Automated Machine Learning: a case study of genomic “image-based” prediction in maize hybrids

Giovanni Galli
University of São Paulo https://orcid.org/0000-0002-3400-7978
Felipe Sabadin
University of São Paulo https://orcid.org/0000-0003-2937-1465
Rafael Massahiro Yassue
University of São Paulo https://orcid.org/0000-0002-7424-2227
Cassia Galves de Souza
University of Campinas https://orcid.org/0000-0002-4694-9946
Humberto Fanelli Carvalho
University of São Paulo https://orcid.org/0000-0003-0745-7583
Roberto Fritsche-Neto (roberto.neto@usp.br)
University of São Paulo https://orcid.org/0000-0003-4310-0047

Research Article

Keywords: non-image to image, non-linear effects, Convolutional Neural Networks, AutoML, accuracy

Posted Date: August 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-840380/v1

Automated Machine Learning: a case study of genomic "image-based" prediction in maize hybrids

Giovanni Galli[a], Felipe Sabadin[a], Rafael Massahiro Yassue[a], Cassia Galves de Souza[b], Humberto Fanelli Carvalho[a] and Roberto Fritsche-Neto[a,c]*

[a]University of São Paulo, Luiz de Queiroz College of Agriculture, Department of Genetics, Piracicaba, São Paulo, Brazil.

[b]Department of Food Engineering, School of Food Engineering, University of Campinas, Campinas, São Paulo, Brazil

[c] International Rice Research Institute (IRRI), Los Baños, Philippines.

*Corresponding author

E-mail: roberto.neto@usp.br

ORCID of the authors:

Roberto Fritsche-Neto (RFN): 0000-0003-4310-0047

Giovanni Galli (GG): 0000-0002-3400-7978

Felipe Sabadin (FS): 0000-0003-2937-1465

Rafael Massahiro Yassue (RMY): 0000-0002-7424-2227

Cassia Galves de Souza (CGS): 0000-0002-4694-9946

Humberto Fanelli Carvalho (HFC): 0000-0003-0745-7583

ABSTRACT

Machine learning methods such as Multilayer perceptrons (MLP) and Convolutional Neural Networks (CNN) have emerged as promising methods for genomic prediction (GP). In this sense, we assess the performance of MLP and CNN on regression and classification tasks in a case study with maize hybrids. The genomic information was provided to the MLP as a relationship matrix and to the CNN as "genomic images". In the regression task, the machine learning models were compared along with GBLUP. Under the classification task, MLP and CNN were compared. In this case, the traits (plant height and grain yield) were discretized in such a way to create balanced (moderate selection intensity) and unbalanced (extreme selection intensity) datasets for further evaluations. An automatic hyperparameter search for MLP and CNN was performed, and the best models were reported. For both task types, several metrics were calculated under a validation scheme to assess the effect of the prediction method and other variables. Overall, MLP and CNN presented competitive results to GBLUP but improved a little using only the additive genomic layer. It is expected that the average effect of allele substitution is mostly linear. Nevertheless, the methodology's potential for GP is unprecedented because we can create "multispectral genome images," including other effects and layers of data, such as dominance, epistasis, g × e, transcriptome, and so on, capturing linear and non-linear effects and boosting prediction accuracies. Hence, we bring new insights on automated machine learning for genomic prediction and its implications to plant breeding.

KEYWORDS: non-image to image; non-linear effects; Convolutional Neural Networks; AutoML; accuracy;

DECLARATIONS

Funding: this work was financially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Availability of data and material: data has not been made available.

Code availability: code has not been made available

COMPLIANCE WITH ETHICAL STANDARDS

Conflicts of interest/Competing interests: the authors declare no conflict of interest.

AUTHOR CONTRIBUTION STATEMENT

GG elaborated on the hypothesis, conducted the analyses, and wrote the manuscript. RFN, HFC, RMY, FS and CGS contributed to interpret the results and writing. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The Allogamous Plant Breeding Laboratory team (Luiz de Queiroz College of Agriculture, University of São Paulo, Brazil) thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support.

INTRODUCTION

Genomic prediction (GP) arose as a breeding tool capable of enabling a considerable increase in the rates of genetic gain. In this sense, three decades of scientific research have shown that the accuracy of this statistical approach might be conditioned to a series of factors, including the quality and pre-processing of the phenotypic data (Galli et al. 2018), the platform used to obtain genomic information and how it is processed (Granato et al. 2018; Sousa et al. 2019), the population mating design (Fristche-neto et al. 2018), the intrinsic genetic architecture of the trait (Alves et al. 2019), the genetic structure of the population (Lyra et al. 2018), how the genotype-by-environment interaction is dealt with (Costa-Neto et al. 2020; Alves et al. 2021), and which prediction methods are used (e.g., BayesA, BayesB (Meuwissen et al. 2001); GBLUP (Bernardo 1994; VanRaden 2008); Reproducing Kernel Hilbert Spaces (de Los Campos et al. 2009); and machine learning (e.g., (deep) neural networks; NN)). Hence, it is the scientist's discretion to adjust the factors mentioned above to their best convenience and knowledge.

Regarding the choice of a GP method, amongst the myriad of options, machine learning-based algorithms, e.g., (deep) neural networks, have emerged as one of the most promising over the past years. Multilayer perceptrons (MLPs; fully connected layers) and Convolutional Neural Networks (CNNs; fully connected layers and convolutional/pooling filters) are two common types of neural networks. These methods are characterized by the sequentially stacking (several) layers, which automatically identifies latent patterns or features from data (Trevisan et al. 2020). For a technically accurate and contextualized explanation of such models, refer to Pérez-Enciso and Zingaretti (2019). This rising interest is fundamentally associated with the increasing availability of computational power (e.g., graphical processing unit computing, cloud computation, web servers); its success in diverse tasks (such as self-driving vehicles, object detection, and context recognition); ability to work on both regression and classification problems; and especially due to the lower-level restrictions compared to standard models. For instance, neural networks can perform predictions without restrictive model assumptions; in the context of genetic studies, it does not require specifying the distribution of variables, priors, and the nature of genetic effects (additive, dominance, and epistasis), being theoretically capable of self-adjusting to the underlying genetic architecture (Pérez-Enciso and Zingaretti 2019).

Initial reports suggest that (deep) neural networks can be competitive to the standard GP methods (e.g., GBLUP) in prediction accuracy. Nevertheless, results are highly inconsistent on this matter (Bellot

et al., 2018; Ma et al., 2018; Montesinos-López et al., 2018a, 2018b; Azodi et al., 2019; Abdollahi-Arpanahi et al., 2020), and its best use and performance is still to be determined on a broader and most representative spectrum of prediction scenarios. In this sense, one of the major challenges for applying this methodology is identifying adequate model structures and hyperparameters (Bellot et al. 2018; Pérez-Enciso and Zingaretti 2019; van Dijk et al. 2021). Hereon, we refer to hyperparameter as, e.g., number of hidden layers, number of neurons per layer, learning rate, filter type and number, activation function, optimization algorithm, regularization type, etc. Since it is an exceptionally flexible algorithm, there is an infinite number of possible configurations. Therefore, automated procedures are required to explore the possibilities and increase the chance of finding near-to-optimal hyperparameters.

Automated Machine Learning (AutoML) has great potential for identifying adequate network structures and hyperparameters for a given task (Jin et al. 2019). These procedures circumvent hand-designing and testing hyperparameters to save time and effort. Numerous platforms have been developed, such as Auto-sklearn (Feurer et al. 2015), Auto-Weka (Kotthoff et al. 2017), and AutoKeras (Jin et al. 2019); each one with its search algorithm. A comprehensive guide and benchmarking study on the most common search platforms is presented by Truong et al. (2019) for further reference. Despite the importance of hyperparameter tuning and the availability of easy-to-use AutoML tools, the number of reports on its use for identifying artificial neural networks for GP is still very limited (Zingaretti et al. 2020).

