frontiers in Genetics

ORIGINAL RESEARCH
published: 17 December 2021
doi: 10.3389/fgene.2021.798840

A New Deep Learning Calibration Method Enhances Genome-Based Prediction of Continuous Crop Traits

Osval A. Montesinos-López[1], Abelardo Montesinos-López[2], Brandon A. Mosqueda-González[3], Alison R. Bentley[4], Morten Lillemo[5], Rajeev K. Varshney[6,7]* and José Crossa[4,8]**

[1]Facultad de Telemática, Universidad de Colima, Colima, Mexico, [2]Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Guadalajara, Mexico, [3]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Esq. Miguel Othón de Mendizábal, Mexico city, Mexico, [4]International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [5]Department of Plant Sciences, Norwegian University of Life Sciences, IHA/CIGENE, As, Norway, [6]Centre of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India, [7]State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Murdoch University, Perth, WA, Australia, [8]Colegio de Postgraduados, Montecillo, Mexico

OPEN ACCESS

Edited by:
Pasquale Tripodi, Council for Agricultural and Economics Research (CREA), Italy

Reviewed by:
Moyses Nascimento, Universidade Federal de Viçosa, Brazil
Shogo Tsuruta, University of Georgia, United States

***Correspondence:**
Abelardo Montesinos-López
aml_uach2004@hotmail.com
Rajeev K. Varshney
rajeev.varshney@murdoch.edu.au
José Crossa
j.crossa@cgiar.org

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *20 October 2021*
Accepted: *18 November 2021*
Published: *17 December 2021*

Citation:
Montesinos-López OA, Montesinos-López A, Mosqueda-González BA, Bentley AR, Lillemo M, Varshney RK and Crossa J (2021) A New Deep Learning Calibration Method Enhances Genome-Based Prediction of Continuous Crop Traits. Front. Genet. 12:798840. doi: 10.3389/fgene.2021.798840

Genomic selection (GS) has the potential to revolutionize predictive plant breeding. A reference population is phenotyped and genotyped to train a statistical model that is used to perform genome-enabled predictions of new individuals that were only genotyped. In this vein, deep neural networks, are a type of machine learning model and have been widely adopted for use in GS studies, as they are not parametric methods, making them more adept at capturing nonlinear patterns. However, the training process for deep neural networks is very challenging due to the numerous hyper-parameters that need to be tuned, especially when imperfect tuning can result in biased predictions. In this paper we propose a simple method for calibrating (adjusting) the prediction of continuous response variables resulting from deep learning applications. We evaluated the proposed deep learning calibration method (DL_M2) using four crop breeding data sets and its performance was compared with the standard deep learning method (DL_M1), as well as the standard genomic Best Linear Unbiased Predictor (GBLUP). While the GBLUP was the most accurate model overall, the proposed deep learning calibration method (DL_M2) helped increase the genome-enabled prediction performance in all data sets when compared with the traditional DL method (DL_M1). Taken together, we provide evidence for extending the use of the proposed calibration method to evaluate its potential and consistency for predicting performance in the context of GS applied to plant breeding.

Keywords: genomic selection, genomic prediction, calibration of predictions, deep learning, GBLUP, plant breeding

INTRODUCTION

Genomic selection (GS) exploits dense genome-wide markers for predicting complex traits. Practically it requires development of a training population (with phenotypic and genotypic information) with which a statistical machine learning algorithm is trained and used for making predictions for individuals of a test breeding population with only genotypic information. Genome-enabled prediction and GS were originally proposed by Meuwissen et al. (2001) as a novel approach

for predicting complex traits for a selection of candidates using predicted phenotypic or breeding values. Zhong et al. (2009) and Heffner et al. (2010) state that GS works given realistic assumptions of selection accuracies, breeding cycle times and selection intensities. In simple terms, GS offers tremendous opportunities to improve rates of genetic gain in plant and animal breeding, and it has been supported by many research articles published in the last 20 years (Bhat et al., 2016; Crossa et al., 2017).

GS is changing the landscape of practical plant breeding, as it is able to predict breeding values earlier and with greater accuracy when compared with conventional selection methods such as mixed models, Ridge regression and Bayesian methods (BayesA, BayesB, BayesC, Bayesian Lasso, etc). Additionally, time is saved by using GS because it is no longer necessary to wait for late filial generations to phenotype complex quantitative traits such as yield, biotic and abiotic stresses, among others. The genotypic data can be obtained from the seed of early generations and used to predict phenotypic performance of later generation individuals without the need for extensive phenotyping evaluation over years and environments (Mellers et al., 2020). Furthermore, it highlights the potential to increase the speed of varietal development across crop species (Bhat et al., 2016; Crossa et al., 2017).

Estimating the genetic worth of the individual in GS is based on a large set of marker information distributed across the whole genome, which contrasts with the relatively few markers used in marker assisted selection (MAS) (Varshney R. K. et al., 2021). Conventional breeding involves hybridization between diverse parents and subsequent selection over a number of generations to develop improved crop varieties. This has several limitations, including the long duration (5–12 years) required to develop a crop variety, the reliance on time-consuming (and traditionally low-throughput) phenotypic selection, high environmental noise and genotype × environment interactions. It is also less effective for complex and low heritability traits (Tuberosa, 2012). For these reasons, several studies have shown GS models to be advantageous for complex quantitative traits like grain yield, quality, biotic and abiotic stresses, etc. (de los Campos et al., 2009; Crossa et al., 2010; Burgueño et al., 2012; González-Camacho et al., 2012; Jannink et al., 2010).

However, there are still numerous opportunities to improve the selection process of candidate individuals in GS. Some of these are: 1) to improve the quality and coverage of marker data; 2) to design optimal training-testing sets; 3) to better identify where in the breeding program GS could be efficiently applied (Crossa et al., 2017); 4) to have sufficient numbers of individuals in the reference (training) population; and 5) to use the most appropriate statistical machine learning model for each data set at hand.

Intensive research has explored different statistical machine learning methods for GS (Varshney R. K. et al., 2021). For example, some of the models/methods used in GS are: 1) linear mixed models and their Bayesian counterpart that includes the so-called Bayesian alphabet [BayesA, BayesB, BayesC, Genomic Best Linear Unbiased Predictor (GBLUP), and Bayesian Lasso]; 2) Random forest for predicting binary, categorical and continuous traits (Montesinos-López et al., 2021a); 3) support vector machine (Montesinos-López et al., 2019a); 4) gradient boosting machine and 5) deep learning algorithms (Montesinos-López A. et al., 2018; Montesinos-López, 2018b; Montesinos-López et al., 2019a; Montesinos-López et al., 2021c). These statistical machine learning methods have been adopted for GS because they can help improve genome-enabled prediction accuracy as they use machine learning advances for analysis, interpretation, prediction and decision-making. One explanation of why many statistical machine learning methods have been implemented in GS is the fact that there is no universal best prediction model that can be used under all circumstances (No free lunch theorem; Wolpert, 1996).

