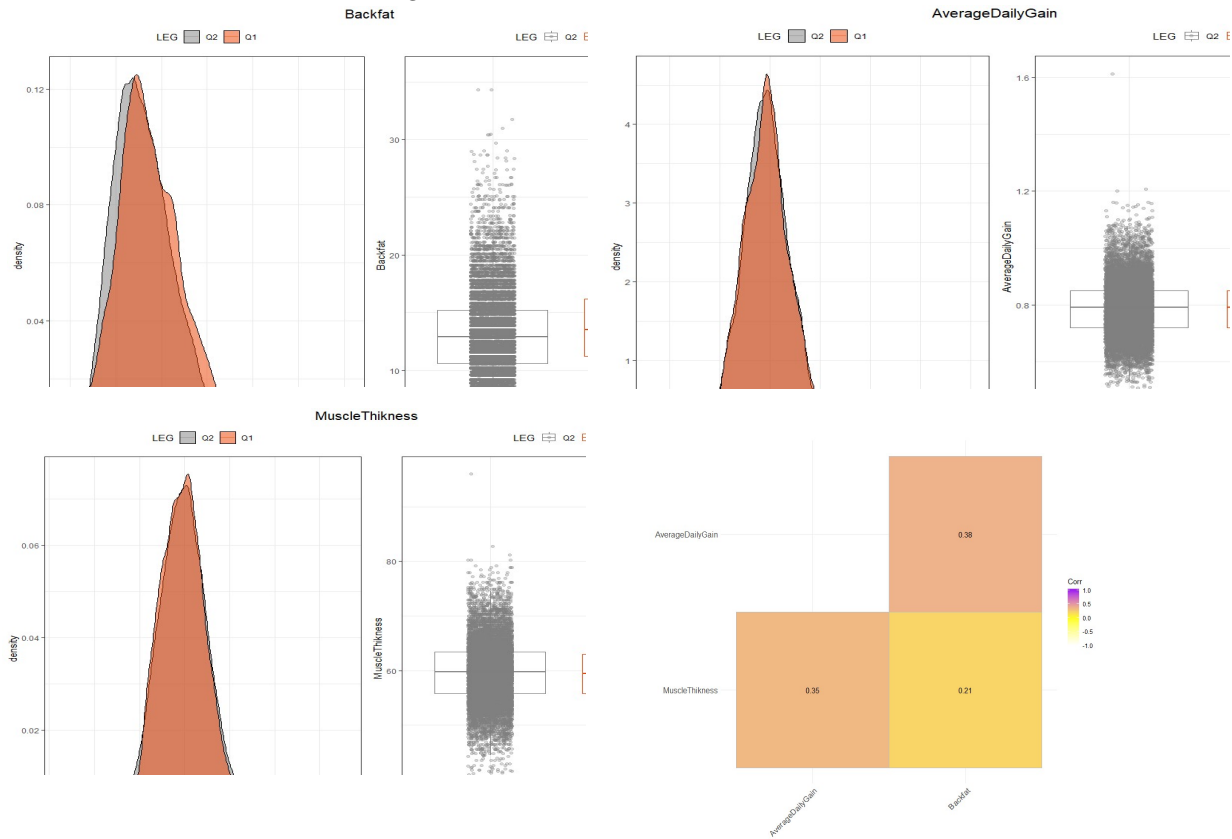266   contain redundant information (Figure 2).

267   **Figure 2**. Visualization of concordance between predictors: Back Fat, Average Daily Gain and
268   Muscle Thickness, showing moderate correlation between the three parameters.
269

270   For the predictive model to be useful, it must have a success rate higher than expected by chance or
271   at a certain base level. In classification problems, the base level is the level obtained if all
272   observations are assigned to the majority class (mod). In our case, given that 89% (80%) of the
273   animals have healthy legs, if Q2 = Yes, it is always predicted, then the success rate will be about 89%

274 (80%) for unbalanced and 50% for balanced data participating in the training set, respectively. This is
275 the minimum percentage that must be overcome with the help of predictive models (strictly
276 speaking, this percentage should be counted only with the help of the training set).

