agronomy

MDPI

Article

Application of Artificial Neural Networks to Predict Genotypic Values of Soybean Derived from Wide and Restricted Crosses for Relative Maturity Groups

Lígia de Oliveira Amaral [1], Glauco Vieira Miranda [2,*], Jardel da Silva Souza [3], Alyce Carla Rodrigues Moitinho [3], Dardânia Soares Cristeli [3], Hortência Kardec da Silva [3], Rafael Silva Ramos dos Anjos [3], Luis Fernando Alliprandini [4] and Sandra Helena Unêda-Trevisoli [3]

[1] BASF Porto Nacional Soybean Station, Porto Nacional 77500-000, Tocantins, Brazil; ligiaoamaral@hotmail.com
[2] Department of Agronomy, Federal Technological University of Paraná, Santa Helena 85892-000, Paraná, Brazil
[3] Laboratory of Biotechnology and Plant Breeding, Department of Agricultural Sciences, São Paulo State University—UNESP/FCAV, Jaboticabal 14884-900, São Paulo, Brazil; jardel.souza@unesp.br (J.d.S.S.); acr.moitinho@unesp.br (A.C.R.M.); dardania.cristeli@unesp.br (D.S.C.); hortencia.silva@unesp.br (H.K.d.S.); rsranjos@gmail.com (R.S.R.d.A.); shu.trevisoli@unesp.br (S.H.U.-T.)
[4] Bayer Crop Science, Rolândia 86600-000, Paraná, Brazil; luis.alliprandini@bayer.com
* Correspondence: glaucovmiranda@utfpr.edu.br

Abstract: The primary objective of soybean-breeding programs is to develop cultivars that offer both high grain yield and a maturity cycle tailored to the specific soil and climatic conditions of their cultivation. Therefore, predicting the genetic value is essential for selecting and advancing promising genotypes. Among the various analytical approaches available, deep machine learning emerges as a promising choice due to its capability to predict the genetic component of phenotypes assessed under field conditions, thereby enhancing the precision of breeding decisions. This study aimed to determine the efficiency of artificial neural networks (ANNs) in predicting the genetic values of soybean genotypes belonging to populations derived from crosses between parents of different relative maturity groups (RMGs). We characterized populations with broad and restricted genetic bases for RMG traits. Data from three soybean populations, evaluated over three different agricultural years, were used. Genetic values were predicted using the multilayer perceptron (MLP) artificial neural network and compared to those obtained using the best unbiased linear prediction from variance components using restricted maximum likelihood (RR-BLUP). The MLP neural network efficiently predicted genetic values for the relative maturity group trait for genotypes belonging to populations of broad and restricted crosses, with an R^2 of 0.999 and root-mean-square error (RMSE) of 0.241, and for grain yield, there was an R^2 of 0.999 and an RMSE of 0.076. While the percentage of coincident superior genotypes remained relatively consistent, a significant difference was observed in their ranking order. The genetic gain with selection estimated using MLP was higher by 30–110% compared to RR-BLUP for the relative maturity group trait and 90–500% for grain yield. Artificial neural networks (ANNs) showed higher efficiency than RR-BLUP in predicting the genetic values of the soybean population. Local selection at intermediate latitudes is conducive to developing lines adaptable for regions at higher and lower latitudes.

Keywords: variance components; *Glycine max*; machine learning; mixed models; REML/BLUP

check for updates

Citation: Amaral, L.d.O.; Miranda, G.V.; Souza, J.d.S.; Moitinho, A.C.R.; Cristeli, D.S.; Silva, H.K.d.; Anjos, R.S.R.d.; Alliprandini, L.F.; Unêda-Trevisoli, S.H. Application of Artificial Neural Networks to Predict Genotypic Values of Soybean Derived from Wide and Restricted Crosses for Relative Maturity Groups. *Agronomy* **2023**, *13*, 2476. https://doi.org/10.3390/agronomy13102476

Academic Editor: Gniewko Niedbała

Received: 22 July 2023
Revised: 4 September 2023
Accepted: 15 September 2023
Published: 25 September 2023

