Machine learning methods for genomic prediction of cow behavioral traits measured by automatic milking systems in North American Holstein cattle

Victor B. Pedrosa,1 Shi-Yi Chen,1,2 Leonardo S. Gloria,1 Jarrod S. Doucette,3 Jacquelyn P. Boerman,1 Guilherme J. M. Rosa,4 and Luiz F. Brito1*

1Department of Animal Sciences, Purdue University, West Lafayette, IN 47907
2Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, Sichuan, 611130, China
3Agriculture Information Technology (AgIT), Purdue University, West Lafayette, IN 47907
4Department of Animal and Dairy Sciences, University of Wisconsin–Madison, Madison, WI, 53706

ABSTRACT

Identifying genome-enabled methods that provide more accurate genomic prediction is crucial when evaluating complex traits such as dairy cow behavior. In this study, we aimed to compare the predictive performance of traditional genomic prediction methods and deep learning algorithms for genomic prediction of milking refusals (MREF) and milking failures (MFAIL) in North American Holstein cows measured by automatic milking systems (milking robots). A total of 1,993,509 daily records from 4,511 genotyped Holstein cows were collected by 36 milking robot stations. After quality control, 57,600 SNPs were available for the analyses. Four genomic prediction methods were considered: Bayesian least absolute shrinkage and selection operator (LASSO), multiple layer perceptron (MLP), convolutional neural network (CNN), and GBLUP. We implemented the first 3 methods using the Keras and TensorFlow libraries in Python (v.3.9) but the GBLUP method was implemented using the BLUPF90+ family programs. The accuracy of genomic prediction (mean square error) for MREF and MFAIL was 0.34 (0.08) and 0.27 (0.08) based on LASSO, 0.36 (0.09) and 0.32 (0.09) for MLP, 0.37 (0.08) and 0.30 (0.09) for CNN, and 0.35 (0.09) and 0.31(0.09) based on GBLUP, respectively. Additionally, we observed a lower reranking of top selected individuals based on the MLP versus CNN methods compared with the other approaches for both MREF and MFAIL. Although the deep learning methods showed slightly higher accuracies than GBLUP, the results may not be sufficient to justify their use over traditional methods due to their higher computational demand and the difficulty of performing genomic prediction for nongenotyped individuals using deep learning procedures. Overall, this study provides insights into the potential feasibility of using deep learning methods to enhance genomic prediction accuracy for behavioral traits in livestock. Further research is needed to determine their practical applicability to large dairy cattle breeding programs.

Key words: accuracy of prediction, automatic milking systems, deep learning, sensor-based systems

INTRODUCTION

Advances in sensor-based high-throughput phenotyping and artificial intelligence have enabled the development of precision livestock farming technologies, especially for deriving novel traits and improving the accuracy of genomic predictions (Pérez-Enciso and Zingaretti, 2019; Fuentes et al., 2022). In this context, machine learning (ML) models excel at handling large amounts of data that would typically require substantial computational capacity and advanced learning algorithms. However, although ML methods, including deep learning, exhibit a high degree of flexibility that allows them to effectively generalize and capture intricate relationships, their implementation can be considerably more complex than traditional methods (Gianola et al., 2011; González-Recio and Forni, 2011; Abdollahi-Arpanahi et al., 2020). Deep learning (DL), a subgroup of ML procedures, was inspired by brain functionality using neural networks with many nodes and layers to establish more efficient and accurate prediction models (Abdollahi-Arpanahi et al., 2020). Deep learning methods consist of a heterogeneous set of algorithms that allow the implementation of promising digital technologies and livestock monitoring systems (Morota et al., 2018; Karthick et al., 2020). Deep learning modeling...
techniques have been used to create practical artificial intelligence applications focused on assessing animal behavior, welfare, and physiological changes (e.g., Neethirajan et al., 2017; Halachmi et al., 2019; Neethirajan and Kemp, 2021).

The development of genomic prediction models applying DL algorithms with multiple neuron layers and numerous hyperparameters has been proposed for the analysis of complex traits in humans as well as agricultural and livestock species (e.g., Bellot et al., 2018; Waldmann, 2018; Passafaro et al., 2020; Varona et al., 2022). Recent studies reported that DL algorithms outperformed the GBLUP and Bayesian least absolute shrinkage and selection operator (LASSO) methods for genomic predictions of wheat, maize, and pig traits, considering multiple layer perceptron (MLP), recurrent neural network, and convolutional neural network (CNN; Ma et al., 2018; Montesinos-López et al., 2018; Abdollahi-Arpanahi et al., 2020). In contrast, other authors pointed out the necessity for clearer evidence on the improvement of prediction accuracy when fitting DL algorithms (Passafaro et al., 2020; Zingaretti et al., 2020; Fuentes et al., 2022), indicating that additional studies should compare DL methods and commonly used genomic prediction techniques.