277 Since the aim of the study is to assess the state (conformation) of legs by means of selected
278 predictors (growth and meat quality), each variable is analyzed with respect to the variable Q2 =
279 "good". By analyzing the data in this way, one can begin to extract ideas about which variables are
280 most associated with "good" legs. Alternatively, to study the importance of predictors, we use the
281 Random Forest package. All studies algorithms have identified that that the most important
282 predictors are Muscle Thickness, Back Fat, Average Daily Gain, while the predictor Breed is not
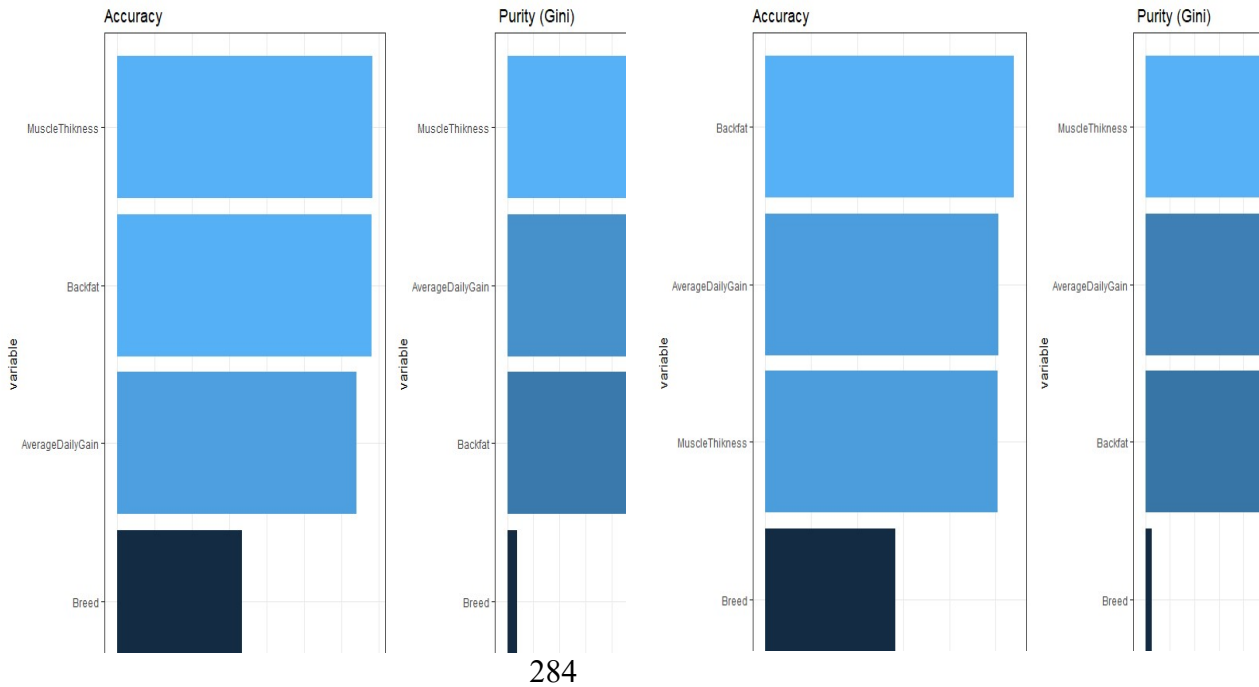283 significant (Figure 3).

284

285 **Figure 3.** Relative importance of leg weakness predictors.

286

287 *4.5 Model training*

288 Figure 2 shows that the measurements follow bell-shaped distributions. Therefore, a
289 standardization of the data was carried out by subtracting the mean and dividing each predictor by
290 its standard deviation, so the data obeys the standard normal distribution.