Deep learning (DL) methods are one of the most recent adoptions of statistical machine learning methods used for GS (Varshney RK. et al., 2021). There is mounting evidence suggesting that these methods outperform conventional methods in terms of predictive power, as well as other advantages (Montesinos-López et al., 2021c). Some of these advantages are: 1) power in capturing complex patterns in the data caused by the inclusion of many neurons communicated in complex ways and via multiple nonlinear transformations through hidden layers (Montesinos-López et al., 2019b; Montesinos-López et al., 2021c); 2) support for raw (not preprocessed) inputs, which is impossible with most statistical machine-learning methods (Montesinos-López et al., 2021c); 3) support for a variety of different inputs that can accommodate pedigree, genomic, environmental and other forms of omics data (e.g., metabolomics, microbiomics, phenomics, proteomics, transcriptomics, etc.) (Montesinos-López et al., 2021c); 4) greater efficiency for handling large and complex data sets compared with most statistical machine-learning methods (Montesinos-López et al., 2021a,c); and 5) a very flexible network architecture permitting a "Lego-like" construction of new models, while an unlimited number of neural network models can be constructed using elements of the core architectural building blocks of existing DL models (Montesinos-López et al., 2021a, c).

While DL methods offer many advantages, their training process is very challenging, especially considering hyper-parameter selection. The correct selection of hyper-parameters are time consuming and complicated to implement largely due to the absence of a unique and efficient optimized methodology. This means that the implementation of DL methods for genome-enabled prediction is not straightforward. Furthermore, DL methods are inefficient when used with small data sets or with simple linear patterns, and as such, research is underway to facilitate the training process of DL methods so that they can be used in these contexts.

One way of improving the training process of DL models is to use a calibration based on a model that is already trained and applied via a post-processing operation. However, in the context of machine learning methods (including DL), calibration methods have only been proposed for binary and categorical response variables, where the predicted probabilities that do not match the expected distribution of the observed probabilities of

the response variable in the data are adjusted (calibrated) to increase this match and the prediction accuracy in the testing set. This is very important in classification problems because the estimated class probabilities reflect the true underlying probability of the sample. For this reason, the predicted class needs to be well-calibrated, which means the probabilities must effectively reflect the true likelihood of the event of interest. This discrepancy between the distribution of observed and predicted values is also very commonly found in continuous response variables, and yet no solution has been proposed. In other words, this means that despite all efforts taken during the training process of the deep neural network, often times, there will still be bias in the predictions. Consequently, methods for calibrating continuous and categorical response variables are of paramount importance to increase the accuracy of your prediction machine.

Based on the previous considerations, the main objective of this study is to present a method that facilitates the calibration of DL outputs in the context of genomic-based prediction in GS. In this vein, we propose a calibration method for continuous response variables that significantly improves the training process for DL methods. We used four existing data sets to compare the prediction accuracy in terms of Mean Squared Error Prediction (MSE) of the popular Genomic Best Linear Unbiased Predictor (GBLUP), for the standard DL method (DL_M1) and the new proposed calibration method (DL_M2). GBLUP was used for comparison, as it is the most used model in genome-enabled prediction. The genome-enabled prediction models and methods were also compared in the absence and presence of genotype × environment interaction.

MATERIALS AND METHODS

Data sets used in previous studies were employed here for assessing the performance of the new calibration methods applied to DL.

Dataset 1. Maize Grain Yield Prediction

As previously reported by Montesinos-López et al., 2016 this dataset consists of a sample of 309 maize lines evaluated for three traits: anthesis-silking interval, plant height and grain yield (GY). Each trait was evaluated in three optimal environments (denoted Env1, Env2 and Env3). It is important to point out that each line was evaluated once in each environment, and as such, each lines has three replications. Additionally, it should be highlighted that we have genotypic information for the 309 lines and phenotypic information for the 927 (309×3) observations, which were all collected in the same year. The field design in each of the three environments was a lattice incomplete block design with two replications. Data were pre-adjusted using estimates of block and environmental effects derived from a linear model that accounted for the incomplete block design within environments and for environmental effects. The lines were genotyped with 681,257 single nucleotide polymorphisms (SNPs). Markers with more than 20% missing values and with minimum allele frequency (MAF) of 0.05 were removed. The remaining missing markers were imputed using observed allelic frequencies resulting in 158,281 SNPs available for further analyses. In the present study, we compared genome-enabled prediction performance for GY.

Dataset 2. Groundnut Seed Yield per Plant (SYPP) Prediction

The phenotypic dataset reported by Pandey et al. (2020) contains information on the phenotypic performance for various traits in four environments. In the present study we assessed predictions using the trait seed yield per plant (SYPP) for 318 lines in four environments denotes as Environment1 (ENV1): Aliyarnagar_Rainy 2015; Environment2 (ENV2):Jalgoan_Rainy 2015; Environment3 (ENV3):ICRISAT_Rainy 2015; Environment4 (ENV4):ICRISAT Post-Rainy 2015.

The dataset is balanced, giving a total of 1272 (318×4) assessments (phenotypic values) with each line included once in each environment (four replications of each line), which were all measured in the same year. Marker data were available for all lines and 8,268 SNP markers remained after quality control (each marker was coded with 0, 1 and 2); however, the makers were obtained only for the 318 lines.

Dataset 3. Chickpea Biomass Prediction

The phenotypic dataset reported by Roorkiwal et al. (2018) contains information for 315 lines evaluated in six environments (denoted as 1, 2, 4, 5, 6, 7) for biomass. The dataset is balanced with all lines assessed in all environments (that is, four replications of each line), giving a complete phenotypic dataset with $315 \times 4 = 1890$ observations. Marker data were available for all 315 lines, with 35,527 SNP markers available following quality control, where each marker was coded with 0, 1 and 2, and all information was collected in the same year.

Dataset 4. Spring Wheat Grain Yield Prediction

Spring wheat data was available from the Global Wheat Program (GWP) at the International Maize and Wheat Improvement Center (CIMMYT) from elite yield trials (EYT) evaluated in four selection environments (denoted Bed5IR, EHT, Flat5IR, FlatDrip). The dataset included the performance data from the 2016-2017 cycle from a total of 980 lines assessed in the four environments, giving 3920 (980×4) observations since each line was repeated four times, once in each environment. The experimental design was an alpha-lattice with the lines sown in 39 sets, each including 28 lines and two checks in six blocks with three replications. Four performance traits were assessed: days to heading (number of days from germination to 50% spike emergence); days to maturity (number of days from germination to 50% physiological maturity or the loss of green color in 50% of the spikes); plant height (measured from the ground to the top of the spike, in centimeters); and grain yield (GY). Genome-wide SNP markers were generated for the 980 lines using genotyping-by-sequencing (GBS; Elshire et al., 2011; Poland et al., 2012) at Kansas State University using an Illumina HiSeq2500. After