In general, genomic prediction models are based on linear functions, such as GBLUP (VanRaden, 2008) and Bayesian regression methods (Goddard and Hayes, 2009; Pérez and de los Campos, 2014), which can limit modeling design when nonlinear functions are intrinsic in the data distribution (Guo et al., 2010; Erbe et al., 2012). Furthermore, in some genomic prediction analyses, the number of markers fitted is higher than the number of individuals, which could cause statistical problems depending on the method used (de los Campos et al., 2013; Pérez and de los Campos, 2014). In practice, a potential benefit of DL algorithms is the ability to possibly predict the complete genetic value of an individual, including dominance and epistasis, and any interactions irrespective of their origin (Ma et al., 2018; Montesinos-López et al., 2021). In this context, the primary objective of this study was to evaluate the predictive performance of DL methods compared with LASSO and GBLUP for genomic prediction of behavioral traits in North American Holstein cattle.

**MATERIALS AND METHODS**

**Ethics Statement**

Animal Care Committee approval was not necessary for the purposes of this study because all the information required was obtained from pre-existing databases.

**Datasets**

A total of 1,993,509 daily records from 4,511 genotyped Holstein cows were collected using 36 milking robot stations. A set of 57,600 SNPs was used for the analyses after quality control (QC). The QC was performed using the BLUPF90+ family programs (Misztal et al., 2018) and removed animals and SNPs with a call rate lower than 0.90, SNPs located on nonautosomal chromosomes, SNPs with minor allele frequency lower than 0.05, and extreme departure from Hardy–Weinberg equilibrium assessed by the difference between observed and expected heterozygosity greater than 0.15. In the end, out of the total markers retained after QC, the 1,000 most significant SNPs (based on their P-value) were selected for milking refusals (MREF) and milking failures (MFAIL) for further analyses using a GWAS approach, as recommended by Pérez-Enciso and Zingaretti (2019), when genomic predictions (i.e., not genomic estimated breeding values) were performed based on DL approaches. Using large numbers of markers with minimal or no effect can significantly diminish the training efficacy of DL methods. Hence, the preselection of markers can enhance training proficiency and optimize the structuring of ML models. The original number of markers (50K) remained the same for GBLUP to replicate as closely as possible what would be done in standard analyses based on this method. Furthermore, for comparison purposes, GBLUP was also implemented using only the 1,000 most significant SNPs included in the DL analyses.

The behavioral indicators MREF and MFAIL were measured in the automatic milking systems (AMS), as described by Pedrosa et al. (2023). Milking refusal is the total number of refused milking activities by the AMS units on a day. A certain number of refusals per cow during the lactation season are normally caused by (1) level of animal adaptation to the AMS; (2) high feed-dispenser speed, leading to leftovers of concentrate in the AMS; and (3) low feed amount provided at the AMS feeder. An average of MREF was considered for each animal to determine the final MREF for the entire studied period. For MFAIL, which is a trait that represents the total number of milking failures on a day, an average of MFAIL was considered for each animal to determine the final MFAIL, as was done for MREF. The main reasons for milking failures are (1) teats not found by the robot; (2) connection time, where teats were detected but the milker robot was unable to connect one of the teats; (3) the AMS stopped because a cow left the AMS unit before the gate was closed; and (4) dead milking time, in which teat cups were successfully connected but there was no milk flow.
from one or more of the udder quarters. Milking failure is constantly related to excessive animal movement inside the AMS. Usually, this excessive movement can be linked to the animal’s behavior associated with stress due to adaptation to the use of the AMS.

**Statistical Methods**

Two DL algorithms (MLP and CNN) were evaluated and compared with 2 traditional genomic prediction methods (LASSO and GBLUP). A brief description of each statistical approach and their methods is presented in the following sections. For additional details about the methods used, please see Tibshirani (1996), Legarra et al. (2011), Li and Sillanpää (2012), Goodfellow et al. (2016), Pérez-Enciso and Zingaretti (2019), and Lorenco et al. (2020). The variables were analyzed based on phenotypic records pre-adjusted for fixed effects. The fixed effects considered in the pre-adjustment model were the concatenation of calving year and season of calving (March to May, June to August, September to November, and December to February), lactation number (1, 2, or 3+), and DIM (5–305 d).