291 Machine learning algorithms were trained and tested based on the following structure for all
292 three features of interest in this study. A random 10% of the data was excluded from the complete
293 data set for the final assessment, which in our study, will be called a set of independent trials. The
294 remaining 90% were randomly divided into 70% for training and 30% for testing 100 times. In every
295 100 training iterations, hyperparameters were set up using a search within the 10-fold
296 cross-validation structure on a random 70% of the training set. The most effective hyperparameters
297 at this stage were used to train each ML model on a training set and were tested on a test set in each
298 iteration. All processes were implemented in R. The performance of the final model has been
299 evaluated on the test set.

300 **5. Conclusions**

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

306 **Supplementary Materials:** The scripts can be found at http://www.compubioverne.group/data-and-software/.

307 **Author Contributions:** Conceptualization, S.B. and L.G.; methodology, S.B. and T.T.; software, S.B.; validation,
308 S.B.; formal analysis, S.B.; investigation, M.K., L.G and O.K.; resources M.K. and O.K.; data curation, M.K and
309 L.G.; writing—original draft preparation, S.B. and T.T; writing—review and editing, T.T and D.C.;
310 visualization, S.B.; supervision, S.B.; project administration, L.G.; funding acquisition, SB, LG, MK, and OK.

316 **Conflicts of Interest:** The authors declare no conflict of interest.

Abbreviations

NGS	Next Generation Sequencing
QTL	Quantitative Trait Locus
GWAS	Genome Wide Association Studies
ML	Machine Learning
ADG	Average Daily Gain
MT	Muscle thickness
BF	Back Fat
RF	Random Forest
KNN	K-Nearest Neighbors
NN	Neural Networks
SVM	Support Vector Machines
NB	Naïve Bayes
GLM	Generalized Linear Models
LDA	Linear Discriminant Analysis

References

319 1. Ekman, S.; Carlson, C.S. The pathophysiology of osteochondrosis. *Vet. Clin. North Am. Small*
320 *Anim. Pract*. **1998**, *28*, 17–32.

321 2. Le, T.H.; Christensen, O.F.; Nielsen, B.; Sahana, G. Genome-wide association study for
322 conformation traits in three Danish pig breeds. *Genet. Sel. Evol.* **2017**, *49*, 12.

323 3. Aasmundstad, T.; Olsen, D.; Sehested, E.; Vangen, O. The genetic relationships between
324 conformation assessment of gilts and sow production and longevity. *Livest. Sci.* **2014**, *167*,
325 33–40.

326 4. Knauer, M.T.; Cassady, J.P.; Newcom, D.W.; See, M.T. Phenotypic and genetic correlations
327 between gilt estrus, puberty, growth, composition, and structural conformation traits with
328 first-litter reproductive measures. *J. Anim. Sci.* **2011**, *89*, 935–942.

329 5. Morota, G.; Ventura, R.V.; Silva, F.F.; Koyama, M.; Fernando, S.C. BIG DATA ANALYTICS
330 AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data
331 mining advance predictive big data analysis in precision animal agriculture1. *Journal of Animal*

332   *Science* 2018, *96*, 1540–1550.

333 6.   Putz, A.M.; Harding, J.C.S.; Dyck, M.K.; Fortin, F.; Plastow, G.S.; Dekkers, J.C.M.; PigGen
334   Canada Novel Resilience Phenotypes Using Feed Intake Data From a Natural Disease
335   Challenge Model in Wean-to-Finish Pigs. *Front. Genet.* **2018**, *9*, 660.

336 7.   Howard, J.T. The use of "Big Data" in a modern swine breeding program now and in the
337   future.

338 8.   Shahinfar, S.; Kahn, L. Machine learning approaches for early prediction of adult wool growth
339   and quality in Australian Merino sheep. *Comput. Electron. Agric.* **2018**, *148*, 72–81.

340 9.   Shahinfar, S.; Kelman, K.; Kahn, L. Prediction of sheep carcass traits from early-life records
341   using machine learning. *Comput. Electron. Agric.* **2019**, *156*, 159–177.