Besides adequate hyperparameter tuning, the performance of a network is also determined by the quality and preparation of the data fed for training. For example, in neural network-based GP models, the genomic information has been input as a genomic relationship/distance matrix (Montesinos-López et al., 2018a, 2018b), or as the genomic matrix (Azodi et al., 2019; Abdollahi-Arpanahi et al., 2020). In the case of CNN, the organization of the matrix is meaningful and might contain valuable information (Pérez-Enciso and Zingaretti 2019). For example, Abdollahi-Arpanahi et al. (Abdollahi-Arpanahi et al. 2020) applied CNN with genomic matrices to exploit linkage disequilibrium (LD) patterns between genetic markers. In this case, meaningful filter movements were restricted to a single direction (chromosome-wise), seizing physical linkage disequilibrium (e.g., neighboring markers). Nevertheless, LD is known to vary across the genome (Bellot et al. 2018); hence, further advancements to this methodology have been proposed, such as using local convolutional layers applying region-specific filters (Pook et al. 2020).

Recently, a work by Sharma et al. (2019) has shown the possibility of transforming non-image data (e.g., a genomic matrix) into "images" (2 or 3-dimensional visual matrices) leveraging dimensionality reduction techniques. In images, data is coherently distributed along with a space pattern, meaning that neighboring pixels share information in all directions are correlated (Sharma et al. 2019). The authors reported the superiority of image-based CNN over original data in machine learning tasks and named the pipeline *DeepInsight*. In context, using genomic images would presumably unlock the potential of CNNs for GP, capturing the relationships between SNP over new dimensions (e.g., non-physical – intrachromosomal - LD).

To test new methodologies in a GP context, a key component of comprehensive and meaningful benchmarking starts with an adequate choice of comparison metrics. In this sense, regression tasks have mainly relied on metrics such as Pearson's product-moment correlation and its variations (e.g., divided by the trait's heritability), Spearman's correlation, repeatability/heritability, reliability, etc. However, some of these metrics cannot represent the core practice of plant breeding, which is ranking and selection (Ornella et al. 2014; Blondel et al. 2015). For example, in the most used metric, Pearson's correlation, all ranking positions are equally important, meaning that interclass distinction (selected or non-selected) is neglected. This problem has been tackled using selection-centered metrics, such as selection coincidence (Matias et al. 2017; Galli et al. 2018; Alves et al. 2019). We add to this matter by unifying ranking and selection by discretizing continuous data and conducting prediction based on classification methods suggested by Ornella et al. (2014). That opens the possibility of comparing methods with a new realm of metrics that better align with the context of plant breeding.

AutoML has a full, yet to be determined, potential application for breeding targeted GP. In this context, we present a comprehensive study on using these technologies for predicting plant height (PH) and grain yield (GY) in maize. The objectives of this research were to: *i)* assess the comparative performance of MLP and CNN with the standard model GBLUP at predicting PH and GY in maize; *ii)* evaluate the performance of neural networks for the GP of PH and GY in maize in regression and classification contexts using MLP and CNN; *iii)* elaborate on the use of AutoML to identify the best hyperparameters to perform GP; *iv)* and verify the usefulness transforming genomic information into images for CNN based GP.

MATERIAL AND METHODS

Dependent variables

149 Field trials

150 The genetic material was composed of 904 maize single-cross hybrids obtained from a partial diallel
151 of 49 tropical inbred lines (Fritsche-Neto et al. 2019). Thorough populational description and statistics have
152 been reported on both the inbred lines and hybrids (Fristche-neto et al. 2018; Alves et al. 2019; Morosini
153 et al. 2020).

154 The genotypes were arranged in unreplicated trials with the augmented block scheme. Each
155 incomplete block was composed of 18 treatments, 16 regular and two checks (common genotypes). The
156 trials were carried out at Piracicaba-São Paulo (22°42'23" S, 47°38'14" W, 535 m) and Anhembi-São Paulo
157 (22°50'51" S, 48°01'06" W, 466 m), during the second growing seasons of 2016 (738 hybrids) and 2017
158 (789 hybrids), under two nitrogen application regimes (ideal: 0.1 Mg ha^{-1} and low: 0.03 Mg ha^{-1}). Each
159 experimental unit was composed of a 7 m row. The single-crosses were phenotyped for GY (Mg ha^{-1}) and
160 PH (cm). GY was estimated as the production of a plot corrected for 13% moisture. PH was obtained as
161 the mean height, measured from soil to flag leaf, of five plants in the plot.

162 Phenotypic analysis

163 The genotypic values of hybrids were obtained with a joint linear mixed model using in ASReml-R
164 (Gilmour et al. 2009) following:

165 $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 b} + \boldsymbol{Z_2 g} + \boldsymbol{\varepsilon}$$

166 where $\boldsymbol{y}$ is the phenotype (PH or GY); $\boldsymbol{\beta}$ is the vector of fixed effects of check, environment (combinations
167 of site, year, and nitrogen regime) and check × environment; $\boldsymbol{b}$ $[\sim N(\mathbf{0}, \boldsymbol{I}\sigma_b^2)]$ is the random effect of block-
168 within-environment; $\boldsymbol{g}$ $[\sim N(\mathbf{0}, \boldsymbol{I}\sigma_g^2)]$ is the random effect of regular genotypes (genotypic values); and $\boldsymbol{\varepsilon}$
169 $[\sim N(\mathbf{0}, diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_8^2)]$ is a vector of residuals structured by environment estimated from the
170 common treatments (checks). $\boldsymbol{X}$, $\boldsymbol{Z_1}$, and $\boldsymbol{Z_2}$ are the incidence matrices of the mentioned factors. Likelihood
171 Ratio Test (LRT) was used to determine the significance of random effects.

172 Additionally, a similar model was fit, having check as fixed and regular genotypes, environment,
173 genotype (checks) × environment, and block-within-environment as random for the estimation of variance
174 components. Repeatability at plot level $R_i = \hat{\sigma}_g^2 / \left(\hat{\sigma}_g^2 + \hat{\sigma}_{ga}^2 + \hat{\sigma}_\varepsilon^2\right)$ was estimated having $\hat{\sigma}_g^2$, $\hat{\sigma}_{ga}^2$, and $\hat{\sigma}_\varepsilon^2$
175 as the genotypic (hybrids), genotypic (checks) × environment, and residual variances, respectively. The
176 residual variance ($\hat{\sigma}_\varepsilon^2$) was regarded as the mean residual across environments.

177 Genotypic values pre-processing

178 The ultimate goal of plant breeding is ranking and selecting the best genotypes. A common practice
179 is to categorize genotypes in groups of selected and non-selected based on their genetic merit. In this sense,
180 the subsequent analysis regards genotypic values in two manners, as continuous or a discrete variable. First,
181 genotypic values were categorized based on the empiric distribution or absolute values depending on the
182 trait, using two selection intensities (SI), moderate and extreme (Fig. S1). The moderate SI was created to
183 mimic a balanced dataset regarding the selected and non-selected classes; while the extreme SI was created
184 for an unbalanced dataset. For GY, the higher-yielding individuals were regarded as the best. For the
185 extreme SI, about 10% of the higher-yielding genotypes were selected; and for the moderate SI, around
186 50% of the higher-yielding were selected for the moderate SI. Second, the selection for PH was based on a
187 hypothetical ideotype. In this case, genotypic values between 1.95 and 2.05 m (~60%; moderate SI) were
188 selected; additionally, genotypic values between 1.90 m and 2.10 (~90%; extreme SI) were regarded as
189 selected. Notice that under the extreme SI, the 10% best were selected for GY, while for PH, the 10% out
190 of type were eliminated. This approach was chosen because selecting the central 10% of hybrids for PH
191 would result in ~0.02 m variation within the "selected" class, which might not be realistic under usual
192 breeding premises.

193 The categorized genotypes were used for classification tasks in the subsequent analysis. For this, the
194 selected genotypes were attributed value 1, while the non-selected had value 0. Notice that the metrics used
195 to evaluate the prediction models might have different meanings for GY and PH. For example, under
196 extreme SI, the individuals regarded as selected compose a low proportion of the samples for GY, while
197 for PH, they are the majority. The original continuous variables were used in regression tasks. In this case,
198 the genotypic values were scaled using $\hat{g} = (\hat{g} - \hat{g}_{min})/(\hat{g}_{max} - \hat{g}_{min})$, were $\hat{g}_{min}$ and $\hat{g}_{max}$ are the
199 minimum and maximum genotypic values, respectively.