**Deep Learning Algorithms**

**Multiple Layer Perceptron.** The neural network (NN) concept was proposed in late 1940s and early 1950s with the purpose of computationally reproducing the brain functions (Rosenblatt, 1962). A DL general structure, based on the NN concept, is made of a chain of several layers composed of various nodes (or neurons) able to approximate complex relationships between input variables (predictors) and an outcome (response variable; Goodfellow et al., 2016). The MLP is known as feed-forward NN because of the sequence of connected information that allows the next layers to take advantage of the previously processed information to maximize the estimates until the output. The MLP is composed of 3 types of layers: input layer, hidden layers, and output layer. Figure 1 shows an example of an MLP with an input layer with 3 predictors, 2 hidden layers, and 1 output layer. For genomic prediction purposes the first layer can be composed of SNP genotypes as an input and the information is processed sequentially by each hidden layer as a weighted nonlinear function of each input plus the bias.

\[
\hat{y} = \sigma (XW_1 + B) W_2,
\]

where \(\hat{y}\) is the vector of predictions, \(W_1\) and \(W_2\) are the weight matrices that connect the input matrix \(X\) (SNP matrix) to the \(\hat{y}\) output layer across the hidden layers. The dimension of the \(W\) matrixes relies on the number of neurons for each specific layer. The parameter \(\sigma\) is the activation function for the connection of 2 consecutive layers, and \(B\) is the bias matrix associated with \(W_1\) and \(W_2\).

The MLP approach was implemented using the Python (v.3.9) software, in which the DL methods were implemented using the Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2016) packages. After a grid search using the scikit-learn tool (Pedregosa et al., 2011), in which the values from the whole dataset were considered, the hyperparameters applied were the optimization algorithm = stochastic gradient descent, epochs = 30, learning rate = 0.01, neurons for the first hidden layer = 96, and neurons for the second hidden layer = 64. The nonlinear activation function for the first hidden layer was the rectifier linear unit and the soft rectifier linear unit was applied for the second one.

**Convolutional Neural Networks.** The CNN method is a type of NN that exhibits consistent patterns among the inputs during the ML processing (LeCun and Bengio, 1995). An example of this consistency, as it relates to genomic predictions, could be seen as the linkage disequilibrium between nearby SNPs (Abdollahi-Arpanahi et al., 2020). In the CNN method, the hidden layers are sequentially organized into convolutional layers, pooling layers, and fully connected layers. In each convolutional layer, a convolutional operation is performed in accordance with the input width and strides. These convolutional operations are often referred to as kernels and can be interpreted as an equivalent of the neurons in the MLP method. The activation functions are applied to each convolution layer to generate the outputs. Pooling operations are employed to refine the results, merging the kernel of different positions considering the mean, maximum,
or minimum values of these positions (Pérez-Enciso and Zingaretti, 2019). Figure 2 illustrates an example of the CNN structure in an ML application.

Similar to MLP, CNN was implemented using the Python (v.3.9) software and the Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2016) packages. The CNN was structured using 1 input layer (genotypic matrix), 1 convolutional layer, 1 pooling layer, 2 layers (neurons = 128 and neurons = 96), 1 dropout layer, and 1 output layer. The additional hyperparameters used were epochs = 30, learning rate = 0.01, and the stochastic gradient descent optimization algorithm. The Tanh activation function was applied in the convolutional layer, and the soft rectifier linear unit function was subsequently used in the fully connected layers.

Traditional Statistical Methods

Genomic Best Linear Unbiased Prediction. Genomic best linear unbiased prediction is a popular method that uses genomic relationships to predict the genetic merit of individuals of interest. In this approach a genomic-relationships matrix is constructed based on a large number of SNP markers distributed across the genome. This matrix specifies the covariance between individuals considering their similarity at the genomic level, seeking for more accurate predictions of the genetic merit of individuals (Clark and van der Werf, 2013). The GBLUP method has been widely adopted in livestock breeding programs, being one of the most commonly used methods in breeding programs since the beginning of the genomic era (VanRaden, 2008; Misztal et al., 2020). The general statistical model of single-trait GBLUP can be written as:

\[ y = 1\mu + Zg + e, \]

where \( y \) is an \( n \)-vector of phenotypes adjusted for systematic effects, \( 1 \) is a \( n \)-vector of ones, \( \mu \) is the population mean, \( g \) is a vector of additive genetic values, \( Z \) is a design matrix that allocates observations to genetic values, and \( e \) is the vector of random residual effects. No pedigree information was available, and therefore, all the analyses were based on a GBLUP approach (instead of single-step GBLUP) using the BLUPF90+ family programs (Misztal et al., 2018).