1. Introduction

Soybean is cultivated worldwide in different latitudinal zones. In Brazil, this model was adapted for local conditions and is presently used by all South American public and private breeding companies [1]. It is recognized as a species with a narrow genetic base, which can hinder the acquisition of genetic sources for economically important traits and populations with significant genetic variability. While each cultivar has traditionally been

associated with a relatively narrow latitudinal zone, its adaptability to diverse producing regions stems from the genetic variability present in crucial gene loci and quantitative trait loci (QTLs) responsible for regulating flowering and maturity. The primary genetic loci, E1–E11 and J, and several QTLs, such as Tof11/Gp11, Tof12/Gp1/qFT12-1, and qDTF-J, have been identified. In general, except for the E6, E9, E11, and J genes, the dominant allele of the E genes confers late flowering and maturity, whereas an increase in the number of recessive alleles leads to early flowering [2,3]. The genetic diversity of soybean has been reduced, mainly because of genetic bottlenecks related to domestication. In contrast, its wild relative *Glycine soja*, which grows under various environmental conditions, has retained significant genetic diversity [4]. Furthermore, soybean lineages have undergone distinct and individual selection based on geographical location, with numerous highly conserved regions among cultivated varieties because of domestication [5]. In countries with vast continental dimensions, such as Brazil, which is the world's largest soybean producer, the segregation of soybean populations for genetic enhancement is often constrained by latitude, leading breeders to focus their efforts on similar genotypes. Consequently, crosses between cultivars with differing relative maturity groups can contribute to genetic diversity by exploring novel QTLs and loci of interest.

Genetic value prediction is paramount in genetic improvement programs as it exclusively encompasses the hereditary component of quantitative trait control that can be inherited by offspring. Consequently, acquiring insights into the genetic value of individuals constitutes a critical facet of breeding programs, ensuring the realization of genetic advancements [6]. During the selection phase, breeders must discern the genetic potential of individual candidates and make decisions regarding the improvement of specific genotypes, all grounded in empirical data. It is imperative to possess accurate selection predictions to systematically evaluate individuals within segregated populations and pinpoint superior genotypes [7].

Genetic variability is estimated using variance components utilizing the restricted maximum likelihood method (REML), and obtaining the best unbiased linear prediction (BLUP) of genetic values is preferred because it maximizes selective accuracy compared to parametric statistical methodologies [8]. Using parametric statistical methods necessitates assumptions related to the probability distribution of variables, often assuming the linear nature of the phenomenon under study. However, this can result in inefficiencies in the analysis since these ideal conditions are not always met when collecting data within genetic improvement programs.

In contrast to both parametric and non-parametric analyses, artificial neural networks (ANNs) offer distinct advantages that render them better suited for particular scenarios. Consequently, they play a valuable role in the selection and development stages, as described in [9], and exhibit a high predictive capacity, as demonstrated in [10]. Artificial neural networks (ANNs) are machine-learning models inspired by the human brain and have shown promise in various areas, such as pattern recognition, natural language processing, and computer vision. They can learn from raw data without prior knowledge of the domain or specific problems and handle incomplete or noisy data.

ANNs have been used in various areas of agriculture, such as the prediction of crop productivity [10–12], soil attributes, and image interpretation. ANNs have shown high efficiency in predicting genetic values compared to other methodologies in several studies [13–15]. In addition, methodologies using fractal analysis have been applied in plant breeding, reducing human error during the breeding process [16].

Therefore, the present study aims to evaluate the efficiency of ANNs in predicting genetic values of soybean genotypes derived from broad and narrow crosses for the trait of the relative maturity group (RMG). These results may provide critical information for developing new, more productive soybean varieties adapted to Brazilian conditions.

2. Materials and Methods

Three soybean populations with different territorial adaptabilities based on their relative maturity group (RMG) classification [1] were evaluated in the Soybean Breeding Program at the "Júlio de Mesquita Filho" State University in Jaboticabal, São Paulo.

The cross between the cultivars BMX Veloz (GMR 5.0) and BRS 278 RR (GMR 9.4) resulted in the "WRMG" population, characterized by wide adaptability and coverage in the critical photoperiod for latitudes between 23° LS (Subtropical) and 0° LS (Tropical). The "Brazilian" population comprised 220 F4, 252 F5, and 252 F6 progenies in 2017, 2018, and 2019, respectively.

The "Subtropical" population was established from the cross between BMX Energia (GMR 5.3) and BMX Potência (GMR 6.7) cultivars with a critical photoperiod for the southern region of Brazil, corresponding to latitudes 23° LS (Subtropical) and 20° LS (Tropical). The "Subtropical" population comprised 120 F5, 168 F6, and 168 F7 progenies in 2017, 2018, and 2019, respectively.

