Machine learning algorithms based on haplotype libraries for classification of stillbirth susceptibility in Holstein cows

Pablo A.S. Fonseca[1]; Massimo Tornatore[2]; Angela Cánovas[1]

[1]Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON N1G 2W1, Canada
[2]Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milano, Italy

Reduced fertility is one of the main causes of economic losses in dairy farms. The cost of a stillbirth is estimated in US$ 938 per case in Holstein herds. Machine learning (ML) is gaining popularity in the livestock sector as a mean to identify hidden patterns and due to its potential to address dimensionality problems. Here we investigate the application of ML algorithms for the prediction of cows with higher stillbirth susceptibility in two scenarios: cows with >25% and >33.33% of stillbirths among birth records. These thresholds correspond to percentiles 75 (still_75) and 90 (still_90), respectively. A total of 10,570 cows and 50,541 birth records were collected to perform a haplotype-based genome-wide association study. Five-hundred significant pseudo single nucleotide polymorphisms (pseudo-SNPs) (False-Discovery Rate<0.05) were used as input features of ML-based predictions to determine if the cow is in the top-75 and top-90 percentiles. Table 1 shows the classification performance of the investigated ML and linear models. The ML models outperformed linear models for both thresholds. In general, still_75 showed higher F1 values compared to still_90, suggesting a lower misclassification ratio when a less stringent threshold is used. We observe that accuracy of the models in our study is higher when compared to ML-based prediction accuracies in other breeds, e.g. compared to the accuracies of 0.46 and 0.67 that were achieved using SNPs for body weight in Brahman and fertility traits in Nellore, respectively. Xgboost algorithm shows the highest balanced accuracy (BA; 0.625), F1-score (0.588) and area under the curve (AUC; 0.688), suggesting that xgboost can achieve the highest predictive performance and the lowest difference in misclassification ratio between classes. The ML applied over haplotype libraries is an interesting approach for the detection of animals with higher susceptibility to stillbirths due to highest predictive accuracy and relatively lower misclassification ratio.

Keywords: Stillbirth; Machine Learning; Haplotypes; Genomic Prediction; Holstein dairy cattle.

Table 1: Performance metrics for the machine learning (ML) and linear algorithms used to classify stillbirth susceptibility in two scenarios (percentiles 75 (still_75) and 90 (still_90) of stillbirth cases among the birth records).

Algorithm	Model	Scenario	Acc	BA	F1	AUC	Prec	RMSE
xgboost	ML	Still_75	0.591	0.625	0.588	0.688	0.476	0.493
xgboost	ML	Still_90	0.710	0.624	0.436	0.690	0.417	0.439
svmLinearWeights	ML	Still_75	0.584	0.609	0.566	0.609	0.469	0.645
svmLinearWeights	ML	Still_90	0.558	0.613	0.446	0.613	0.323	0.665
naivebayes	ML	Still_75	0.659	0.630	0.532	0.630	0.557	0.583
naivebayes	ML	Still_90	0.675	0.615	0.430	0.615	0.380	0.570
RF	ML	Still_75	0.663	0.590	0.392	0.590	0.622	0.581
RF	ML	Still_90	0.757	0.519	0.088	0.519	0.627	0.492
glm	Linear	Still_75	0.665	0.618	0.490	0.618	0.582	0.578
glm	Linear	Still_90	0.748	0.578	0.322	0.578	0.482	0.502
bayesglm	Linear	Still_75	0.667	0.620	0.492	0.620	0.585	0.577
bayesglm	Linear	Still_90	0.748	0.577	0.312	0.577	0.479	0.502

Legend: Extreme Gradient Boosting (xgboost), L2 Regularized Linear Support Vector Machines with Class Weights (svmLinearWeights), Multi-Layer Perceptron (mlpWeightDecayML), Naive Bayes (naivebayes), Random Forest (RF), Generalized Linear Model (glm), Bayesian Generalized Linear Model (bayesglm), Accuracy (Acc), Balanced Accuracy (BA), F1-score (F1), Area Under the Curve (AUC), Precision (Prec), and Residual Mean Squared Error (RMSE).