Bayesian LASSO. The LASSO method (de los Campos et al., 2009) was also used to perform genomic prediction because it is a traditional approach used in livestock breeding programs and, therefore, an important comparison method to the ML methods tested here. In quantitative genetics, multiple linear regression approaches are frequently used to explain the relationship existing between phenotypes and markers. Based on the marker-effect estimates, genomic predictions can be made for a new set of genotyped animals (typically young animals with no phenotypic measurements). Because the number of SNPs can be larger than the number of individuals tested, the model tends to be overparameterized (de los Campos et al., 2009). To minimize this issue, a penalized regression can alternatively be employed as proposed in the LASSO approach. It provides sparse estimation of regression coefficients by adding \( \ell_1 \) penalty functions to the traditional least squares (Tibshirani, 1996). In LASSO, the penalty term \( \sum_{j=1}^{p} |\beta_j| \) with a coefficient \( \lambda \), promotes sparsity in the regression coefficients. It is worth noting that LASSO can be connected to a Bayesian framework by considering double exponential (Laplace) prior distributions for the regression coefficients. In this context, the \( \lambda \) parameter in LASSO can be related to the scale parameter of the Laplace prior. The Bayesian LASSO estimates can then be seen as posterior mode estimates under Laplace prior distributions (Tibshirani, 1996), as described below:
$$p(\beta) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp \left( -\lambda | \beta_j | \right).$$

This connection between LASSO and Bayesian LASSO provides a Bayesian interpretation of the regularization technique, where the penalty term induces sparsity in the coefficients, similar to how the Laplace prior results in sparsity in a Bayesian context. Differently from the standard LASSO in which the shrinkage factor ($\lambda$) must be explicitly selected, in the Bayesian LASSO, a prior distribution is assigned to $\lambda^2$, so that the value of $\lambda$ can be estimated together with the other model parameters. A common option is to use a conjugate Gamma(a,b) prior for $\lambda^2$, in which $a > 0$ and $b > 0$ are predetermined hyperparameters (Li and Sillanpää, 2012). Based on that, a biological interpretation can be determined by the shape of the distribution of the SNP effects in the LASSO model (Legarra et al., 2011).

**Comparison Methods.** The accuracy of genomic predictions for all the methods analyzed was calculated considering the correlation between the predicted and observed values (pre-adjusted phenotypes), as well as the mean squared error of prediction ($\text{MSE}$) utilizing a 5-fold cross-validation. In addition, we also presented the slope of the regression of observed values on predicted values ($b_1$). Selected and culled animals were evaluated by comparing all pairs of models, considering a selection threshold of 20% for the animals with the best genomic prediction for each trait. Thus, a scatter plot was constructed based on (1) animals selected based on both methods; (2) animals culled based on both methods, and (3) animals selected by one method but discarded in the other method, to identify the clustering of coinciding animals selected according to the different methods evaluated.

**RESULTS AND DISCUSSION**

**Validation Accuracy of Genomic Predictions**

Moderate cross-validation genomic prediction accuracies were obtained with all methods evaluated in this study, for both MREF and MFAIL, with very similar MSE between approaches. Table 1 shows the prediction accuracy and the MSE for the 4 methods evaluated. The highest accuracy of prediction for MREF was found in CNN (0.369), followed by MLP (0.363), GBLUP (0.352), and LASSO (0.343). As shown in Table 1, slightly higher accuracy was found in the analyses using ML methods, evidencing the prediction potential of current artificial intelligence tools. For MFAIL, the upper genomic prediction accuracy was obtained for MLP (0.317), followed by GBLUP (0.307), CNN (0.296), and LASSO (0.274).

<table>
<thead>
<tr>
<th>Parameter and method</th>
<th>MREF</th>
<th>MFAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>LASSO 0.343±0.032</td>
<td>0.274±0.033</td>
</tr>
<tr>
<td></td>
<td>MLP 0.363±0.033</td>
<td>0.317±0.032</td>
</tr>
<tr>
<td></td>
<td>CNN 0.369±0.032</td>
<td>0.296±0.033</td>
</tr>
<tr>
<td></td>
<td>GBLUP 0.352±0.032</td>
<td>0.307±0.031</td>
</tr>
<tr>
<td>MSE</td>
<td>LASSO 0.081±0.007</td>
<td>0.083±0.008</td>
</tr>
<tr>
<td></td>
<td>MLP 0.087±0.008</td>
<td>0.089±0.009</td>
</tr>
<tr>
<td></td>
<td>CNN 0.077±0.007</td>
<td>0.089±0.008</td>
</tr>
<tr>
<td></td>
<td>GBLUP 0.092±0.008</td>
<td>0.090±0.009</td>
</tr>
<tr>
<td>Regression slope ($b_1$)</td>
<td>LASSO 0.299±0.028</td>
<td>0.268±0.031</td>
</tr>
<tr>
<td></td>
<td>MLP 0.329±0.025</td>
<td>0.305±0.028</td>
</tr>
<tr>
<td></td>
<td>CNN 0.366±0.027</td>
<td>0.287±0.030</td>
</tr>
<tr>
<td></td>
<td>GBLUP 0.338±0.029</td>
<td>0.301±0.031</td>
</tr>
</tbody>
</table>

For both traits, LASSO was the method with the lowest prediction performance. Out of the traditional methods, GBLUP yielded the best predictive performance and outperformed CNN for MFAIL. As a point of comparison, GBLUP was also evaluated using the same 1,000 SNPs from the DL methods, which resulted in reduced prediction accuracy ($\pm$ SE) of 0.239±0.044 for MREF and 0.207±0.048 for MFAIL.