342 10.   Gorczyca, M.T.; Milan, H.F.M.; Maia, A.S.C.; Gebremedhin, K.G. Machine learning algorithms
343   to predict core, skin, and hair-coat temperatures of piglets. *Comput. Electron. Agric.* **2018**, *151*,
344   286–294.

345 11.   Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

346 12.   Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information*
347   *Theory* 1967, *13*, 21–27.

348 13.   Lantz, B. *Machine Learning with R: Expert techniques for predictive modeling to solve all your data*
349   *analysis problems*; Packt Publishing Ltd, 2015; ISBN 9781784394523.

350 14.   Ripley, B.D.; Hjort, N.L. *Pattern Recognition and Neural Networks*; Cambridge University Press,
351   1996; ISBN 9780521460866.

352 15.   Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.

353 16.   Rennie, J.D.; Shih, L.; Teevan, J.; Karger, D.R. Tackling the poor assumptions of naive bayes text
354   classifiers. In Proceedings of the Proceedings of the 20th international conference on machine
355   learning (ICML-03); 2003; pp. 616–623.

356 17.   Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several
357   independent variables. *Biometrika* **1967**, *54*, 167–179.

358 18.   Breiman, L. Arcing classifier (with discussion and a rejoinder by the author). *The Annals of*
359   *Statistics* 1998, *26*, 801–849.

360 19.   Fisher, R.A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS.
361   *Annals of Eugenics* 1936, *7*, 179–188.

362 20.   Team, R.C.; Others R: A language and environment for statistical computing. **2013**.

363 21.   Kuhn, M.; Others Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**,
364   *28*, 1–26.

365 22.   Torgo, L. *Data mining with R: learning with case studies*; Chapman and Hall/CRC, 2011;.

366 23.   Shitikov, V.K.; Mastitsky, S.E. Classification, regression, Data Mining algorithms using R 2017.

367 24.   Vapnik, V.; Chervonenkis, A. Theory of pattern recognition 1974.

368 25.   Ripley, B.; Venables, W. nnet: Feed-forward neural networks and multinomial log-linear
369   models. *R package version* **2011**, *7*.

370 26.   Friedman, J.H. Greedy function approximation: a gradient boosting machine. Department of
371   Statistics. *University of Stanford: Stanfors, CA, USA* **1999**.

372 27.   Mason, L.; Baxter, J.; Bartlett, P.L.; Frean, M.R. Boosting Algorithms as Gradient Descent. In
373   *Advances in Neural Information Processing Systems 12*; Solla, S.A., Leen, T.K., Müller, K., Eds.; MIT
374   Press, 2000; pp. 512–518.

375 28. Jørgensen, B.; Andersen, S. Genetic parameters for osteochondrosis in Danish Landrace and
376   Yorkshire boars and correlations with leg weakness and production traits. *Anim. Sci.* **2000**, *71*,
377   427–434.
378 29. Nakano, T.; Brennan, J.J.; Aherne, F.X. LEG WEAKNESS AND OSTEOCHONDROSIS IN
379   SWINE: A REVIEW. *Can. J. Anim. Sci.* **1987**, *67*, 883–901.
380 30. Lundeheim, N. Genetic Analysis of Osteochondrosis and Leg Weakness in the Swedish Pig
381   Progeny Testing Scheme. *Acta Agriculturae Scandinavica* 1987, *37*, 159–173.
382 31. Van der Wal, P.G.; Van der Valk, P.C.; Goedegebuure, S.A.; Van Essen, G. Osteochondrosis in
383   six breeds of slaughter pigs: II. Data concerning carcass characteristics in relation to
384   osteochondrosis. *Vet. Q.* **1980**, *2*, 42–47.
385 32. Draper, D.D.; Rothschild, M.F.; Christian, L.L. Effects of divergent selection for leg weakness
386   on muscle and bone characteristics in Duroc swine. *Genet. Sel. Evol.* **1992**, *24*, 363.