200 *Independent variables*

201 A graphical summary of the procedures explained hereon is presented in Fig.1.

202

203 Fig.1. Summarized general exemplification of the employed methodology. A) Genomic data obtained from
204 the pre-processing step; B) Additive Genomic Relationship Matrix (GRM) obtained with VanRaden's

205 method using the genomic matrix (A); C) *DeepInsight* pipeline: t-SNE decomposition of the genomic
206 matrix (A) with original (dark gray) and rotated (light gray) marker coordinates; D) Genomic images
207 obtained with *DeepInsight*; one image is produced for each hybrid and all markers are represented in the
208 image; E) representation of the genotypes (0, 1, and 2, also white, light gray, and dark gray, respectively)
209 in a genomic image; each pixel might comprise a single or multiple markers depending on the level of
210 linkage disequilibrium in the population; F) Genomic prediction methods: genomic BLUP (GBLUP),
211 multilayer perceptron (MLP), and convolutional neural network (CNN); GBLUP and MLP used the GRM
212 (B) as independent variable, while CNN used genomic images (D); the neural networks were used as both
213 regression and classification tasks; AutoKeras was used for hyperparameter search in MLP and CNN; G)
214 Simplified representation of the nested validation procedure.

215 Genomic data pre-processing

216 The parental inbred lines were genotyped with the Affymetrix® Axiom® Array of 614k SNPs
217 (Unterseer et al. 2014). The genomic data pre-processing was performed following the procedure presented
218 in Galli et al. (2020) by: removing markers with low call rate (<95%); removing markers with at least one
219 heterozygote in the population; imputing missing (homozygous) data with the Synbreed-R package
220 (Wimmer et al. 2012); pruning with Plink v. 1.9 (Chang et al. 2015) so the maximum linkage disequilibrium
221 between markers is 0.9 to avoid high-level redundancy between marker information; building the hybrids
222 synthetic genomic matrix; and, removing markers with minor allele frequency lower than 5%. After pre-
223 processing, a total of 34,571 markers remained for further analysis. Principal component analysis (Lyra et
224 al. 2017; Morosini et al. 2017), linkage disequilibrium decay (Morosini et al. 2017), distribution of minor
225 allele frequency, and heterozygosity (Morosini et al. 2020) have been reported for this dataset.

226 Genomic Relationship Matrix (GRM)

227 The genomic information was transformed into two types of data for inclusion in prediction methods.
228 The first type utilized was the additive GRM. We opted for VanRaden's (2008) baseline method to
229 determine the genomic relationship between genotypes. The relationship was obtained as $\boldsymbol{G} = \frac{\boldsymbol{XX'}}{trace(\boldsymbol{XX'})/n}$,
230 where $\boldsymbol{X}$ is the matrix of genotypic information and n is the number of individuals. The GRM was obtained
231 using the *G.matrix* function of the *snpReady* (Granato et al. 2018) R library.

232 Obtaining images from genomic data

233 The second type of data transformation was performed by converting the structured genotype matrix
234 by marker into images. This was achieved using the *DeepInsight* algorithm proposed by Sharma et al.
235 (2019). In summary, the algorithm applies a similarity measuring/dimensionality reduction technique (e.g.,
236 t-SNE, kPCA) to obtain a Cartesian representation of the similarity between genomic markers in the
237 population. At this step, one graph is produced, and each point represents a marker (Fig.1 C, dark gray). In
238 this sense, if two markers are somehow related due to, e.g., linkage disequilibrium, they should have similar
239 coordinates. Then, the algorithm finds the smallest rectangle containing all the points and applies a rotation
240 to the graph, so the rectangle is vertically or horizontally oriented (Fig.1 C, light gray). At this point, the
241 graph is converted to an image, and the genomic marker information (e.g., 0, 1, or 2) is mapped to its
242 corresponding position (Fig.1 E). This procedure produces one image per hybrid (Fig.1 D).

243 Using *DeepInsight*, images were generated for the 904 genotypes (Fig.1 D). The genomic matrix
244 mapped to images had 0, 1, and 2 parametrizations (Fig.1 E), commonly used to estimate additive effects
245 of markers or additive GRMs in genomic prediction. The Cartesian plane marker coordinates were
246 estimated using kPCA and t-SNE; no relevant difference was found on preliminary tests, and the latter was
247 selected. The 120 × 120 pixels resolution presented adequate results regarding image size, given the number
248 of features (markers) and the available computational power. The *DeepInsight* algorithm is implemented
249 in MATLAB and available at http://www.alok-ai-lab.com.

250 *Genomic prediction*

251 Prediction scenarios

252 The GP methods used were GBLUP (standard method), MLP (using the GRM), and CNN (using the
253 genomic images obtained with *DeepInsight*) (Fig.1 F). GP was performed as regression and classification
254 tasks, i.e., the dependent variable (GY or PH) was continuous or discrete, respectively. For the regression
255 task, the evaluated scenarios were: 1) GBLUP; 2) MLP; and 3) CNN. For the classification task, the
256 scenarios were: 1) MLP under moderate SI; 2) MLP under extreme SI; 3) CNN under moderate SI; and 4)
257 CNN under extreme SI. Thus, these scenarios enabled estimating the effect of prediction methods (GBLUP
258 vs. MLP vs. CNN) in the regression task; the impact of selection intensity (moderate vs. extreme) in the

classification task; and the effect of and data type/prediction method (MLP (GRM) vs. CNN (genomic image)) on both regression and classification.

GBLUP

GBLUP is a standard regression task and was performed using ASReml-R (Gilmour et al. 2009) following the given linear model:

$$\widehat{\boldsymbol{g}} = \mathbf{1}\mu + \boldsymbol{Z_3 h} + \boldsymbol{e}$$

where $\widehat{\boldsymbol{g}}$ is the scaled vector of genotypic values of hybrids; μ is the overall mean; $\boldsymbol{h}$ $[\sim N(\mathbf{0}, \boldsymbol{G}\sigma_h^2)]$ is the vector of genomic estimated breeding values, considering that $\boldsymbol{G}$ is the VanRaden's (2008) additive relationship matrix; and $\boldsymbol{e}$ $[\sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)]$ is the residual; $\mathbf{1}$ and $\boldsymbol{Z_3}$ are the incidence matrices for the mentioned factors.

(Deep) Neural networks

We call the attention that this work is not focused on an in-depth explanation of neural networks as an algorithm despite the need for a basic understanding of neural networks. If the reader is not familiarized with the subject, we encourage the reading of González-Camacho et al. (2016) and Pérez-Enciso and Zingaretti (Pérez-Enciso and Zingaretti 2019) for a thorough comprehension of key concepts.

Neural networks were performed for regression and classification. In this sense, the python AutoML system *AutoKeras* (Jin et al. 2019) was used. AutoML libraries perform neural architecture search with minor manual intervention and enable the automated finding of population-specific machine learning models. In this sense, regression was implemented with the *ImageRegressor* and the *StructuredDataRegressor* functions to search suitable CNNs and MLPs, respectively. The loss function was the mean squared error (MSE; $\frac{1}{N}\sum_{i=1}^{N} e_i^2$) and the metrics were the mean absolute error (MAE; $\frac{1}{N}\sum_{i=1}^{N}|e_i|$), and Pearson's product-moment correlation (r; $\sum_{i=1}^{N}(gt_i - \mu_{gt})(gp_i - \mu_{gp}) / \sqrt{\sum_{i=1}^{N}(gt - \mu_{gt})^2 \sum_{i=1}^{N}(gp - \mu_{gp})^2}$), given that $e_i = gt_i - gp_i$ is the residual for hybrid i, gt_i is the genotypic value of hybrid i, gp_i is the predicted value of hybrid i, N is the number of observations, $\mu_{gt} = \frac{1}{N}\sum_{i=1}^{N} gt_i$ is the mean of genotypic values, and $\mu_{gp} = \frac{1}{N}\sum_{i=1}^{N} gp_i$ is the mean of predicted values.

The classification was performed with the *ImageClassifier* and the *StructuredDataClassifier* functions for identifying CNNs and MLPs, respectively. In this sense, positives (p) are genotypes that would have been selected based on their genotypic value, and negatives (n) are non-selected genotypes. A loss function and several metrics were estimated based on the number of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn). The loss function was the binary cross-entropy, and the metrics were true negative rate (TNR; tn/n), precision [$tp/(tp+fp)$], recall (or true positive rate; TPR) [$tp/(tp+fn)$], F_1 score [$2tp/(2tp+fp+fn)$], accuracy [$(tp+tn)/(tp+tn+fp+fn)$], balanced accuracy [$(TPR+TNR)/2$], and area under the receiver operating characteristic curve (AUC). As the datasets were imbalanced, especially for the extreme selection intensity, the weights (w) of classes (selected and non-selected) were fed to the model given that $w_i = 1/f_i$, where f_i is the frequency of class i.