The results of the regression slopes ($b_1$) presented a similar pattern as that observed in the accuracy of predictions. In this context, it is important to strike a balance when selecting the optimal genomic prediction method. This decision must consider not only the similarity of the results but also the complexity of the analyses required by each approach, as well as the potential benefits of using the entire dataset available. Behavioral traits are often complex and influenced by numerous factors, including genetics, environment, and their interactions. The complex inheritance of behavioral traits, sample size, and low heritability of these behavioral traits (Pedrosa et al., 2023) are potential explanations for the low slopes found here.

Most of the early genomic prediction methods have succeeded in calculating genetic merits and other estimates related to the genetic potential of individuals in efficient ways (VanRaden, 2008; Goddard and Hayes, 2009; Zhu et al., 2021). More recently, ML-based genetic prediction methods, despite being in an early stage of testing and with still uncertain effectiveness, have been evaluated as a new alternative for performing genomic predictions. The use of DL algorithms for implementing multilayer artificial neural networks may potentially enhance genomic selection and livestock management at the herd level (Weigel et al., 2017). Furthermore, there is...
also the possibility of using DL methods to identify candidate genes influencing biological structures or functions related to livestock, crops, and humans (Fu et al., 2020; Le, 2020; Bi and Hu, 2021). An additional strength of GBLUP compared with other methods is its potential for easy adaptation in obtaining GEBV for non-genotyped individuals (Lourenco et al., 2020; Garrick and Fernando, 2022). This adaptability could involve leveraging ML techniques like fitting the Cholesky decomposition or singular value decomposition of the $H$ matrix, although this concept still demands further investigation and validation. Moreover, GBLUP and single-step GBLUP can be implemented in a relatively straightforward manner in multitrait analyses. Another point is the issue of the high computational demand usually required by ML analyses, which depending on the sample size, can be impractical (Eraslan et al., 2019; Pérez-Enciso and Zingaretti, 2019).

Few genomic prediction studies have compared ML methods for genomic predictions in livestock or plants. In a Holstein cattle study evaluating sire conception rate based on different genomic prediction methods (Abdollahi-Arpanahi et al., 2020), GBLUP (0.33) presented the highest genomic prediction accuracy in comparison to CNN (0.29) and MLP (0.26). The authors reported that the MSE for GBLUP was lower compared with CNN and MLP, representing a better control of the genomic prediction bias (Abdollahi-Arpanahi et al., 2020). In another study using simulated data, the authors reported that the genomic prediction accuracy for 2 traits (one in cattle and one in pigs) was slightly higher based on GBLUP than DL models (Han et al., 2021). The authors mentioned that these results were not surprising because GBLUP is well suited for a substantial number of small effects that approximate the infinitesimal model (Han et al., 2021). The magnitude of genomic prediction accuracies is also dependent on the heritabilities of the evaluated traits. In our study, both traits (MREF and MFAIL) are lowly heritable (0.09 and 0.02, respectively; Pedrosa et al., 2023). However, it is not known whether even greater accuracies would be obtained based on ML methods when evaluating traits with higher heritability.

**Genetic Dispersion and Selection Characterization for MREF**

Figure 3 shows the scatter plots for MREF, indicating the dispersion resulting from each estimate calculated according to the method used. Higher dispersions were observed for the LASSO approach compared with the other methods. In contrast, the distribution between observed and predicted values for the DL methods and GBLUP was quite similar, demonstrating a greater central clustering of inference points, which denotes less dispersion. In terms of the traditional genomic prediction methods, similar results were obtained in crop studies, where the genetic variation was slightly lower when GBLUP was compared with LASSO and other Bayesian approaches (Bhering et al., 2015). Regardless of the method, genomic prediction constantly suffers variation with great dependence on the sample size and heritability estimates, resulting in under- or overestimation of the true predictive accuracy (Ould Estaghvirou et al., 2013; Holland et al., 2020). In general, traditional genomic prediction methods do not consider nonadditive genetic effects, which limits the modeling of the total genetic variance (Vitezica et al., 2017; Varona et al., 2018). The importance of accounting for the contribution of other genetic factors, potentially expressing epistatic effects, is often ignored, highlighting the importance of considering this source of variation within the genomic prediction models, as is usually done in DL prediction methods (Santantonio et al., 2019; Zingaretti et al., 2020).