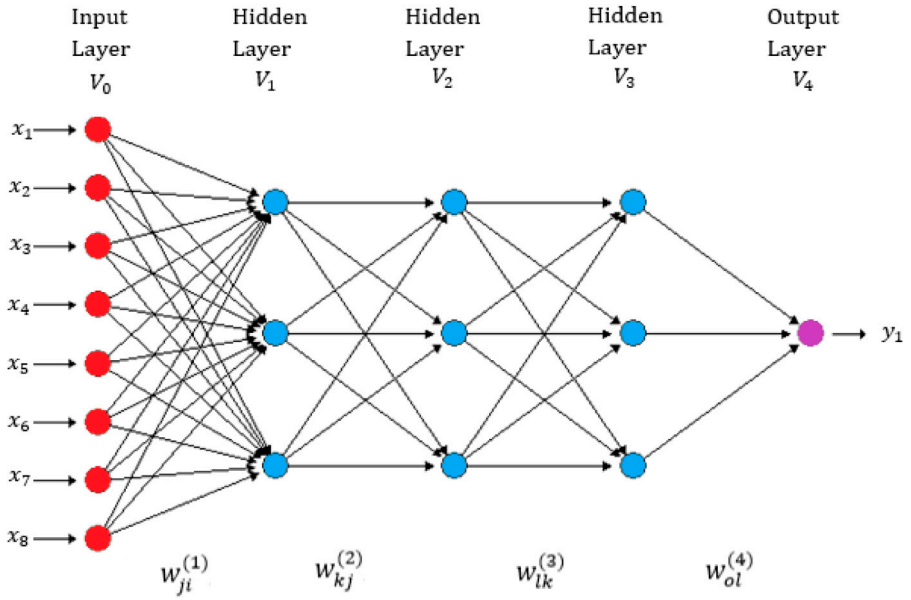

FIGURE 1 | A feedforward deep neural network with one input layer, three hidden layers and one output layer (Montesinos-López, et al., 2021d). There are eight neurons in the input layer that correspond to the input information and four neurons in each of three hidden layers, with only one neuron in the output layer that corresponds to the count trait that will be predicted (Intercepts = biases not shown in this figure).

filtering, 2,038 markers remained. Imputation of missing marker data was done using LinkImpute (Money et al., 2015) implemented in TASSEL V5 (Bradbury et al., 2007). In the current study we assessed predictions using GY.

GBLUP Model

The model assumed for the response variable was

$$Y_{ij} = \mu + Loc_i + g_j + gL_{ij} + \varepsilon_{ij} \quad (1.1)$$

where Loc_i are the fixed effects of locations, g_j, $j = 1, \ldots, J$, are the random effects of lines, gL_{ij} are the random effects of location-line interaction, and ε_{ij} are random error components assumed to be independent normal random variables with mean 0 and variance σ^2. Furthermore, it is assumed that $\mathbf{g} = (g_1, \ldots, g_J)^T \sim N_J(\mathbf{0}, \sigma_{\mathbf{g}}^2\mathbf{G})$, $\mathbf{gL} = (gL_{11}, \ldots, gL_{1J}, \ldots, gL_{IJ})^T \sim N_{IJ}(\mathbf{0}, \sigma_{\mathbf{gL}}^2(\mathbf{I} \otimes \mathbf{G}))$, where $\mathbf{G}$ is the genomic relationship-matrix as computed by VanRaden (2008) and $\otimes$ denotes the Kronecker product. The implementation of this model was done in the BGLR library of Pérez and de los Campos, 2014.

Conventional Deep Learning (DL_M1)

We implemented the most popular deep neural network architecture called densely connected networks (multilayer perceptron) (Chollet and Allaire, 2017). This network does not assume a specific structure in the input features. In general, the basic structure of a densely connected network consists of an input layer, one output layer (for uni-trait modeling) and multiple hidden layers between both layers. This type of neural network is also known as a feedforward deep neural network (See **Figure 1**). The implementation of this deep neural network is challenging because it requires many hyper-parameters, like number of units, number of layers, number of epochs, type of regularization method and type of activation function. Based on available literature, we used the rectified linear activation unit (ReLU) as the activation function in the hidden layers, the linear activation function in the output layer and the dropout type of regularization method for training the models (Chollet and Allaire, 2017).

The dataset was divided into training (80%) and testing (20%). Then each training set was divided into *inner-training* ($80 \times 0.8 = 64\%$) and *validation set* ($80 \times 0.2 = 16\%$). With the *inner-training,* we trained the 8 resulting models (grid of eight values) by combining the following hyper-parameters: two neurons ($1.5 \times$ Number of independent variables of each dataset, $3 \times$ Number of independent variables of each dataset), two values of hidden layers (with 1 and 4), two values of dropout (0.15 and 0.3), one learning rate equal to 0.001, and one value of epoch that was fixed at 1,000. From these eight combinations, we selected the best hyper-parameter combination in terms of prediction performance (with mean square error or prediction (MSE) in the *validation* set. Then, with the best hyper-parameter combination obtained from the *validation* set, a model was refitted with the whole information of the *training* (*inner-training* + *validation*) *set.* Then with this refitted model, predictions of the corresponding testing set were made. Finally, the average of the five folds in terms of MSE was reported as prediction performance of the conventional deep learning method (DL_M1). This model was evaluated with (GE) and without (NO GE) the genotype × environment interaction. When the GE was taken into consideration, the predictor

contained the design matrix of environments (X_E), genotypes (X_G; this matrix contains the raw design matrix of genotypes post multiplied by the Cholesky decomposition of the genomic relationship matrix) and the design matrix of the GE interaction (X_{GE}; this matrix was built by combining matrices X_E and X_G), that is, the predictor contained the following concatenated information: predictor=(X_E, X_G, X_{GE}). Conversely, when the GE was ignored (NO GE), the predictor only took into account the design matrices of X_E and X_G, that is, predictor=(X_E, X_G). However, it is important to point out that because deep neural networks apply more than one hidden layer, with many units (neurons) and with nonlinear transformations (activation functions) without explicitly giving the interaction term, X_{GE}, they can capture complex interactions due to the way neurons interact with each other (See **Figure 1**).

Calibration Method for Outputs of Deep Learning (DL_M2)

Next, we implemented the proposed new calibration method (DL-M2) for continuous outcomes. This involved eight steps, as follows. First, the data was divided into training (80%) and testing (20%) sets, as previously mentioned. Then the training data was divided into 1) an *inner-training* ($80 \times 0.8 = 64\%$) set and 2) a *validation* ($80 \times 0.2 = 16\%$) set. Following this, the *inner-training* was divided into *inner-inner-training* ($64 \times 0.8 = 51.2\%$) and *inner-validation* ($64 \times 0.2 = 12.8\%$). With the *inner-inner-training* we trained the 8 resulting models (grid of eight values) of combining the following hyper-parameters: two neurons (1.5 × Number of independent variables of each data set, 3 × Number of independent variables of each data set), two values of hidden layers (with 1 and 4), two values of dropout (0.15 and 0.3), one learning rate equal to 0.001, and one value of epoch that was fixed at 1,000.

From these eight hyper-parameter combinations, we selected the best in terms of prediction performance in the *inner-validation* set. Then the best hyper-parameter combination obtained from the *inner-validation* set was refitted to a model with the information of the *inner-training* (*inner-inner-training* + *inner-validation*). We then used the fitted model with the *inner-training* set to make predictions of the *validation* data set and for the testing data set. A linear model was then adjusted using the observed response variable of the *validation* set as the response variable and the predicted values (of the validation set) obtained in the fitting of the inner-training set for the *validation* data set as the independent variable. This step fits a linear model of the *observed validation* data with the *predicted validation* data previously obtained. Finally, we used this fitted linear model for making adjusted predictions of the testing set using only the predicted values of the testing set as input. This step provides adjusted predictions (calibrated) of the testing set and are the final predictions which are calibrated in this step. The steps in this process were repeated for each training-testing partition. In this case, there were five folds, and the average MSE of the five folds was reported as the prediction performance.