For both classification and regression, the maximum number of models tried by AutoKeras was 50; the number of epochs was set to 150; the batch size was set to eight; seeds were utilized for reproducibility. Finally, the objective of the search was to identify the hyperparameters that minimized the validation loss.

Classification/Regression performance

A random sampling validation scheme assessed the prediction performance under each validation scenario (Fig.1 G). For the neural networks (MLP and CNN), an adaptation of the validation was applied to steps 1 and 2, enabling the identification of the best set of hyperparameters for each replication within a scenario, a process called inner validation, similar to the utilized by Montesinos-López et al. (2018a, 2018b). The steps that were exclusively performed for neural networks are represented by lowercase letters. The overall validation procedure was performed as follows:

1. Allocation of randomly sampled genotypes to training (80%; TS) and validation sets (20%, VS);
    a. Random assignment of samples from the training set into inner training (ITS; 80%) and inner validation sets (IVS; 20%);
    b. Identification of the best set of hyperparameters with AutoKeras using the inner training and inner validation sets.
2. Models were trained using the individuals from the inner set (ITS and IVS);
3. Prediction of the outer validation set;

4. Estimation of comparison metrics; and
5. Repeat steps 1 to 4 five times, considering equal set sampling between scenarios.

Comparisons between scenarios were made using the metrics estimated on the validation process. Values are presented as mean and standard deviation across the five replications.

RESULTS

Phenotypic analysis

According to the joint phenotypic analysis, the plot level repeatabilities were 0.23 for GY and 0.59 for PH, revealing the traits as lowly and moderately heritable, respectively. The LRT test ($P<0.05$) showed the effects of the environment, block within environment, genotype, and genotype (check) by environment to significantly affect both GY and PH. The BLUPs of GY averaged 6.79 Mg ha^{-1} ranging from 4.90 to 8.36 Mg ha^{-1}. For PH, the mean was 199.02 cm, with values varying between 170.69 and 217.74 cm (Fig. S1). More information on these genotypes can be found in Alves et al. (2019), Fritsche-Neto et al. (2019), and Galli et al. (2020).

Regression performance

The regression metrics for the prediction of PH and GY using GBLUP, MLP, or CNN are presented in Table 1. Considering each prediction scenario (e.g., PH with MLP), the metrics (MAE or MSE) were consistent across the inner training, inner validation, and outer validation sets. Therefore, scenario comparisons were performed, having the outer validation set as a reference. The values of MAE varied from 0.0635 to 0.0770 for PH and from 0.0797 to 0.0850 for GY. The loss function (MSE) varied from 0.0083 to 0.0109 for PH and from 0.0114 to 0.0128 for GY. The correlations varied from 0.68 to 0.75 for PH and from 0.53 to 0.59 for GY.

Patterns arose from comparisons between the studied scenarios. Contrasting the GBLUP (standard model) with MLP and CNN, GBLUP yielded superior results across all metrics for PH. For GY, a similar pattern was observed when comparing GBLUP with MLP, except for MSE. However, CNN outperformed GBLUP for this trait concerning all metrics. Regarding MLP or CNN, the latter presented better results for both traits considering all estimated metrics, except for r in PH.

Classification performance

The classification metrics for the prediction of PH and GY using MLP or CNN are presented in Table 2 and Supplemental Table S1. Under each prediction scenario, the loss decreased from inner training to inner validation to outer validation. Regarding the other metrics (with few exceptions), values were (in average across scenarios) greater in inner training, followed by inner validation and outer validation sets.

Nevertheless, no drastic changes were observed between sets. Hence, comparisons between scenarios were performed using the outer validation set as a reference.

The observed values of metrics varied depending on the prediction scenario. The TNR varied from 0.26 to 0.65 for PH and 0.66 to 0.97 for GY; the recall ranged from 0.54 to 0.96 for PH and from 0.17 to 0.67 for GY; the precision presented values from 0.65 to 0.92 for PH and from 0.35 to 0.70 for GY; the F1score showed results from 0.57 to 0.93 for PH and from 0.22 to 0.67 for GY; the accuracy varied largely presenting values from 0.52 to 0.88 for PH and 0.66 to 0.90 for GY; the variation of balanced accuracy varied from 0.55 to 0.61 for PH and from 0.57 to 0.68 for GY; at last, the AUC ranged from 0.61 to 0.67 for PH and from 0.71 to 0.74 for GY.

The effect of selection intensity (moderate and extreme) presented tendencies to the estimated metrics. TNR, precision, recall, and F1 score were higher at extreme selection intensity for PH. For GY, the opposite was observed. The accuracy was higher at extreme selection intensity for both traits. The balanced accuracies using GRM were equal for both selection intensities for PH and GY. However, when CNN was used, this metric was lower at extreme selection intensity for both traits.

At last, regarding the AUC, moderate-intensity presented better values for both traits. Regarding the effect of the prediction method, for predicting PH, using MLP showed better values of precision, accuracy, balanced accuracy, and AUC. However, for recall and F1 score, this was only observed at extreme intensity. For GY, using MLP generally presented better results at extreme selection intensity, while CNN was superior at moderate selection intensity. The exceptions were precision, where image-based models were better for both intensities, and recall, which presented the opposite behavior.

Auto-machine learning model tuning

The neural network structures that minimized the loss function for each replication under each scenario are presented on Supplemental File S1. The classification scenarios were constitutionally composed of an input layer as the first, a dense layer as the second-last summarizing all the neurons of the previous layer, and an activation layer with the sigmoid function to generate the output probabilities. Similarly, the regression scenarios had an input layer as the first and a dense layer as the last to summarize all neurons to one output. Nevertheless, the network structures were generally different, with few exceptions. Among these coincidences, seven out of nine were of the same task type (regression or classification), four were of the same trait (PH or GY), and three were of the same selection intensity

373 (extreme or moderate). However, the number of parameters varied greatly (from 24,533 to 23,589,764),
374 typically higher when images were used.

375 Dealing with GRM or images requires networks with specific internal layers. The scenarios with
376 MLPs presented a varying number of dense layers (1 to 4); normalization layers (0 to 4; present in about
377 half of the networks); ReLU activation function (positioned after dense layers except the last one); and
378 dropout (0 to 4; present in about 2/3 of networks). The CNNs were composed of 2-dimensional
379 convolutions (1 to 4 in classifications and 2 to 6 in regressions; present in all networks); normalization
380 layers (0 or 1; present in about 2/3 of the networks); 2-dimensional max/global max or average pooling (0
381 to 3; present in about 2/3 of networks); dropout (0 to 3; present in about 2/3 of networks); image processing
382 filters (resize, random flip, contrast, rotation, translation, and concatenation; 0 to 4; present in about half of
383 the networks; being more common for PH). Also, ResNet50 and Xception networks appeared within 1/3 of
384 the classification networks (more common for PH).

385 Finally, the preferred optimizer was Adam; Adadelta and SGD also appeared, in a limited number
386 of cases. The most common learning rate was 0.001, followed by 0.01, 0.00001, 0.0001, and 0.1. The
387 dropout regularization had values of 0.5 (most common) and 0.25.