For MREF, among the DL methods evaluated, the genomic prediction accuracy of CNN was 0.006 higher than MLP, which represents practically equal prediction effectiveness between the 2 DL approaches. However, when comparing CNN with GBLUP and LASSO, the prediction variation increased by 0.017 and 0.026, respectively. Although the increase seems relevant, when evaluating the absolute numbers, the genomic prediction accuracies are similar. This reinforces the conclusion that, in addition to observing only small differences in genomic prediction accuracies, other factors must be considered when choosing the genomic prediction method to be used for practical applications in breeding programs. In Holstein cattle, the genomic prediction accuracy for sire conception rate was 12% higher for CNN compared to MLP (Abdollahi-Arpanahi et al., 2020), similar to the results found here for MREF. Artificial NN are mathematical models that do not aim to provide biologically realistic representations (Emmett-Streib et al., 2020). Instead, these models tend to be used for analyzing data and making predictions based on the information provided. Thus, there is no prior prediction about which DL model will be the most suitable for each situation because they are data dependent and vary according to the structure of the data and the purpose for which the ML is processed (Schmidhuber, 2015).

In terms of genetic selection, DL techniques have been tested to predict genetic merit based on pre-adjusted phenotypic values, classify animals based on their genotype, and identify genomic regions associated with important traits (González-Recio et al., 2014). However, the most widely used statistical methods in genetic selection, such as GBLUP and Bayesian approaches (e.g., LASSO), have been shown to be effective in predicting GEBV and enabling genetic progress in animal and plant breeding programs (VanRaden, 2008; Legarra et al., 2011). In this
study, we evaluated the selective potential of traditional methods of genomic predictions compared with DL methods to identify the methods better able to select the best individuals for breeding purposes. The Cartesian selection graph for MREF in Figure 4 demonstrates the top 20% individuals selected for each method (evaluated 2 × 2), divided in categories as selected simultaneously for both methods, selected in only one of the methods, and not selected by any of the methods.

The methods that presented the highest number of commonly selected individuals for MREF were MLP × CNN (60%), followed by LASSO × CNN (56%), CNN × GBLUP (50%), LASSO × MLP (49%), LASSO × GBLUP (46%), and MLP × GBLUP (45%). The CNN method was the one with the highest number of individuals commonly selected with other approaches. In contrast, of the traditional genomic prediction methods GBLUP showed the lowest number of commonly selected animals compared with the other genomic prediction models. This demonstrates that, even moderately, there is a disagreement in the genomic prediction results between the DL methods (which were closer to each other, MLP × CNN) compared with the more traditional methods. Similarly, Han et al. (2021) reported that the DL methods have demonstrated similarity among them in the potential of genetic prediction compared with more traditional approaches. It should be emphasized that these results, however, are not sufficient to determine that one method is superior to another in terms of genomic predictive ability. However, the results reveal some divergence between the approaches, which will ultimately affect the genetic gain when choosing one or another genomic prediction model, highlighting the importance of validating DL methods for genomic predictions before commercial implementation.

Figure 3. Scatter plots of predicted versus observed breeding values for Bayesian LASSO, MLP, CNN, and GBLUP for milking refusals in North American Holstein cattle in automatic milking systems.
The predictive performance of DL models, as observed through the process of retraining, revealed notable variations, suggesting overfitted predictions, which might affect selection (Abdollahi-Arpanahi et al., 2020; Han et al., 2021). In this context, genomic predictions based on DL models are dependent on 2 primary factors. First, DL models are initialized with random weights at the starting point and, second, the sample size for training can play a crucial role in the final values obtained (Bellot et al., 2018). Because of this, DL methods may be overestimated and incorrectly select the best animals, which might cause incongruity when comparing them with the traditional prediction methods, which are less prone to overfitting (Pérez-Enciso and Zingaretti, 2019). In addition, livestock traits such as behavior, well-being, and efficiency measures, among others that are difficult or costly to measure, may have their predictions impaired due to the smaller amount of information available for dense training of models based on DL.

Figure 4. Scatter plots of commonly selected animals for milking refusals (MREF) for (A) Bayesian LASSO versus MLP, (B) Bayesian LASSO versus CNN, (C) Bayesian LASSO versus GBLUP, (D) MLP versus CNN, (E) MLP versus GBLUP, and (F) CNN versus GBLUP. NOT_S = nonselected for both methods, S_BOTH = selected for both methods, S_CNN = selected only for CNN method; S_MLP = selected only for MLP method; S_GBLUP = selected only for GBLUP method; S_LASSO = selected only for Bayesian LASSO method.
**Genetic Dispersion and Selection Characterization for MFAIL**