The "Tropical" population was obtained by crossing the cultivars BRS 245 RR (GMR 7.3) and BRS 278 RR (GMR 9.4) with a critical photoperiod for the northern region of Brazil, corresponding to latitudes of 20° LS (Tropical) and 0° LS (Tropical). The "Tropical" population comprised 60 F5, 60 F6, and 104 F7 progenies in 2017, 2018, and 2019, respectively.

The evaluations of the three populations across the three agricultural years took place at the Teaching, Research, and Extension Farm (FEPE) situated within São Paulo State University (UNESP) on the Campus of Jaboticabal (FCAV), São Paulo, Brazil. The location is positioned at a latitude of 21°15′19″ south and a longitude of 48°19′21″ west, with an altitude of 615 m. This region offers ideal photoperiod conditions for soybean genotypes, falling within the 6–8 growing degree month (GMR) range. This suitability is attributed to the prolonged rainy season in the region, spanning from November (spring) to April (fall), which permits the cultivation of soybean genotypes with a maturity cycle ranging from 90 to 150 days.

The control varieties for each population were their parents, as well as TMG 7262 RR (GMR 6.2), TMG 1174 RR (GMR 7.4), and TMG 1179 RR (GMR 7.9), for all agricultural years. The experimental design used was the augmented block design described by [17], in which the progenies were arranged in plots of a five-meter-long row with a spacing of half a meter between rows. Controls, parents of each population, and two other commercial cultivars were randomly allocated to each experimental block. The planting density was 15 seeds per meter, and all cultivation practices followed the technical recommendations for soybean culture [18]. Data were collected from five visually selected plants per plot.

The evaluated traits were total crop cycle (MATURITY), assessed by the number of days counted from germination until harvest at the R8 development stage [19]; grain yield (GY), obtained by the weight in grams of the grains from the five selected plants in each plot after harvest and processing; the number of days to flowering (NDF), counted from germination until full flowering of the field (stage R^2—[19]; the height of first pod insertion (AIV), i.e., the height measured from the ground to the first productive pod of the plant, in cm; and plant height at maturity (APM), i.e., height from the ground to the last fertile pod of the plant, also expressed in cm.

Data were analyzed for each population using R version 4.0.2 [20] via the mixed model approach proposed by [21].

The variance components were estimated using the restricted maximum likelihood (REML) approach [22]. Fixed effects were checked for significance using the F-test, and significance of variances associated with random effects was verified using the likelihood ratio test. Heritability, experimental coefficient of variation (CV), and selection accuracy (rgg) were calculated as described by [22].

The developed neural network is a multilayer perceptron (MLP) with an input layer, two hidden layers, and an output layer. The number of units in the input layer was composed of the five agronomic years, population identification, and the three agricultural years, totaling nine input layers. It was necessary to convert the categorical variables (years)

into a single-point representation. The input layers also included the population (POP) and three agricultural years.

The input layer has nine neurons, the hidden layers have 64 and 128 neurons, and the output layer has one neuron corresponding to MATURITY and the other corresponding to grain yield (GY).

The dataset consisted of 6158 examples. The MLP network was built in Python 3.6 using Keras as a front end, TensorFlow 2.3.0 as a back end, and Scikit-learn 0.22.2. The dataset was split into k, or ten partitions, where k-1 sections were used for training, and k was used to test the model (k-fold cross-validation). Thus, ten models were created in which the training and test data were changed for each iteration [23].

The final evaluation of the models was based on the correlation between the observed and predicted values using network (R^2) and RMSE parameters [24]. The selected activation function was a logistic sigmoid function. The output layer used the softmax function [25].

Backpropagation is the standard algorithm for updating the weights in this type of network and is used to train the network. Backpropagation is an efficient method for calculating the partial derivatives of each layer. The weights are updated using gradient descent, which aims to minimize the error produced by the network. The algorithm generally applies the data and passes it forward to the successive layers, called the "forward pass". Then, it calculates the error in the output layer and propagates it backwards, called the "backward pass". These steps are repeated until the error is as small as possible [26].

Stochastic gradient descent (SGD) is an efficient method for calculating gradient descent SGD [27]. In practice, the optimizers used are variations of the SGD. This study used an adaptive moment estimation (ADAM) optimizer [28]. The number of training cycles was set to 600. Care was taken to limit the number of iterations such that it did not become excessive, which could lead to a loss of generalization power.