Table 1. Regression metrics for (dependent variables; DV) plant height (PH) and grain yield (GY) using genomic BLUP (GBLUP), Multilayer Perceptrons (MLP), and Convolutional Neural Networks (CNN). Mean absolute error (MAE), mean squared error (MSE; loss function), and Pearson's product-moment correlation (r) are presented for the inner training, inner validation, and outer validation sets. The values are the mean and standard deviations (in parenthesis) across five replications

Scenario		MAE						MSE (loss)						r	
DV	Method	Inner training		Inner validation		Outer validation		Inner training		Inner validation		Outer validation		Outer validation	
PH	GBLUP	0.0607	(0.0006)	-	-	0.0635	(0.0017)	0.0076	(0.0002)	-	-	0.0083	(0.0008)	0.75	(0.03)
	MLP	0.0946	(0.0552)	0.0693	(0.0087)	0.0770	(0.0164)	0.0191	(0.0227)	0.0086	(0.0022)	0.0109	(0.0039)	0.70	(0.06)
	CNN	0.0693	(0.0053)	0.0642	(0.0077)	0.0750	(0.0107)	0.0087	(0.0013)	0.0077	(0.0024)	0.0103	(0.0023)	0.68	(0.08)
GY	GBLUP	0.0736	(0.0009)	-	-	0.0836	(0.0038)	0.0098	(0.0003)	-	-	0.0128	(0.0011)	0.56	(0.05)
	MLP	0.0922	(0.0142)	0.0823	(0.0043)	0.0850	(0.0048)	0.0147	(0.004)	0.0114	(0.0017)	0.0125	(0.0018)	0.53	(0.04)
	CNN	0.0776	(0.0117)	0.0778	(0.004)	0.0797	(0.0057)	0.0106	(0.0031)	0.0102	(0.0014)	0.0114	(0.0012)	0.59	(0.01)

Table 2. Classification metrics for (dependent variables; DV) plant height (PH) and grain yield (GY) using genomic BLUP (GBLUP), Multilayer Perceptrons (MLP), or Convolutional Neural Networks (CNN) under moderate and extreme selection intensities (SI). True negative rate (TNR), recall (or true positive rate; TPR), precision, F1 score, accuracy, balanced accuracy, AUC, and binary cross-entropy (BC; loss function) are presented for the inner training, inner validation, and outer validation sets. The values are the mean and standard deviations (in parenthesis) across five replications

Scenario			TNR		Recall (TPR)		Precision		F1 score		Accuracy		Balanced accuracy		AUC		BC (loss)	
DV	Method	SI	Inner training															
PH	MLP	Extreme	0.91	(0.20)	0.95	(0.05)	0.99	(0.02)	0.97	(0.03)	0.95	(0.06)	0.93	(0.11)	0.94	(0.13)	0.0005	(0.0007)
		Moderate	0.55	(0.17)	0.75	(0.13)	0.72	(0.05)	0.73	(0.05)	0.67	(0.05)	0.65	(0.05)	0.69	(0.07)	0.0014	(0.0001)
	CNN	Extreme	0.75	(0.21)	0.72	(0.23)	0.95	(0.04)	0.81	(0.16)	0.73	(0.22)	0.74	(0.22)	0.77	(0.21)	0.0022	(0.0029)
		Moderate	0.61	(0.12)	0.68	(0.08)	0.72	(0.07)	0.70	(0.07)	0.65	(0.09)	0.64	(0.09)	0.68	(0.12)	0.0014	(0.0002)
GY	MLP	Extreme	0.76	(0.26)	0.91	(0.12)	0.46	(0.27)	0.58	(0.28)	0.78	(0.24)	0.84	(0.17)	0.86	(0.18)	0.0009	(0.0007)
		Moderate	0.72	(0.04)	0.74	(0.01)	0.73	(0.03)	0.73	(0.02)	0.73	(0.02)	0.73	(0.02)	0.80	(0.03)	0.0012	(0.0001)
	CNN	Extreme	0.84	(0.13)	0.88	(0.12)	0.51	(0.32)	0.61	(0.28)	0.85	(0.13)	0.86	(0.12)	0.91	(0.10)	0.0007	(0.0006)
		Moderate	0.74	(0.02)	0.76	(0.02)	0.75	(0.01)	0.75	(0.01)	0.75	(0.01)	0.75	(0.01)	0.83	(0.02)	0.0011	(0.0001)
DV	Method	SI	Inner validation															
PH	MLP	Extreme	0.27	(0.18)	0.94	(0.04)	0.92	(0.03)	0.93	(0.02)	0.88	(0.03)	0.61	(0.08)	0.68	(0.16)	0.3454	(0.0951)
		Moderate	0.50	(0.16)	0.76	(0.12)	0.68	(0.06)	0.71	(0.03)	0.64	(0.04)	0.63	(0.04)	0.70	(0.06)	0.6003	(0.0399)
	CNN	Extreme	0.12	(0.16)	0.99	(0.02)	0.91	(0.03)	0.95	(0.02)	0.90	(0.03)	0.55	(0.07)	0.73	(0.14)	0.2863	(0.0640)
		Moderate	0.47	(0.13)	0.80	(0.10)	0.67	(0.06)	0.73	(0.07)	0.66	(0.08)	0.63	(0.08)	0.70	(0.05)	0.6135	(0.0402)
GY	MLP	Extreme	0.94	(0.05)	0.38	(0.23)	0.37	(0.22)	0.37	(0.21)	0.88	(0.04)	0.66	(0.10)	0.86	(0.08)	0.2621	(0.0755)
		Moderate	0.75	(0.06)	0.71	(0.06)	0.74	(0.05)	0.72	(0.04)	0.73	(0.02)	0.73	(0.02)	0.80	(0.02)	0.5427	(0.0159)
	CNN	Extreme	0.97	(0.02)	0.31	(0.20)	0.44	(0.27)	0.35	(0.21)	0.90	(0.02)	0.64	(0.09)	0.84	(0.07)	0.2651	(0.0508)
		Moderate	0.71	(0.04)	0.72	(0.09)	0.70	(0.02)	0.71	(0.05)	0.72	(0.03)	0.72	(0.03)	0.79	(0.02)	0.5486	(0.0188)
DV	Method	SI	Outer validation															
PH	MLP	Extreme	0.26	(0.21)	0.96	(0.03)	0.92	(0.02)	0.93	(0.01)	0.88	(0.01)	0.61	(0.09)	0.66	(0.11)	0.7612	(0.0904)
		Moderate	0.65	(0.18)	0.57	(0.32)	0.71	(0.05)	0.57	(0.28)	0.59	(0.14)	0.61	(0.07)	0.67	(0.07)	1.1120	(0.4637)
	CNN	Extreme	0.32	(0.27)	0.79	(0.35)	0.90	(0.02)	0.79	(0.29)	0.73	(0.28)	0.55	(0.05)	0.63	(0.08)	1.0639	(0.4752)
		Moderate	0.49	(0.15)	0.63	(0.22)	0.65	(0.06)	0.63	(0.16)	0.58	(0.07)	0.56	(0.04)	0.61	(0.02)	1.0838	(0.4774)
GY	MLP	Extreme	0.95	(0.04)	0.40	(0.23)	0.35	(0.22)	0.37	(0.21)	0.90	(0.03)	0.67	(0.10)	0.71	(0.13)	0.6838	(0.3174)
		Moderate	0.66	(0.10)	0.67	(0.05)	0.67	(0.07)	0.67	(0.04)	0.66	(0.04)	0.67	(0.04)	0.72	(0.04)	2.2193	(2.5988)
	CNN	Extreme	0.97	(0.01)	0.17	(0.13)	0.37	(0.19)	0.22	(0.13)	0.90	(0.01)	0.57	(0.06)	0.72	(0.08)	0.7877	(0.4112)
		Moderate	0.72	(0.07)	0.64	(0.04)	0.70	(0.08)	0.67	(0.03)	0.68	(0.02)	0.68	(0.03)	0.74	(0.04)	1.6882	(0.7615)

DISCUSSION

Regression analysis – the standard

Benchmark studies suggest the inconsistent performance of neural networks compared to standard GP methods, which depends on a series of factors. In this sense, we contrasted our findings to reference studies and explored how these factors might have affected the results. Concerning the regression analysis, the GBLUP method outperformed MLP for both traits. One of the factors that are reported to determine the best methodology is how modeling is performed. In this sense, GP was carried out as a two-stage analysis. Hence the genotypic value of hybrids across environments was obtained before prediction. Therefore, the environmental source of variation was absent and could not be captured by the ML models. For instance, it has been extensively shown that linear models (e.g., GBLUP or BMTME) tend to be outperformed by MLP in a multi-environmental joint analysis if the genotype by environment factor is not modeled for the prediction of PH and GY in maize. This holds under both single (Montesinos-lópez et al. 2018) and multi-trait (Montesinos-López et al. 2018) modelling contexts. Accordingly, MLP was outperformed by GBLUP for both traits in our study, supporting the suggested effect of modeling to the comparative outcome for the studied GP methods.