The scatter plots for MFAIL are shown in Figure 5. Similar to MREF, it can be observed that the LASSO method presented a lower correlation between predicted and observed values compared with other methods. The MLP approach delivered the higher association between predicted and observed values, followed by GBLUP and CNN, respectively. Again, those 3 latter methods established a quite similar pattern in terms of distribution along observed and predicted estimates, indicating a greater central clustering of inference points. Bayesian models, such as LASSO, use more flexible prior distributions characterized by fatter tails, emphasizing a higher probability allocation to effects compared with thinner tails. By leveraging observed trait data, these models derive posterior distributions of marker effects, facilitating predictive modeling (de los Campos et al., 2013; Holland et al., 2020). This approach demonstrates enhanced versatility compared with GBLUP, which assumes a uniform variance across all markers, particularly when confronted with traits containing markers with larger effects due to major genes (Clark et al., 2011). In essence, these models not only accommodate larger effects but also do so by employing priors with wider tails without imposing overly strong shrinkage, unlike the normal distribution. In another study, using several crop datasets, GBLUP and DL methods yielded higher predictive ability compared with LASSO (Ubbens et al., 2021). The authors mentioned that although the predictive ability of GBLUP and DL methods was similar, the shape...
of the scatter plots was different, as seen in our results (Figures 3 and 5), causing variations between predicted and observed values.

For MFAIL, within the ML methods evaluated, the genomic prediction accuracy of MLP was 0.021 higher than that of CNN, which demonstrates that for MFAIL the difference in the prediction ability between the DL methods was higher than for MREF. Comparing MLP with GBLUP the prediction variation decreased by 0.01, indicating a closer prediction accuracy compared with MLP and CNN. However, when MLP was compared with LASSO the prediction variation increased as much as 0.043, demonstrating a higher superiority of the prediction ability of MLP compared with the LASSO method. Finally, for MFAIL, the prediction abilities between the DL and GBLUP methods proved to be similar but highly superior to the LASSO method. Again, because the accuracies based on DL and GBLUP methods were essentially similar, other factors should be considered when choosing the genomic prediction method to be implemented in practice. Computational efficiency and the predictive capacity of nongenotyped animals (Gar-rick and Fernando, 2022), which can be contemplated in a single-step GBLUP approach, may be integrated into genomic predictions as a priority, considering the coverage and efficiency of the results obtained.

Figure 6 is a Cartesian selection graph for MFAIL indicating the top 20% of animals selected by both methods, as well as the selected individuals by only one of the methods and not selected by any of the methods evaluated 2 × 2. The methods that presented the highest number of commonly selected individuals for MFAIL were MLP × CNN (60%), followed by LASSO × CNN (54%), LASSO × MLP (45%), LASSO × GBLUP (37%), CNN × GBLUP (32%), and MLP × GBLUP (32%). Once again, the disagreement was evident in the prediction ability between the DL methods compared with the more traditional methods, especially DL versus GBLUP. This fact further highlights that even though the DL methods and GBLUP have shown similar genomic prediction accuracy, this is not reflected in the number of individuals commonly selected between both groups. In this context, one must be careful when opting for prediction via ML because the recommendation of superior individuals can differ substantially in relation to the more traditional methods that have proven to be effective. Selection errors, whether for replacing females or for choosing males to be used for breeding, can significantly influence genetic gain over subsequent generations (Weigel et al., 2015; Moreira et al., 2019; Labroo and Rutkoski, 2022). The selection divergence observed in the present work between the different methods studied must have been accentuated by the low heritability of MFAIL, which diminishes the predictive ability due to the low genetic variability. Numerous studies have demonstrated that the efficacy of genetic prediction in connecting phenotypic to genomic varia-
tions is inherently tied to the genetic architecture of those traits (Meuwissen et al., 2001; Momen et al., 2018; Rice and Lipka, 2019). As a result, prediction methods have been designed to align with specific genetic architectures by adjusting preconceived notions about the variance and impact of genomic markers (Sun et al., 2014; Galli et al., 2020). Therefore, considering the complexity of the traits evaluated, as well as their heritability, not only serves as a critical practice in breeding but also provides valuable insights into the optimization and application of genomic prediction methods.

**General Discussion**

The assessment of animal behavior, including emotion-
al states, may be highly subjective and often represented by multiple interacting factors (Broom et al., 1993; Miller et al., 2021). For this reason, it is difficult to identify a specific pattern of animal behavior, which results in high variation in the measured performance and low heritability estimates due to the difficulty in capturing the genetic component over the total variation of the behavior indicator traits (Pedrosa et al., 2023). In terms of statistical modeling, DL focuses on determining a prediction without essentially assuming a functional distribution of the data trained, seeking an optimal predictive performance (Valletta et al., 2017). In other words, DL prediction methods will seek the maximum predictive efficiency in numerical terms, which can make them less efficient when the evaluated traits are difficult to precisely record. In these cases, there is even more dependence on a large volume of data and replications to increase the prediction effectiveness, which, in turn, makes the process too slow and, sometimes, computationally inefficient. In addition, deep NN can capture complex effects, including nonadditive effects, which may provide advantages for genomic prediction, especially for management purposes such as precision livestock management. However, the ability of DL methods to better predict “genetic merit” based on pre-adjusted phenotypes, and their advantage for breeding purposes, remains unclear.