Applying a Cross-Validation Strategy

To evaluate the predictive performance, we used a 5-fold cross-validation, with four folds used for training and one for testing. The average mean square error (MSE), was computed with the five folds, which was used to assess prediction performance in each data set under study. For deep learning models (DL_M1 and DL_M2), a 5-fold cross-validation was also implemented to select the best combination of hyper-parameters. For the conventional deep learning model (DL_M1), the 5-fold cross-validation was implemented with a training set that was divided into inner-training and validation, while for the proposed calibration method (DL_M2), the 5-fold cross-validation was implemented with the inner-training set that was divided into inner-inner-training and inner-validation. This strategy of cross validation mimics real applications where some lines are missing in some environment's, but are present in at least another environment. This means that our approach does not mimic scenarios where we use previous generation to predict next generations as training. On the other hand, as pointed out by one reviewer, other metrics can be used for evaluating the prediction accuracy like the Person´s correlation, even though in this application only the MSE was used. Furthermore, in this case since we have available phenotypes (because a 5-fold cross-validation approach was used), it was not necessary to compare predictions with parent averages from early generations.

RESULTS

Dataset 1 (Maize Data set)

Analysis of Dataset 1 showed that the best prediction performance when including genotype × environment (GE) interaction was observed using GBLUP (**Figure 2**). Across the sites, the best predictions were observed under environment KAK while the worst were observed under environment KTI. When comparing the conventional deep learning method (DL_M1) with our proposed method for calibrating deep learning models (DL_M2), we observed that the proposed calibration method improved the genome-enabled prediction performance based on MSE. Ignoring the GE interaction term in environments EBU, KAK and KTI, DL_M2 reduced the MSE with regard to DL_M1 by 17.188, 4.273 and 19.469%, respectively. However, the DL_M2 prediction performance in terms of MSE was worse than the GBLUP method, which outperformed DL_M2 at all sites by 16.179% (EBU), 18.298% (KAK) and 12.370% (KTI).

When the GE interaction was taken into account, the DL_M2 method reduced the MSE compared to DL_M1 by 8.354, 1.487 and 16.258% in environments EBU, KAK and KTI, respectively. However, although the improvement of DL_M2 over DL_1 is significant, the GBLUP method still outperformed DL_M2 method by 15.268, 20.753 and 13.848% in environments EBU, KAK and KTI, respectively.

Finally, across environments (**Figure 3**), the best prediction performance was observed from the GBLUP and the worst from the DL_M1 method. GBLUP outperformed DL_M2 by 12.296% and DL_M1 by 37.616%, whilst DL_M2 outperformed DL_M1 by 18.399%. Accounting for the GE interaction, GBLUP

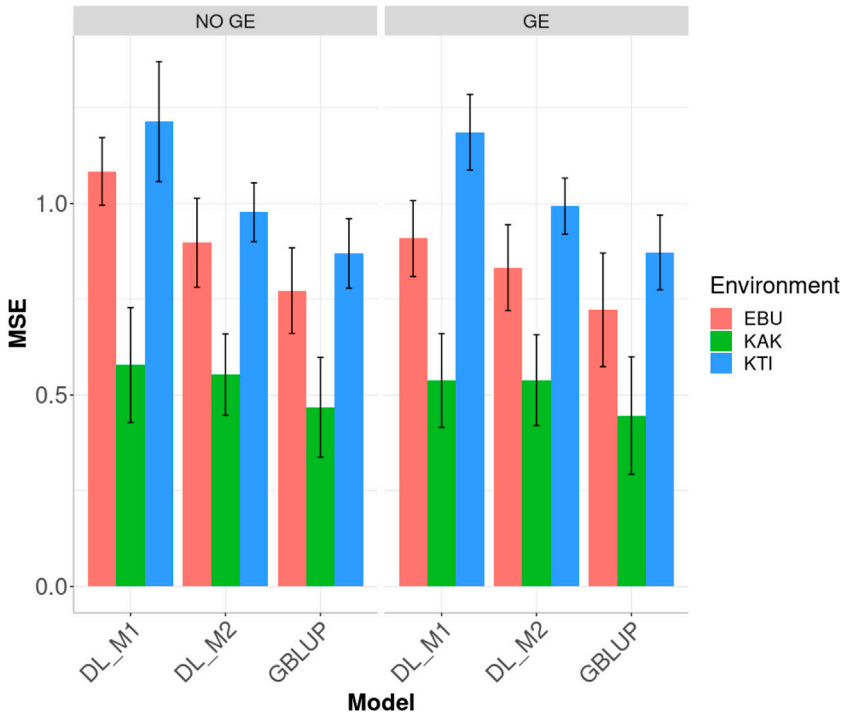

FIGURE 2 | Dataset 1. Prediction performance in terms of mean square error (MSE) for each environment of the maize dataset under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype ×environment interaction. GE means that the model takes the genotype × environment interaction into account.

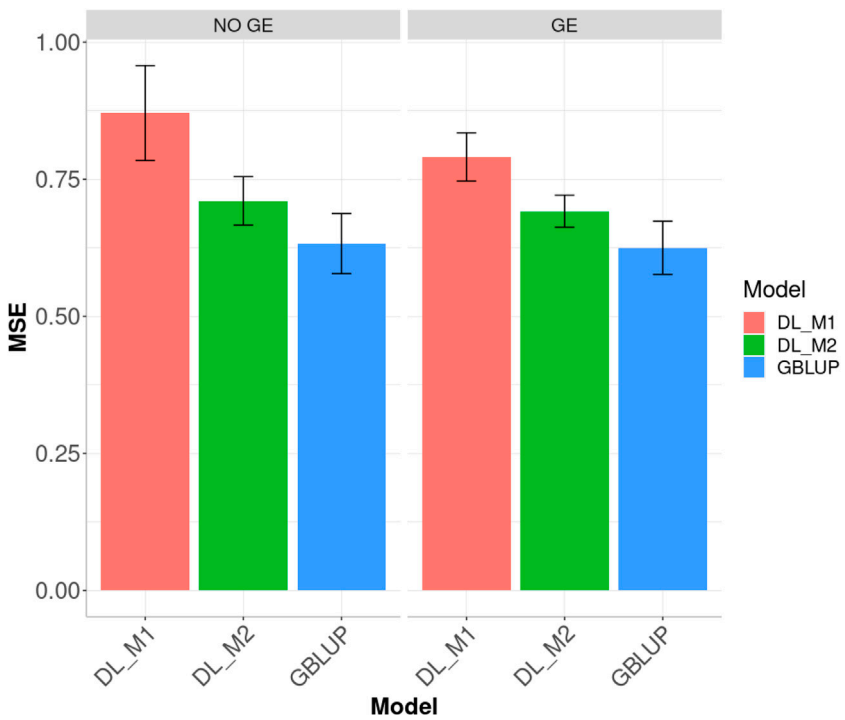

FIGURE 3 | Dataset 1. Prediction performance in terms of mean square error (MSE) across environments for the maize data set under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype ×environment interaction. GE means that the model takes the genotype × environment interaction into account.