The efficiency comparison in predicting genetic values between the mixed models REML/BLUP and artificial neural networks was performed using the coincidence index (%) of the 10 and 20% best genotypes for each trait, according to each methodology, and the gain with selection, considering a selection intensity of 20%. For the total cycle trait (MATURITY), genotypes with lower additive genetic values were selected to choose earlier genotypes, and for grain yield, genotypes with higher additive values were selected to increase the estimates.

3. Results

The "WRMG" population exhibited significant genetic variance for GY and MATURITY across the three agricultural years, corresponding to the various generations (as shown in Table 1). Five of the six estimated heritabilities fell within the medium-to-high range, spanning from 0.33 to 0.82. Regarding the MATURITY trait, the average ranged from 130 days in the F4 generation to 121 days in the F6 generation, following the selective process applied to the earliest plants. In the case of GY, there was a remarkable increase of over 100% by the conclusion of the selection process. In the F4 generation, the average GY per plant was 17 g, and this increased to 39 g per plant in the F6 generation.

The "Subtropical" population showed significant genetic variance for GY and MATURITY, except for GY in F6. The heritabilities of the six estimates were moderate to high, ranging from 0.31 to 0.68. The "Subtropical" population with earlier and less productive progenies in F5 was selected for the cycle of 115 days and 39 g per plant, which is ideal for the evaluated region.

The "Tropical" population presented significant genetic variance for GY only in the first generation and for MATURITY only in the last generation, starting from two non-significant generations. Only the heritability of GY and MATURITY in generation F7 and GY in F6 were medium to high. The "Tropical" population with lines with high maturities in F5 (131 days) and optimal productivity (39 g/plant) was selected for lower maturities, with 104 days in generation F7 and a productivity of 35 g/plant.

Table 1. Genetic and phenotypic parameter estimates and means for MATURITY (MAT, days) and grain yield (GY, g/plant) of soybean genotypes belonging to the "WRMG", "Subtropical", and "Tropical" populations, in three agricultural years in Jaboticabal, SP, and Brazil, respectively.

Parameters	Population Brazil					
	F4—2017		F5—2018		F6—2019	
	MAT	GY	MAT	GY	MAT	GY
σ^2_g	17.96 *	22.10 **	26.22 **	141.08 **	28.38 **	97.74 **
σ^2_e	80.11	10.95	19.64	28.27	28.87	66.05
h^2	0.82	0.33	0.43	0.17	0.50	0.40
Accuracy (%)	90.5	57.4	65.5	41.2	70.7	63.2
CV (%)	6.9	19.3	3.3	16.2	4.4	21.0
Mean	130	17	133	33	121	39
Parameters	Population South					
	F5—2017		F6—2018		F7—2019	
	MAT	GY	MAT	GY	MAT	GY
σ^2_g	82.81 **	15.84 **	27.81 **	36.57 ^{ns}	15.41 **	51.07 **
σ^2_e	73.66	9.66	8.64	79.22	6.91	60.14
h^2	0.47	0.38	0.24	0.68	0.31	0.54
Accuracy (%)	68.60	61.60	49.00	82.50	55.70	73.50
CV (%)	8.20	20.10	2.50	29.10	2.30	19.70
Mean	104	15	119	31	115	39
Parameters	Population North					
	F5—2017		F6—2018		F7—2019	
	MAT	GY	MAT	GY	MAT	GY
σ^2_g	19.19 ^{ns}	130.60 **	0.00 ^{ns}	69.13 ^{ns}	33.15 **	15.09 ^{ns}
σ^2_e	5.24	3.44	25.43	157.16	15.30	45.80
h^2	0.21	0.03	0.00	0.69	0.32	0.75
Accuracy (%)	45.8	17.3	0.00	83.10	56.6	80.60
CV (%)	1.70	4.70	3.80	23.00	3.70	19.60
Mean	131	39	132	54	104	35

**/*/^{ns} Significant at 1%, 5% probability, and non-significant, respectively, using the maximum likelihood ratio test; σ2g: genetic variance; σ2e: environmental variance; h2: heritability; CV: environmental coefficient of variation.

Experimental precision, verified through accuracy and environmental coefficient of variation (CV) estimators, varied among generatio no ns and evaluated traits. Accuracy was higher for MATURITY than for GY for the "Brazilian" population. For the "Subtropical" and "Tropical" populations, the accuracy estimates for generations F7 and F6 were higher for GY than for MATURITY.

The coefficient of environmental variation was consistently higher for GY than for MATURITY in the same year, which was consistent with these traits when evaluated on a per plant basis and not on the plot mean. In addition, CVs are inherent to the traits themselves.