**Potential Implications and Limitations**

The results obtained demonstrated that the DL models could potentially predict the genetic merit of individual animals with moderate accuracies and enhance animal selection, demonstrating the potential implications for the use of artificial intelligence tools in livestock breeding programs. The ML-based methods demonstrated
slightly higher prediction accuracies than traditional methods, suggesting that DL algorithms can be tested to be incorporated in animal breeding programs if validated in different populations and traits. However, our study showed some divergence between DL and traditional methods under a selection perspective, which may ultimately affect genetic gain when choosing one or another genomic prediction method. In addition, the accuracy of prediction may be influenced by the heritability of the evaluated traits. In this context, one limitation of this investigation is that the evaluated traits had low heritability estimates, which may have influenced the magnitude of the prediction accuracies obtained. Higher prediction accuracies might be obtained when evaluating traits with moderate-to-high heritability. It is also unclear whether ML methods would provide even greater accuracies for traits with more robust genetic components. In the end, when choosing the best genomic prediction method, factors other than only prediction accuracies should be considered, such as computational demands and the potential

Figure 6. Scatter plots of commonly selected animals for milking failures for (A) Bayesian LASSO versus MLP, (B) Bayesian LASSO versus CNN, (C) Bayesian LASSO versus GBLUP, (D) MLP versus CNN, (E) MLP versus GBLUP, and (F) CNN versus GBLUP. NOT_S = nonselected for both methods, S_BOTH = selected for both methods, S_CNN = selected only for CNN method; S_GBLUP = selected only for GBLUP method; S_MLP = selected only for MLP method; S_LASSO = selected only for Bayesian LASSO method.
of using the information of nongenotyped animals, which can be a bottleneck in implementing DL methods, especially when dealing with large datasets.

**CONCLUSIONS**

This study provides insights into the potential of DL methods to enhance genomic prediction accuracy for complex behavioral traits in dairy cattle. The DL methods presented slightly higher genomic prediction accuracies compared with traditional genomic methods. In addition, the MLP and CNN methods had lower reranking of selection than other approaches for both MREF and MFAIL. However, the results reveal some selection divergence between the DL and traditional genomic prediction approaches, which will affect the genetic gain, potentially resulting in less progress across generations. The importance of validation of DL methods for genetic prediction before possible commercial use is evident. Additional factors, such as high computational demand and the difficulty of performing genomic prediction for nongenotyped animals, must be considered, which may limit the practical applicability of DL methods in animal breeding programs.

**NOTES**

This research was funded by the Agriculture and Food Research Initiative Competitive Grant number 2022-67021-37022 from the USDA National Institute of Food and Agriculture (Washington, DC). We are very grateful for the support received from the Purdue College of Agriculture and Purdue Ag Data Services (West Lafayette, IN) for supporting the development of the Purdue Animal Sciences Data Ecosystem (PASDE), which is a platform for automatically gathering, integrating, and storing large-scale datasets from precision dairy farms to be used for this type of research. The authors also thank Homestead Dairy LLC (Plymouth, IN) and especially Brian and Jill Houin for inviting us on their farm and for allowing our research group to use their farm records for research purposes. We also thank Lely (Maassluis, the Netherlands) for giving us access to their system for extracting the datasets. Animal Care Committee approval was not necessary for the purposes of this study because all the information required was obtained from pre-existing databases. The authors have not stated any conflicts of interest.

**Abbreviations used:** AMS = automated milking systems; CNN = convolutional neural network; DL = deep learning; LASSO = Bayesian LASSO; MFAIL = milking failure; ML = machine learning; MLP = multiple layer perceptron; MREF = milking refusal; MSE = mean squared error of prediction; NN = neural network; NOT_S = nonselected for both methods; QC = quality control; S_BOTH = selected for both methods; S_CNN = selected only for CNN method; S_GBLUP = selected only for GBLUP method; S_MLP = selected only for MLP method; S_LASSO = selected only for LASSO method.

**REFERENCES**


**ORCIDS**

Victor B. Pedrosa @ https://orcid.org/0000-0001-8966-2227

Shi-Yi Chen @ https://orcid.org/0000-0002-3955-0101

Leonardo S. Gloria @ https://orcid.org/0000-0002-2756-5939

Jarrod S. Doucette @ https://orcid.org/0000-0003-4027-2417

Jacquelyn P. Boerman @ https://orcid.org/0000-0002-0336-8295

Guilherme J. M. Rosa @ https://orcid.org/0000-0001-9172-6461

Luiz F. Brito @ https://orcid.org/0000-0002-5819-0922