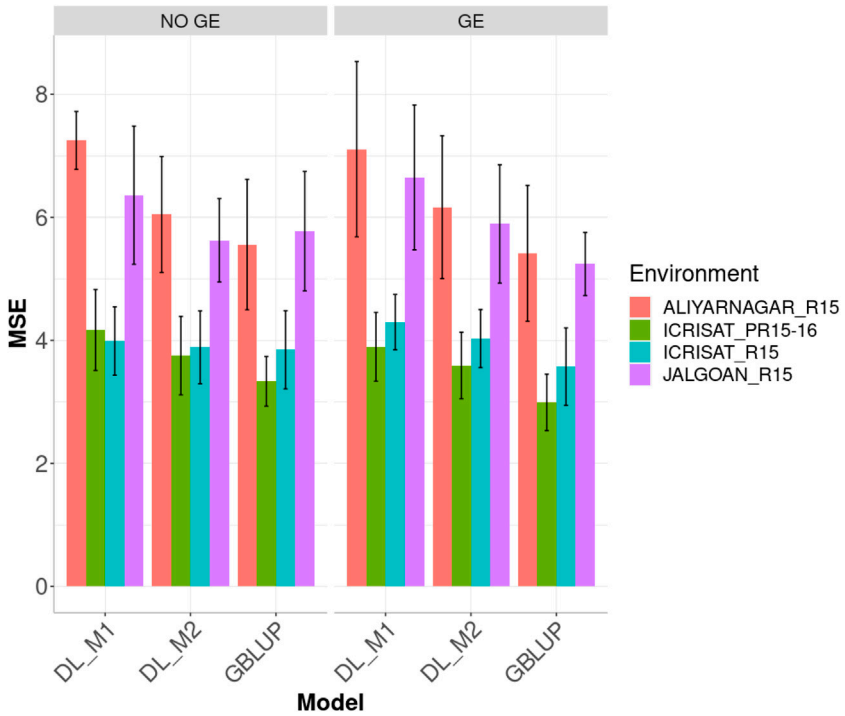

FIGURE 4 | Dataset 2. Prediction performance in terms of mean square error (MSE) for each environment of the groundnut dataset under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype ×environment interaction. GE means that the model takes the genotype × environment interaction into account.

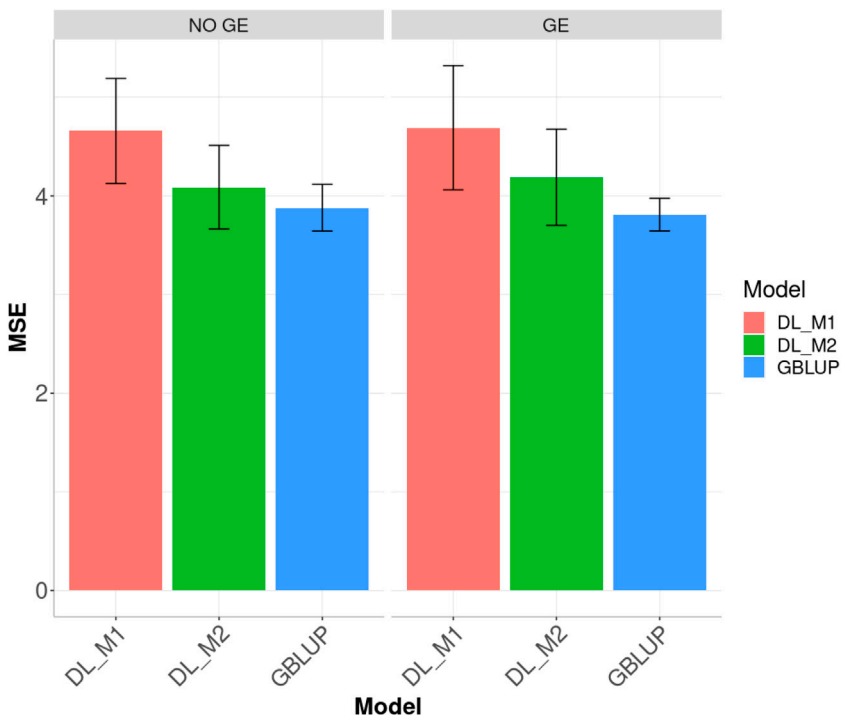

FIGURE 5 | Dataset 2. Prediction performance in terms of mean square error (MSE) across environment of the Groundnut dataset under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype × environment interaction. GE means that the model takes the genotype × environment interaction into account.

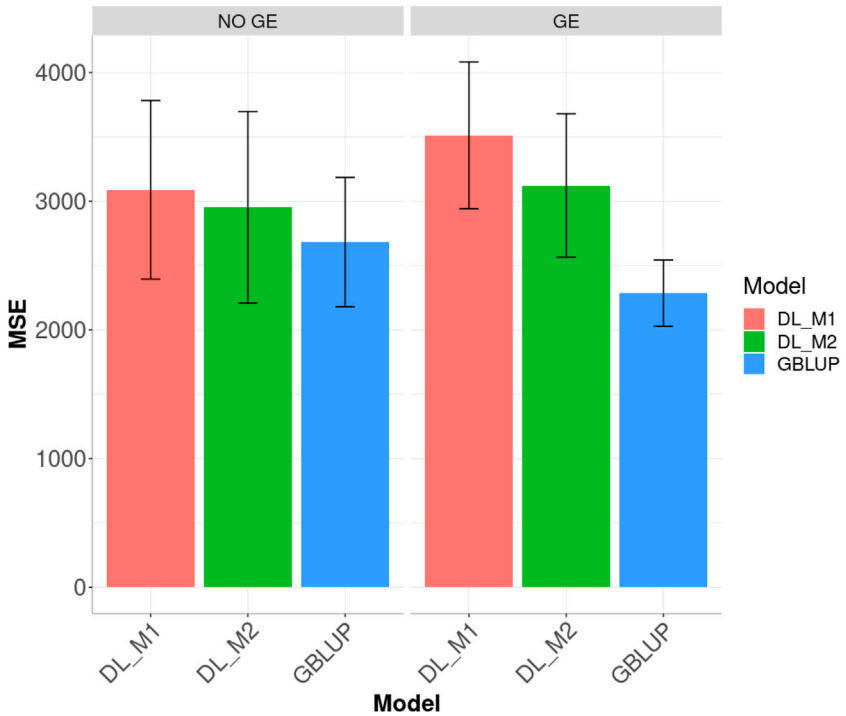

FIGURE 6 | Dataset 3. Prediction performance in terms of mean square error (MSE) across environments for the chickpea dataset under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype × environment interaction. GE means that the model takes the genotype × environment interaction into account.