The use of CNN presented better results than GBLUP and MLP for predicting GY. This contrasts the findings of Azodi et al. (Azodi et al. 2019), who suggested that ridge regression BLUP, a GBLUP-equivalent method, outperforms both MLP and CNN for predicting several traits on numerous crops, including PH and GY in maize inbred lines. In this case, CNN was reported to be the poorest performing method for both traits. The inconsistency between the results of these studies regarding the performance of the CNNs for GY could be attributed to the restrictive search space for hyperparameters given the computational requirements for the analysis of an astonishing amount of studied traits and species by Azodi et al. (Azodi et al. 2019); we tailored ML models to each scenario within each trait, which is known to improve NN performance (Montesinos-López et al. 2018). Another factor that might have led to this discrepancy was the use of pre-processed genomic information (genomic images) in our CNNs, while they opted for using the raw genomic matrix. At last, inbred lines were used in their work, while hybrids were used in ours; studies suggested that CNNs tend to have better performance (than linear methods) when strong nonlinear (e.g., dominance) effects are present (Bellot et al. 2018; Abdollahi-Arpanahi et al. 2020); which is the case of GY in population we studied (Alves et al. 2019).

Regarding the underperformance of NN methods at predicting PH, tangible reasons could be pointed out. It has been hypothesized that the occurrence of extreme allelic frequencies (e.g., only two genotypes are present for a given *locus*) favors linear models by enabling the capture dominance and epistatic variance (Azodi et al. 2019); however, this does not hold for this dataset (Morosini et al. 2020). Also, PH is predominantly governed by additive allelic interactions (Alves et al. 2019), which enables linear models to capture a considerable proportion of the genotypic variance; nevertheless, regardless of the nature of the effects governing the traits under study, ML should always be (at least) as good as linear models given their ability to model linear relationships (Azodi et al. 2019), which was not the case. At last, an unknown and unverifiable cause could be the number of training samples, which might not have been enough for modeling linear and nonlinear interactions between markers by the NN (Montesinos-López et al. 2020). This is a problem of common occurrence in plant breeding given the usually low number of samples, a large number of markers, and heterogeneity of data (Abdollahi-Arpanahi et al. 2020; Pook et al. 2020).

Overall, CNN presented better results compared to MLP. This advantage might have been due to the processing of the genomic matrix into the additive GRM, in the case of MLP. In this case, only the linear relationship between genotypes was modeled. This might reduce the potential of MLP to identify nonlinear effects since they might have been lost during processing. The genomic matrix could be used for further studies at the expense of computational time to overcome this issue.

Classification analysis – the alternative

Given their complex genetic nature, most plant traits present continuous phenotypes. As a matter of fact, traits that were previously discretized by means of measurement ease, such as resistance to biotic stresses, have had their continuous nature better explored by high-throughput phenotyping (Galli et al. 2020). Therefore, regression tasks are an adequate fit for genetic analysis, including GP. Nevertheless, plant breeding is globally a classification problem in which genotypes are assigned classes (Ornella et al. 2014), usually selected and non-selected. Hence, we elaborate on this problem, unifying ranking and selection by using classifying predictors. The evaluation of these prediction machines was performed using many metrics that evaluate the model's ability to distinguish which genotypes should be selected.

A critical step on classification tasks is the discretization of the continuous variable; when applicable. Discretizing traits has its inherent degree of subjectivity, regarding, e.g., the number of classes and which threshold values are used to classify the data (Montesinos-López et al. 2019). Accordingly, these choices have

been reported to influence the performance of prediction models (Ornella et al. 2014; González-Camacho et al. 2016). Furthermore, a greater level of subjectivity is introduced when genotype classification as true positives or negatives before the prediction is based on the empirical distribution rather than the absolute value of the trait. Predicting genotypes from a related population, classification would not be tied to the percentiles of the distribution on the training population but to the genotypic values and genetic variants under each class and the genetic similarities across populations. In this sense, the algorithm might be targeting, e.g., plants with a height between 1.95 and 2.05 m, but not the 10% or 50% best yielding hybrids since the distribution of genotypic values of a new population is likely to differ from the training population.

The classification problem was approached considering two scenarios: one highly imbalanced, where the size of the classes differed substantially (extreme SI), and one nearly balanced, where each class contained about half of the individuals (moderate SI). Both scenarios are plausible and of common occurrence in plant breeding, depending on the program stage. Nevertheless, imbalanced datasets should be evaluated with further cautiousness (Fernández et al. 2011). TNR, precision, recall, F1 score, accuracy, and AUC are examples of metrics sensitive to class imbalance, meaning their results might not be directly interpretable for comparing predictions with differing selection intensities. This is also evidenced by the discrepancy between the accuracy and the balanced accuracy at extreme selection intensity. The selection intensities presented little influence over the balanced accuracy for the same trait and independent variable, except for GY when images were used.

The balanced accuracy is calculated by averaging the proportion of correct predictions in each class, meaning that the label (selected or non-selected) is not relevant. This metric varied from 0.55 to 0.61 for PH and from 0.57 to 0.68 for GY. These results are inconsistent with the regression analysis, which showed higher predictability for PH according to all metrics, following the higher heritability of this trait. We postulate that this is associated with the region of the empirical density of genotypic values from which genotypes were regarded as "selected". For PH, the distinction between the best and the worst individuals was non-directional, which might have difficulted the distinction between which hybrids should or not be selected by the models. For GY, as the selection is directional, this was not an issue. Overall, balanced accuracies were closer to 0.5 (random guess) than to 1 (all correct) for both traits, meaning that further improvements are required. Nevertheless, our results suggest the possibility of non-directional selection, as for PH, which is highly relevant for breeding programs.

Unlike the regression task, where the use of CNN usually presented the best results between machine learning methods, there was considerable inconsistency regarding the superiority of MLP or CNN in the classification task. The comparative performance of the neural network methodologies seemed highly conditioned to trait and selection intensity. Generally, MLP presented the best results for PH, while for GY, the best method heavily depended on the selection intensity. In this sense, it is reasonable to assume that the discretization process of PH and GY impacted the performance of CNN more than that of MLP; but further investigation is warranted. GP prediction regression tasks with machine learning models are already common, but studies comparing methods for predicting discretized variables are still limited.

Choosing machine learning architectures

The choice of the neural network hyperparameters has been reported to be a critical step for NN-based GP by extensive benchmarking (Azodi et al. 2019). Ergo, network search for a given task and dataset has been applied in recent ML-based GP studies (Montesinos-lópez et al. 2018; Montesinos-López et al. 2018; Azodi et al. 2019; Abdollahi-Arpanahi et al. 2020). However, model tuning has been primarily performed using naïve approaches such as random (values sampled from distribution) or grid (discrete values) search, which may limit the number of hyperparameter combinations based on a set of user-defined *a priori* information (Jin et al. 2019). Due to recent advancements in computer science and technology, less restrictive, free, and easy-to-use hyperparameter search algorithms have been made available. In this sense, we used Auto-Keras, an AutoML search algorithm with Bayesian optimization to identify (suitable) models. Overall, the algorithm yielded adequate performing neural networks despite the absence of the commonly required human intervention for adjustments (Azodi et al. 2020).

It has been previously reported that different neural network hyperparameters can be obtained from network search algorithms for a given task (Bellot et al. 2018; Huang et al. 2020). The neural networks selected by the AutoML algorithm presented idiosyncrasies within replications of the same scenario (File S1). The lack of similarity between structures might arise from the ability of AutoML to adapt the network to the dataset (Jin et al., 2019), which changes due to sampling in repeated validation. Huang et al. (Huang et al. 2020) suggest this event to be a consequence of insufficient data, but further confirmation is required. Additionally, this may also be associated with the sampling nature of the hyperparameter search system (Jin et al., 2019). Despite the inconsistency between structures, systematic regularities are suggested by the within scenario low standard deviation of the estimated metrics (Table 1, Table 2). Hence, the networks might be capturing similar features,

yielding consistent predictions. This has relevant implications for the choice of (deep) neural networks, meaning that distinct but adequate network structures result in similar outcomes.