The correlation estimates (R^2) obtained using the MLP algorithm of the ANN between the observed and predicted data exceeded 0.999, indicating a remarkably high predictive capacity for both GY and MATURITY, as detailed in Table 2. The models exhibited the lowest RMSE values for GY, with 0.077 during training and 0.076 during validation. For MATURITY, these values were 0.2407 during training and 2.6106 during validation. It is worth noting that these RMSE values share the same units as the variables under investigation. The overall mean GY of 33.56 g per plant corresponds to an error of merely 0.46%. In the case of maturity, where the mean was 121 days, the error in the validated models amounted to 4.2%.

Table 2. Performance of artificial neural networks in the training and validation phase with estimates of RMSE and correlation (R^2) between observed and predicted values using the artificial neural network for grain yield (GY) and maturity (MAT).

	Training			**Validation**		
	Loss Function	**RMSE**	**R^2**	**Loss Function**	**RMSE**	**R^2**
1	0.006	0.077	0.99994	0.007	0.083	0.99996
2	0.013	0.113	0.9999	0.014	0.118	0.99992
3	0.011	0.104	0.99995	0.013	0.113	0.99995
4	0.012	0.108	0.99994	0.011	0.107	0.99994
5	0.007	0.084	0.99992	0.008	0.09	0.99991
6	0.009	0.093	0.99992	0.009	0.096	0.99993
7	0.011	0.103	0.99989	0.015	0.121	0.99988
8	0.006	0.077	0.99994	0.006	0.076	0.99995
9	0.011	0.106	0.99989	0.012	0.112	0.99987
10	0.005	0.073	0.99994	0.006	0.077	0.99993
	Loss Function	RMSE	R^2	Loss Function	RMSE	R^2
1	0.0713	0.267	0.9993	52.533	22.920	0.9625
2	0.0699	0.2644	0.9992	40.269	20.067	0.9706
3	0.1329	0.3646	0.9983	30.974	17.599	0.9743
4	0.0579	0.2407	0.9994	68.127	26.101	0.9608
5	0.079	0.2811	0.9993	53.709	23.175	0.9683
6	0.0865	0.2941	0.9992	19.446	13.945	0.9857
7	0.0639	0.2529	0.9994	31.333	17.701	0.9756
8	0.0731	0.2703	0.9992	57.704	24.022	0.9667
9	0.0677	0.2602	0.9993	39.577	19.894	0.9637
10	0.1528	0.3909	0.9989	60.132	24.522	0.9537

The predictive capacity through artificial intelligence provided by MLP-ANN was superior to that based on RR-BLUP in predicting the genetic value of plants for both grain yield (GY) and MATURITY, as shown in Table 3.

Table 3. Predicted mean values for grain yield and maturity (MAT).

	MAT	**GY**	**Mean MAT**	**Mean GY**
ANN-MLP Validation	2.61	0.08	132.83	32.82
RR-BLUP	10.68	9.76	128.03	29.56
RR-BLUP WRMG	6.02	6.05	127.64	30.67
RR-BLUP Subtropical	4.00	7.02	113.78	29.99
RR-BLUP Tropical	4.56	2.85	117.49	36.67

The similarity in the classification of the best genotypes indicated by both methodologies was observed using the coincidence index with two percentages (Table 4). For the MATURITY variable, the percentages ranged from 30.77% (F5 population "Subtropical") to 100% (F7 population "Subtropical"), considering the selection intensity of 10%, and from 63.16% (F4 population "WRMG") to 92% (F5 population "WRMG"), considering the selection intensity of 20%. This demonstrates that, for MATURITY, a lower selection intensity may allow for similar genetic gains. For GY, the lowest coincidence for the 10% intensity occurred for F5 in the "WRMG" population (68.18%). The highest for F4 in the "WRMG" population was 89.47%. Considering a selection intensity of 20%, the coincidence percentages were 68.18% (F7 population "Tropical") and 87.50% (F5 population "Subtropical"). This demonstrates that genetic gains for GY can be similar when applying both methodologies.

Table 4. Percentage of coincidence between the 10% and 20% best genotypes for MATURITY (MAT) and grain yield (GY) according to BLUP and ANN for three soybean populations in 2017, 2018, and 2019, in Jaboticabal, SP, and Brazil.