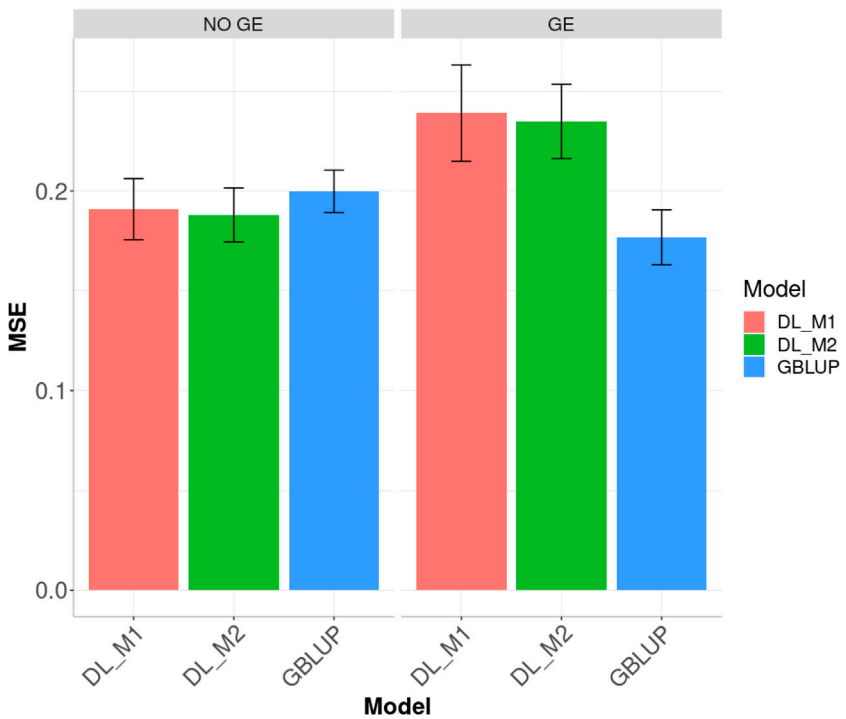

FIGURE 7 | Dataset 4. Prediction performance in terms of mean square error (MSE) across environments for the elite wheat yield trial (EYT) year 2016–2017 dataset under models: DL_M1, DL_M2 and GBLUP. No GE means that the model ignores the genotype × environment interaction. GE means that the model takes the genotype × environment interaction into account.

outperformed both the DL_M1 and DL_M2 by 26.491 and 10.686%, respectively. With GE interaction, DL_M2 reduced the MSE compared with the DL_M1 by 12.495%.

Dataset 2 (Groundnut Dataset)

Figure 4 displays the genome-enabled prediction performance (MSE) including (or not) the GE interaction under the GBLUP, DL_M1 and DL_M2 methods in the four environments. The proposed method of calibration (DL_M2) improved the prediction performance of conventional deep learning method (DL_M1). When ignoring the GE interaction term, the DL_M2 method reduced the MSE compared with DL_M1 by 16.591, 10.004, 2.538 and 11.514% in environments ALIYARNAGAR_R15, ICRISAT_PR15-16, ICRISAT_R15 and JALGOAN_R15, respectively. The GBLUP method outperformed the DL_M2 method in three out of the four environments by 8.80% (ALIYARNAGAR_R15), 12.519% (ICRISAT_PR15-16) and 1.096% (ICRISAT_R15). The worst predictions were observed under environments ALIYARNAGAR_R15 and JALGOAN_R15, while the best were observed under environments ICRISAT_PR15-16 and ICRISAT_R15.

Considering GE interaction, the DL_M2 method improved the prediction performance compared with the DL_M1 method in environments ALIYARNAGAR_R15, ICRISAT_PR15-16, ICRISAT_R15 and JALGOAN_R15 by 13.256, 7.766, 6.238 and 11.368%, respectively. The GBLUP method overcame the DL_M2 method by 13.864, 20.078, 12.763 and 12.423% in environments ALIYARNAGAR_R15, ICRISAT_PR15-16, ICRISAT_R15 and JALGOAN_R15, respectively (**Figure 5**).

When ignoring the GE interaction across environments (**Figure 5**), the worst predictions were observed under the DL_M1 method and the best under the GBLUP method. The GBLUP outperformed the DL_M2 method by 5.535%, while the DL_M2 outperformed the DL_M1 by 12.229%; the GBLUP was also better than the DL_M1 by 20.045%. When including GE interaction, results showed that DL_M1 was the worst method and the GBLUP was the best, but now the GBLUP outperformed the DL_M1 and DL_M2 methods by 23.116 and 9.929%, respectively. However, for this dataset, DL_M2 outperformed the DL_M1 by 10.711%.

Dataset 3 (Chickpea)

Ignoring GE in the across environments case, **Figure 6** indicates that the best predictions were observed under the GBLUP method and the worst under the DL_M1 method. Furthermore, the GBLUP outperformed the DL_M2 method by 10.082%, while the DL_M2 outperformed the DL_M1 by 4.402% and the GBLUP outperformed the DL_M1 by 15.152%. Considering the GE interaction, the GBLUP was the best method and the DL_M1 the worst, where the GBLUP outperformed the DL_M1 and DL_M2 by 53.679 and 36.616%, respectively. Results show that compared with the DL_M1 method, the DL_M2 reduced the MSE by 11.102%.

Dataset 4 [Elite Wheat Yield Trial (EYT) Year 2016–2017]

Results for across environments are displayed in **Figure 7**. When no GE was included, the best predictions were observed under the DL_M2 method and the worst under the GBLUP method; the DL_M2 outperformed the GBLUP method by 5.906%, while the DL_M2 outperformed the DL_M1 by 1.519% and the DL_M1 outperformed the GBLUP by 4.454%. When considering the GE interaction, the GBLUP was the best and the DL_M1 the worst, where GBLUP outperformed the DL_M1 and DL_M2 by 35.18 and 32.805%, respectively. Compared with the DL_ M1, the DL_M2 reduced the MSE by 1.757%.

DISCUSSION

Genomic selection helps save significant resources for the early selection of candidate genotypes because instead of phenotyping and genotyping all the candidate lines, only a sample of them are phenotyped and genotyped. For the remaining individuals that were only genotyped, genome-enabled predictions of the phenotypic values are performed. This means that the accuracy of GS is linked to the quality of the predictions, and the better the predictions, the more accurate the GS methodology. Thus, continuing research to improve the quality of the predictions using GS is of paramount importance. For this reason, this research proposed a simple and novel calibration method to improve the predictions resulting from deep learning methods.

The proposed method was evaluated in four datasets and we found that in three out of the four datasets, the proposed calibration method improved the predictions over conventional deep learning methods. The increase in prediction performance in these four datasets was between 1.519 and 18.39% across environments, which empirically reflects that the proposed calibration method (DL_M2) is quite efficient for improving the prediction power of deep learning models. However, it is important to point out that we did not find that the proposed calibration method outperformed the predictions of the GBLUP method, one of the most popular genomic prediction models. In fact, the GBLUP method outperformed the proposed calibration method (DL_M2) across environments between 5.535 and 36.616% in the four data sets.

However, taking into account the standard errors in most of the scenarios under study no statistical differences were observed between the proposed DL_M2 and the GBLUP method. Two reasonable explanations as to why the GBLUP method outperformed the proposed deep learning method even with the proposed calibration method could be that the four data sets are: 1) small and, as pointed out above, the deep learning methods are data hungry, and 2) they do not have complex nonlinear patterns. It is also important to highlight that our results only used markers (not pedigree) information, whereas some researchers have reported similar results using both pedigree and markers (Ankamah-Yeboah et al., 2020; Calleja-Rodriguez, et al., 2020). Furthermore, the training process with the two deep learning methods (DL_M1 and DL_M2) was considerable slower than the conventional GBLUP method.