Further observations can be drawn from the chosen network structures: *i)* Although limited, the cases where structures did match (within and between scenarios) suggest that: type of task (regression or classification) is determinant over structure since most matches were of the same type; matches across traits were common, suggesting that similar sources of information might have been captured, which is probably intrinsically associated to the genetic correlation between PH and GY in maize. *ii)* Also, when images were used as the input for prediction, augmentation algorithms (e.g., resize, flip, rotation) were allocated in the structure of about half of the chosen models despite the spatial structure in the genomic images created by the decomposition performed by *DeepInsight*; further inferences on this matter would require studying the implications of such algorithms to the original images (Azodi et al. 2020), which is not in the scope of this study. *iii)* At last, the depth and number of parameters of the networks within scenarios were highly variable for both MLP and CNN, suggesting that simple architectures were as effective as the complicated ones. Simpler models also have the additional advantage of being less prone to overfitting and are generally quicker to train (van Dijk et al. 2021). *iv)* Overfitting, which is the tendency of a model to perform well on training but not on unseen data (van Dijk et al. 2021), was not an apparent issue of the chosen models. No drastic changes were observed (in average, within scenario) between the metrics of inner training, inner validation, and outer validation sets for both types of tasks (regression and classification). The dropout regularization, temporarily setting a percentage of random neurons to zero (Srivastava et al. 2014), was present on 2/3 of the chosen models, presumably handling the overfitting problem (Montesinos-lópez et al. 2018).

Further considerations

Overall, based on the empirical and experimental evidence, neural networks are especially competitive under the presence of strong nonlinear factors and interactions and hidden relationships between pieces of information. Accordingly, it is also dependent on the population type (e.g., lines or hybrids) and the consequent, non-mutually exclusive, genetic architecture of the trait (Bellot et al. 2018; Abdollahi-Arpanahi et al. 2020). The performance of NN is certainly conditioned on the choice of hyperparameters (Bellot et al. 2018; Zingaretti et al. 2020) and neural network type (MLP or CNN). It depends on how the input data is processed before prediction; in this sense, special attention should be given to this step since valuable information could be lost. Also, it is presumably dependent on the number of samples and the sample to parameter ratio (Montesinos-

López et al. 2018; Azodi et al. 2019; Pérez-Enciso and Zingaretti 2019; Abdollahi-Arpanahi et al. 2020). Therefore, it is the scientist's discretion to test and identify the best performing method for their task. To this day, the only identified consistency regarding GP benchmarking is that no model performs best for all situations.

From experience, inferences on the use of images for GP could be drawn. In the original work by Sharma et al. (Sharma et al. 2019), *DeepInsight* was used for transforming RNA-seq, text, and artificial datasets into images. Our work is the first to apply such methodology in a GP context, and it is noteworthy that: *i)* the algorithm can create images of different sizes. Image size, which is a hyperparameter, should be adapted to the available dataset and computational power. With the increasing size of the genomic matrix, there is a greater chance that a considerable amount of information would be lost as correlated markers would be tightly grouped, so larger images should be used (Sharma et al. 2019). Additionally, increasing the size of images consequently increases the number of parameters estimated in the neural network, requiring greater computational power. In this work, using 120 by 120 images seemed to be an adequate fit for ~ 30,000 genomic markers; *ii)* different dimensionality reduction techniques can be used: t-SNE and kPCA are implemented in the algorithm, but any other of interest can be implemented; further testing should elaborate on this matter; *iii)* images can have multiple channels: neural networks can model linear and nonlinear relationships between neurons, including other effects and layers of data, such as dominance, epistasis, g × e, transcriptome, and so on.; *iv)* the cost-benefit in terms of predictive gain and additional work, the use of images as input is arguable. Nevertheless, the methodology's potential for GP is unprecedented; *v)* simulations should provide new valuable and unbiased information.

At last, we discussed two prediction alternatives: regression and classification. Under the regression context, MLP and CNN presented competitive results. Under the classification context, we expected better performances. Nevertheless, we believe that the latter has great potential for plant breeding since it simplifies the pipeline. Neural networks are self-adaptable and aimed at prediction alone. This statement implies that understanding and exposing the events underlying the relationship between phenotypes and genotypes are not of particular interest but could be done if necessary (Azodi et al. 2020). This also implies that limited genetic knowledge of the trait is not a constraint for prediction. Coupled with a simpler processing, direct classification opens new possibilities regarding selecting traits where the ideotype points to intermediate phenotypes, e.g., plant height, ear height, and flowering time (under some circumstances) in maize. Hence, we believe this methodology deserves attention since it could further enhance the GP pipeline in breeding programs.

568 REFERENCES

569 Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble
570 methods for genomic prediction of complex phenotypes. Genet Sel Evol. doi: 10.1186/s12711-020-
571 00531-z

572 Alves FC, Galli G, Matias FI, et al (2021) Impact of the complexity of genotype by environment and dominance
573 modeling on the predictive accuracy of maize hybrids in multi-environment prediction models. Euphytica
574 217:37. doi: 10.1007/s10681-021-02779-y

575 Alves FC, Stefanine Í, Granato C, et al (2019) Bayesian analysis and prediction of hybrid performance. Plant
576 Methods 1–18. doi: 10.1186/s13007-019-0388-x

577 Azodi CB, Bolger E, Mccarren A, et al (2019) Benchmarking Parametric and Machine Learning Models for
578 Genomic Prediction of Complex Traits. 9:3691–3702. doi: 10.1534/g3.119.400498

579 Azodi CB, Tang J, Shiu S (2020) Opening the Black Box : Interpretable Machine Learning for Geneticists.
580 Trends Genet 1–14. doi: 10.1016/j.tig.2020.03.005

581 Bellot P, de los Campos G, Pérez-Enciso M (2018) Can deep learning improve genomic prediction of complex
582 human traits? Genetics 210:809–819. doi: 10.1534/genetics.118.301298

583 Bernardo R (1994) Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related
584 Hybrids. Crop Sci 34:20. doi: 10.2135/cropsci1994.0011183X003400010003x

585 Blondel M, Onogi A, Iwata H, Ueda N (2015) A Ranking Approach to Genomic Selection. PLoS One
586 10:e0128570. doi: 10.1371/journal.pone.0128570

587 Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation PLINK: rising to the challenge of larger and
588 richer datasets. Gigascience 4:7. doi: 10.1186/s13742-015-0047-8

589 Costa-Neto G, Fritsche-Neto R, Crossa J (2020) Nonlinear kernels, dominance, and envirotyping data increase
590 the accuracy of genome-based prediction in multi-environment trials. Heredity (Edinb). doi:
591 10.1038/s41437-020-00353-1

592 de Los Campos G, Gianola D, Rosa GJ (2009) Reproducing kernel Hilbert spaces regression: a general
593 framework for genetic evaluation. J Anim Sci. doi: 10.2527/jas.2008-1259

594 Fernández A, García S, Herrera F (2011) Addressing the Classification with Imbalanced Data: Open Problems
595 and New Challenges on Class Distribution. In: Corchado E, Kurzyński M, Woźniak M (eds) Hybrid
596 Artificial Intelligent Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–10