The proposed method is attractive for four reasons: 1) its implementation is straightforward, 2) it is a post-processing method, 3) it helps increase the prediction performance of deep

learning methods and 4) even with a small grid for the tuning process of the DL model, the proposed calibration method will provide reasonable predictions. We observed that the proposed DL method works better with smaller data sets, which is of paramount importance because the lower the data set, the harder the training process of deep learning methods become, since it is well documented that deep learning methods are data hungry (Chollet and Allaire, 2017; Chollet, 2018). In deep learning methods, the prediction accuracy is strongly influenced by the sample size, the heritability, the genetic architecture of the trait of interest, the genome structure of the species under study (Daetwyler et al., 2010; Schopp et al., 2017) and the mating design and family structure of the training set (Hickey et al., 2014). The relatedness between the training and testing sets also plays an important role (Habier et al., 2007; Saatchi et al., 2011; Clark et al., 2012; Lorenz and Nice 2017).

Note that the results presented in this study are not completely definitive, as more empirical evidence is required to be able to claim that the proposed method really helps improve the prediction performance of deep learning methods. The current results are attractive since we observed that the proposed calibration method (DL_M2) helped to more efficiently train deep learning models with an increase in prediction accuracy over the conventional DL methods (DL_M1) that is not negligible. Although the conventional GBLUP method behaves as the best model for the medium to large size data sets included in this study, there are cases where the DL_M1 and the DL_M2 overcame GBLUP genome-enabled prediction, as in the case of data set 4 (wheat data set) and provide similar results to those obtained by Montesinos-Lopez et al. (2019a,b).

DL methods should be used over conventional linear models (like the GBLUP) when it is suspected that the data contain nonlinear patterns. In this vein, one important advantage of DL over conventional methods for genome-enabled prediction is that the whole genetic merit, including all non-additive effects, can potentially be predicted without the need to partition all effects (Zingaretti et al., 2020). It should also be noted that DL consists of a number of layers of neurons and is a hierarchical information extraction process, which is exemplified by the classifications of objects by DL with images (Lee et al., 2009; Chollet, 2018). In the first layer, neurons detect simple and basic features of objects; in the intermediate layers, they detect parts of objects (Chollet and Allaire, 2017; Chollet, 2018); and in the top layers, they code for objects. For this reason, as pointed out in, DL offers many areas of opportunities that should be explored in order to take the full advantage of this technology. Some of these areas of opportunities are:

1) modifying, adapting, or inventing new DL architectures, activation functions, and tuning strategies for the specific context of GS;
2) adapting, improving, and developing more user-friendly software for DL applications in GS;
3) performing greater benchmarking studies to compare the prediction performance of existing DL methods to those that are the standard genome-enabled predictions in GS;
4) exploring transfer learning for GS. The goal of transfer learning is to use the knowledge learned from one specific set of environments to ease the learning tasks in another different but similar environment;
5) exploring how to use reinforcement learning in the context of GS;
6) exploring deep generative models (generative adversarial networks (GANs) and variational auto-encoder (VAE) methods to generate new inputs (fictitious markers or independent variables) that are indistinguishable from the original training set;
7) training or retraining breeders and people involved in genomic prediction in these new frameworks for DL, as exemplified by Keras (Chollet and Allaire, 2017);
8) exploring the deep compression methods in GS to reduce the computation and storage required by neural networks;
9) increasing our efforts for data sharing in platforms to create large data sets for each species containing not only phenotypic and markers data, but also environmental information and other omics data (Montesinos-López et al., 2021a,c,d);
10) taking advantage of DL tools to include in the predictor imaging information that is being collected in plants and to also measure using computer vision tools phenotypic properties that are fast, non-invasive and low-cost (Fahlgren et al., 2015).

Nevertheless, researchers and practitioners must be aware that DL is not always the right method, and for this reason, we need to be open to trying other models. As pointed out in Montesinos-López et al. (2021a), there is still not enough empirical evidence that deep learning methods outperform conventional genomic prediction models or that DL methods are more computationally demanding. However, we need to be conscious that deep learning is just starting to be used in genetics and plant breeding and is not well researched, especially for its optimal implementation in this area of research (Varshney R. K. et al., 2021). In this study, we used a feed-forward deep neural network also known as multilayer perceptron neural network and, for this reason, our results are limited to this type of architectures (topologies). More empirical evaluations are needed to corroborate that the proposed method is also efficient for other deep learning architectures.

CONCLUSION

In this paper we proposed a simple calibration method for outputs from deep learning methods with continuous response variables. We found that the proposed calibration methods help to significantly improve prediction accuracy obtained from deep learning methods, and even greater improvement in smaller data sets. The proposed DL method contributes to the training process in the context of small data sets. However, we suggest performing more empirical evaluations to accumulate more evidence of the utility of the proposed calibration method. Another advantage of the proposed calibration method is that it is a post-processing method that is very simple to implement, as it involves only a few and simple steps. In general, results demonstrated that no unique model/method

exists for producing the most accurate genome-enabled predictions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/brandon-mosqueda/dlc-datasets.

AUTHOR CONTRIBUTIONS

OM-L, and AM-L had the original idea of developing the new DL calibration method. OM-L, JC, and AM-L wrote the first version of the manuscript and revised and correct all the other version. Authors BM-G, AB, ML, and RV contribute revising several versions of the manuscripts at different stages of the research-writing process. AB prepared the last version of the article.

FUNDING

We are thankful for the financial support provided by the Bill and Melinda Gates Foundation (INV-003439, BMGF/FCDO, Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AG2MW)), the USAID projects (USAID Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, AGG-Maize Supplementary Project, AGG (Stress Tolerant Maize for Africa), and the CIMMYT CRP (maize and wheat). We acknowledge the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR Grant 267806.

ACKNOWLEDGMENTS

We thank all scientists, field workers, and lab assistants from the National Programs who collected the data used in this study.

REFERENCES

Ankamah-Yeboah, T., Janss, L. L., Jensen, J. D., Hjortshøj, R. L., and Rasmussen, S. K. (2020). Genomic Selection Using Pedigree and Marker-By-Environment Interaction for Barley Seed Quality Traits from Two Commercial Breeding Programs. *Front. Plant Sci.* 11, 539. doi:10.3389/fpls.2020.00539

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., et al. (2016). Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Front. Genet.* 7, 221. doi:10.3389/fgene.2016.002210.3389/fgene.2016.00221

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics* 23, 2633–2635. doi:10.1093/bioinformatics/btm308

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction Using Pedigree and Dense Molecular Markers. *Crop Sci.* 52, 707–719. doi:10.2135/cropsci2011.06.0299

Calleja-Rodriguez, A., Pan, J., Funda, T., Chen, Z., Chen, Z., Baison, J., et al. (2020). Evaluation of the Efficiency of Genomic versus Pedigree Predictions for Growth and wood Quality Traits in Scots pine. *BMC Genomics* 21, 796. doi:10.1186/s12864-020-07188-4

Chollet, F., and Allaire, J. J. (2017). *Deep Learning with R.Manning Early Access Program (MEA)*. first edition. Manning Publications.

Chollet, Francois. (2018). *Deep Learning with Python*. Manning Publication Co.

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (2012). The Importance of Information on Relatives for the Prediction of Genomic Breeding Values and the Implications for the Makeup of Reference Data Sets in Livestock Breeding Schemes. *Genet. Sel Evol.* 44, 4. doi:10.1186/1297-9686-44-4

Crossa, J., Campos, G. d. l., Pe´rez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics* 186, 713–724. doi:10.1534/genetics.110.118521

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22 (11), 961–975. doi:10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-wide Evaluation Methods. *Genetics* 185 (3), 1021–1031. doi:10.1534/genetics.110.116855

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. *Genetics* 182, 375–385. doi:10.1534/genetics.109.101501

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A Robust, Simple Genotyping-By-Sequencing (GBS) Approach for High Diversity Species. *PLoS One* 6, e19379. doi:10.1371/journal.pone.0019379

Fahlgren, N., Gehan, M. A., and Baxter, I. (2015). Lights, Camera, Action: High-Throughput Plant Phenotyping Is Ready for a Close-Up. *Curr. Opin. Plant Biol.* 24, 93–99. doi:10.1016/j.pbi.2015.02.006

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled Prediction of Genetic Values Using Radial Basis Function Neural Networks. *Theor. Appl. Genet.* 125, 759–771. doi:10.1007/s00122-012-1868-9

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* 177 (4), 2389–2397. doi:10.1534/genetics.107.081190

Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant Breeding with Genomic Selection: Gain Per Unit Time and Cost. *Crop Sci.* 50, 1681–1690. doi:10.2135/cropsci2009.11.0662

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Sci.* 54, 1476–1488. doi:10.2135/cropsci2013.03.0195

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic Selection in Plant Breeding: from Theory to Practice. *Brief. Funct. Genomics* 9, 166–177. doi:10.1093/bfgp/elq001

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional Deep Belief Networks for 414 Scalable Unsupervised Learning of Hierarchical Representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Canada, 609–616. doi:10.1145/1553374.1553453

Lorenz, A., and Nice, L. (2017). "Training Population Design and Resource Allocation for Genomic Selection in Plant Breeding," in *Genomic Selection for Crop Improvement*. Editors R. Varshney, M. Roorkiwal, and M. Sorrells (Cham: Springer), 7–22. doi:10.1007/978-3-319-63170-7_2

Mellers, G., Mackay, I., Cowan, S., Griffiths, I., Martinez-Martin, P., Poland, J. A., et al. (2020). Implementing Within-cross Genomic Prediction to Reduce Oat Breeding Costs. *Plant Genome* 13, e20004. doi:10.1002/tpg2.20004

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., and Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 Genes|Genomes|Genetics* 5, 2383–2390. doi:10.1534/g3.115.021667

Montesinos-Lopez, O. A., Montesinos-Lopez, J. C., Salazar, E., Barron, J. A., Montesinos-Lopez, A., Buenrostro-Mariscal, R., et al. (2021d). Application of a Poisson Deep Neural Network Model for the Prediction of Count Data in Genome-based Prediction. *Plant Genome* 29, e20118. doi:10.1002/tpg2.20118

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018a). Multi-environment Genomic Prediction of Plant Traits Using Deep Learners with a Dense Architecture. *G3: Genes|Genomes|Genetics* 8 (12), 3813–3828. doi:10.1534/g3.118.200740

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019b). New Deep Learning Genomic-Based Prediction Model for Multiple Traits with Binary, Ordinal, and Continuous Phenotypes. *G3: Genes|Genomes|Genetics.* 9 (5), 1545–1556. doi:10.1534/g3.119.300585

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F., Pérez-Hernández, O., Eskridge, K. M., et al. (2016). A Genomic Bayesian Multi-Trait and Multi-Environment Model. *G3: Genes|Genomes|Genetics* 6 (9), 2725–2744. doi:10.1534/g3.116.032359

Montesinos-López, O. A., Montesinos-López, A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018b). Multi-trait, Multi-Environment Deep Learning Modeling for Genomic-Enabled Prediction of Plant. *G3: Genes|Genomes|Genetics* 8 (12), 3829–3840. doi:10.1534/g3.118.200728

Montesinos-López, O. A., Vallejo, M., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019a). A Benchmarking between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding. *G3: Genes|Genomes|Genetics.* 9 (2), 601–618. doi:10.1534/g3.118.200998

Montesinos-López, O. A., Montesinos-López, A., Mosqueda-Gonzalez, B. A., Montesinos-López, J. C., Crossa, J., Ramirez, N. L., et al. (2021a). A Zero Altered Poisson Random forest Model for Genomic-Enabled Prediction. *Genes|Genomes|Genetics* 11 (2), jkaa057. doi:10.1093/g3journal/jkaa057

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W. R., Fajardo-Flores, S. B., et al. (2021c). A Review of Deep Learning Applications for Genomic Selection. *BMC Genomics* 22, 19. doi:10.1186/s12864-020-07319-x

Pandey, M. K., Chaudhari, S., Jarquin, D., Janila, P., Crossa, J., Patil, S. C., et al. (2020). Genome-based Trait Prediction in Multi- Environment Breeding Trials in Groundnut. *Theor. Appl. Genet.* 133, 3101–3117. doi:10.1007/s00122-020-03658-1

Pérez, P., and de los Campos, G. (2014). BGLR: a Statistical Package for Whole Genome Regression and Prediction. *Genetics* 198 (2), 483–495. doi:10.1534/genetics.114.164442

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-By-Sequencing Approach. *PLoS One* 7, e32253. doi:10.1371/journal.pone.0032253

Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., et al. (2018). Genomic-enabled Prediction Models Using Multi-Environment Trials to Estimate the Effect of Genotype × Environment Interaction on Prediction Accuracy in Chickpea. *Sci. Rep.* 8, 11701. doi:10.1038/s41598-018-30027-2

Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., et al. (2011). Accuracies of Genomic Breeding Values in American Angus Beef Cattle Using K-Means Clustering for Cross-Validation. *Genet. Sel Evol.* 43, 40. doi:10.1186/1297-9686-43-40

Schopp, P., Müller, D., Technow, F., and Melchinger, A. E. (2017). Accuracy of Genomic Prediction in Synthetic Populations Depending on the Number of Parents, Relatedness, and Ancestral Linkage Disequilibrium. *Genetics* 205 (1), 441–454. doi:10.1534/genetics.116.193243

Tuberosa, R. (2012). Phenotyping for Drought Tolerance of Crops in the Genomics Era. *Front. Physio.* 3, 347. doi:10.3389/fphys.2012.00347

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M. E. (2021b). Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci.* 26, 631–649. doi:10.1016/j.tplants.2021.03.010

Varshney, R. K., Bohra, A., Roorkiwal, M., Barmukh, R., Cowling, W. A., Chitikineni, A., et al. (2021a). Fast-forward Breeding for a Food-Secure World. *Trends Genet.* 37, 1124–1136. doi:10.1016/j.tig.2021.08.002

Wolpert, D. H. (1996). The Lack of A Priori Distinctions between Learning Algorithms. *Neural Comput.* 8 (7), 1341–1390. doi:10.1162/neco.1996.8.7.1341

Zhong, S., Dekkers, J. C. M., Fernando, R. L., and Jannink, J.-L. (2009). Factors Affecting Accuracy from Genomic Selection in Populations Derived from Multiple Inbred Lines: a Barley Case Study. *Genetics* 182, 355–364. doi:10.1534/genetics.108.098277

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Spsecies. *Front. Plant Sci.* 11, 25. doi:10.3389/fpls.2020.00025

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.