597 Feurer M, Klein A, Eggensperger K, et al (2015) Efficient and robust automated machine learning. In: Advances
598 in Neural Information Processing Systems.
599 Fristche-neto R, Akdemir D, Jannink J (2018) Accuracy of genomic selection to predict maize single-crosses
600 obtained through different mating designs. 131:1153–1162. doi: 10.1007/s00122-018-3068-8
601 Fritsche-Neto R, Galli G, Mendonça L de F, et al (2019) USP tropical maize hybrid panel. Mendeley Data 3:1–
602 15. doi: 10.17632/tpcw383fkm.3
603 Galli G, Horne DW, Fritsche-neto R, Rooney WL (2020) Optimization of UAS-based high-throughput
604 phenotyping to estimate plant health and grain yield in sorghum. 1–14. doi: 10.1002/ppj2.20010
605 Galli G, Lyra DH, Alves FC, et al (2018) Impact of phenotypic correction method and missing phenotypic data
606 on genomic prediction of maize hybrids. Crop Sci 58:1481–1491. doi: 10.2135/cropsci2017.07.0459
607 Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International,
608 Hemel Hempstead
609 González-Camacho JM, Crossa J, Pérez-Rodríguez P, et al (2016) Genome-enabled prediction using
610 probabilistic neural network classifiers. BMC Genomics 17:1–16. doi: 10.1186/s12864-016-2553-1
611 Granato ISC, Galli G, de Oliveira Couto EG, et al (2018) snpReady: a tool to assist breeders in genomic analysis.
612 Mol Breed 38:102. doi: 10.1007/s11032-018-0844-8
613 Huang GH, Lin CH, Cai YR, et al (2020) Multiclass machine learning classification of functional brain images
614 for Parkinson's disease stage prediction. Stat Anal Data Min 508–523. doi: 10.1002/sam.11480
615 Jin H, Song Q, Hu X (2019) Auto-Keras : An Efficient Neural Architecture Search System. 1946–1956. doi:
616 10.1145/3292500.3330648
617 Kotthoff L, Thornton C, Hoos HH, et al (2017) Auto-WEKA 2.0: Automatic model selection and
618 hyperparameter optimization in WEKA. J Mach Learn Res. doi: 10.1007/978-3-030-05318-5_4
619 Lyra DH, de Freitas Mendonça L, Galli G, et al (2017) Multi-trait genomic prediction for nitrogen response
620 indices in tropical maize hybrids. Mol Breed 37:80. doi: 10.1007/s11032-017-0681-1
621 Lyra DH, Granato ÍSC, Morais PPP, et al (2018) Controlling population structure in the genomic prediction of
622 tropical maize hybrids. Mol Breed. doi: 10.1007/s11032-018-0882-2
623 Ma W, Qiu Z, Song J, et al (2018) A deep convolutional neural network approach for predicting phenotypes
624 from genotypes. Planta 248:1307–1318. doi: 10.1007/s00425-018-2976-9
625 Matias FI, Galli G, Correia Granato IS, Fritsche-Neto R (2017) Genomic Prediction of Autogamous and

626 Allogamous Plants by SNPs and Haplotypes. Crop Sci 57:2951. doi: 10.2135/cropsci2017.01.0022
627 Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense
628 marker maps. Genetics 157:1819–1829. doi: 11290733
629 Montesinos-lópez A, Montesinos-lópez OA, Gianola D, et al (2018) Multi-environment Genomic Prediction of
630 Plant Traits Using Deep Learners With Dense Architecture. 8:3813–3828. doi: 10.1534/g3.118.200740
631 Montesinos-López OA, Martín-Vallejo J, Crossa J, et al (2019) New Deep Learning Genomic-Based Prediction
632 Model for Multiple Traits with Binary, Ordinal, and Continuous Phenotypes. G3:
633 Genes|Genomes|Genetics 9:1545–1556. doi: 10.1534/g3.119.300585
634 Montesinos-López OA, Montesinos-López A, Crossa J, et al (2018) Multi-trait, Multi-environment Deep
635 Learning Modeling for Genomic-Enabled Prediction of Plant Traits. G3 (Bethesda) 8:3829–3840. doi:
636 10.1534/g3.118.200728
637 Montesinos-López OA, Montesinos-López JC, Singh P, et al (2020) A multivariate poisson deep learning model
638 for genomic prediction of count data. G3 Genes, Genomes, Genet 10:4177–4190. doi:
639 10.1534/g3.120.401631
640 Morosini JS, Mendonça LDF, Vidotti MS, Fritsche-neto R (2017) Association mapping for traits related to
641 nitrogen use efficiency in tropical maize lines under field conditions. Plant Soil. doi: 10.1007/s11104-
642 017-3479-3
643 Morosini S, Fritsche-neto R, Id GG, Alves FC (2020) On the usefulness of parental lines GWAS for predicting
644 low heritability traits in tropical maize hybrids. 1–15. doi: 10.17632/tpcw383fkm.3
645 Ornella L, Pérez P, Tapia E, et al (2014) Genomic-enabled prediction with classification algorithms. Heredity
646 (Edinb) 112:616–626. doi: 10.1038/hdy.2013.144
647 Pérez-Enciso M, Zingaretti LM (2019) A guide for using deep learning for complex trait genomic prediction.
648 Genes (Basel) 10:1–19. doi: 10.3390/genes10070553
649 Pook T, Freudenthal J, Korte A, Simianer H (2020) Using Local Convolutional Neural Networks for Genomic
650 Prediction. Front Genet. doi: 10.3389/fgene.2020.561497
651 Sharma A, Vans E, Shigemizu D, et al (2019) DeepInsight: A methodology to transform a non-image data to
652 an image for convolution neural network architecture. Sci Rep 9:1–7. doi: 10.1038/s41598-019-47765-6
653 Sousa MB, Galli G, Lyra DH, et al (2019) Increasing accuracy and reducing costs of genomic prediction by
654 marker selection. Euphytica 215:18. doi: 10.1007/s10681-019-2339-z

Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: A simple way to prevent neural networks from overfitting.

Trevisan RG, Pérez O, Schmitz N, et al (2020) High-Throughput Phenotyping of Soybean Maturity Using Time Series UAV Imagery and Convolutional Neural Networks. doi: 10.20944/preprints202009.0458.v1

Truong A, Walters A, Goodsitt J, et al (2019) Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. In: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI.

Unterseer S, Bauer E, Haberer G, et al (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823. doi: 10.1186/1471-2164-15-823

van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D (2021) Machine learning in plant science and plant breeding. iScience 24:101890. doi: 10.1016/j.isci.2020.101890

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

Wimmer V, Albrecht T, Auinger HJ, Schon CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087. doi: 10.1093/bioinformatics/bts335

Zingaretti LM, Gezan SA, Ferrão LF V., et al (2020) Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. Front Plant Sci 11:1–14. doi: 10.3389/fpls.2020.00025

Figures

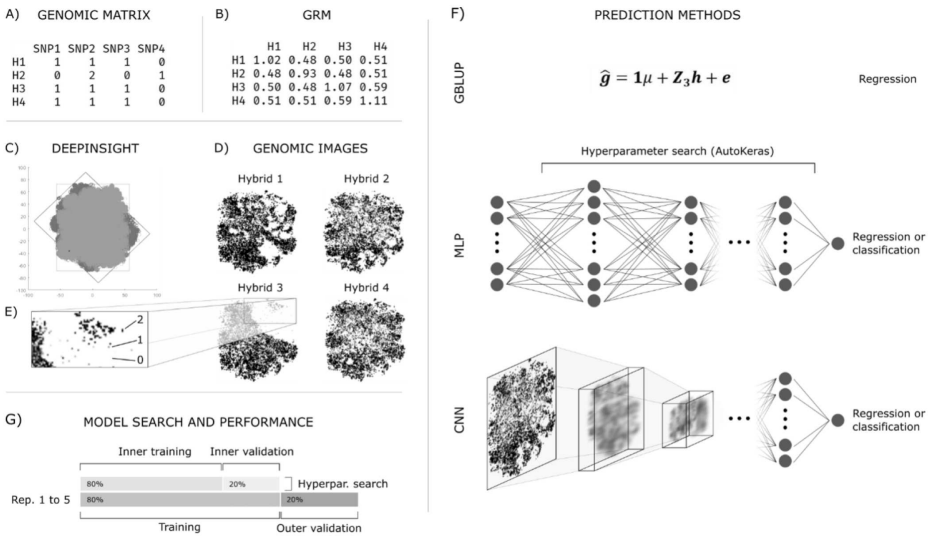

Figure 1

Summarized general exemplification of the employed methodology. A) Genomic data obtained from the pre-processing step; B) Additive Genomic Relationship Matrix (GRM) obtained with VanRaden's method using the genomic matrix (A); C) DeepInsight pipeline: t-SNE decomposition of the genomic matrix (A)

with original (dark gray) and rotated (light gray) marker coordinates; D) Genomic images obtained with DeepInsight; one image is produced for each hybrid and all markers are represented in the image; E) representation of the genotypes (0, 1, and 2, also white, light gray, and dark gray, respectively) in a genomic image; each pixel might comprise a single or multiple markers depending on the level of linkage disequilibrium in the population; F) Genomic prediction methods: genomic BLUP (GBLUP), multilayer perceptron (MLP), and convolutional neural network (CNN); GBLUP and MLP used the GRM (B) as independent variable, while CNN used genomic images (D); the neural networks were used as both regression and classification tasks; AutoKeras was used for hyperparameter search in MLP and CNN; G) Simplified representation of the nested validation procedure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigureS1.docx
- SupplementalFileS1.docx
- SupplementalTable1.docx