Population	10%					
	2017		2018		2019	
	MAT	GY	MAT	GY	MAT	GY
WRMG	84	89	82	68	80	80
Subtropical	31	77	76	71	100	78
Tropical	40	80	-	80	73	82
Population	**20%**					
	2017		2018		2019	
	MAT	GY	MAT	GY	MAT	GY
WRMG	63	84	79	84	92	80
Subtropical	68	87	91	85	91	83
Tropical	70	70	-	70	91	68

It was possible to observe a similarity between the genotypes indicated as the best by both methodologies, although there was a significant divergence in the ranks they occupied. For the trait MATURITY in the "Tropical" population in 2018, there was no ordering of the best genotypes due to zero genetic variance by the analysis using mixed models. According to the ANN analysis, even with minor differences, genotypes were ordered.

The expected gains from the selection, considering an intensity of 20%, are listed in Table 5. For GY, the highest gains were obtained by the progenies ranked according to the artificial neural network prediction, reaching 11.91% for the "Tropical" population, while for the BLUP prediction, the highest gain was 4.43% for the "WRMG" population. For MATURITY, in which the goal is to reduce the total crop cycle, differences were observed between the methodologies, with the highest reduction being −5.42 (ANN, "WRMG" population) and the lowest reduction being −1.49 (BLUP, "Subtropical" population). The expected gains with ANN-MLP compared to RR-BLUP were 30–110% higher for MATURITY and 90–500% higher for grain productivity.

Table 5. Expected gains with selecting the top 20% progenies for MATURITY—MAT (expressed in days) and grain yield—GY (expressed in g/plant) in the agricultural year 2019.

Population	BLUPs		ANN	
	MAT	GY	MAT	GY
WRMG	−2.53	4.43	−5.42	8.47
SUBTROPICAL	−1.49	3.81	−1.99	9.35
TROPICAL	−1.64	1.98	−2.57	11.91

4. Discussion

The results of the predictive capability of ANN-MLP reveal its aptitude for capturing common nonlinear interactions in quantitative genetic traits. This proficiency stems from its capacity to account for non-additive effects, particularly in the context of grain yield and MATURITY. Importantly, the predictive ability of ANN-MLP extends to soybean populations exhibiting both wide genetic variability and a narrow genetic base. Various authors have demonstrated the superiority of ANN over mixed models. For instance, a study on flowering traits in beans [29] highlighted the effectiveness of ANN. Similarly, simulated data were used to showcase how ANN excels in capturing epistatic effects [30].

The MLP neural network was used to estimate soybean yield through its production components, such as plant height (PH), number of branches per plant (B), number of pods per plant (P), number of seeds per pod (S), and weight of 1000 seeds (WTS) [31]. In this

supervised training MLP neural network, the correlation was 0.848 for the validation of grain productivity, with considerable accuracy, using the information on the agronomic traits of the plant, growth habits, and population density of soybean crops. According to [10], MLP has proven to be more efficient in using a relatively small dataset and generalist or unsupervised problems; furthermore, MLP has efficiency for one or few layers, as well as shallow neural networks. The best RNA model tested was highly accurate and able to correctly classify all genotypes, replicating the selection made by the geneticist during the BLUP simulation [32]. This indicates that ANN can be a valuable tool in plant breeding, assisting in the selection of genotypes with greater efficiency and accuracy.

Corn productivity was predicted using an artificial neural network and the construction of multilayer perceptron (MLP) models using public data and experimental networks of corn [10]. The models with data imputation were more accurate than those without imputation, and the model with climatic data/SWB had the lowest RMSE of 71 kg ha^{-1}.

ANN has also been used to predict soybean and maize yields by comparing the prediction capacities of models at the state, regional, and local levels; it was concluded that the ANN models for maize had a correlation of 0.877 and an RMSE of 1036 kg/ha, and for soybean, the correlation was 0.64 and the RMSE was 1356 kg/ha [33].

The partial similarity in indicating the best genotypes for MATURITY and GY between the two prediction methodologies indicates the high efficiency of ANN as the prediction by mixed models is based on assumptions and considers several genetic and environmental parameters [15].

The R^2 values were 4 times higher for RR-BLUP than for ANN-MLP validation for days to maturity and 128 times higher for yield per plant, showing the efficiency of ANN-MLP compared to RR-BLUP, according to [29]. The same authors also identified the efficiency of ANN-MLP compared with RR-BLUP in predicting the capacity for flowering traits in black beans.

The efficiency of the predictive model created using the neural network was verified according to the R^2 parameter (correlation between observed and predicted values), which can range from 0 to 1, indicating a higher correlation the closer it is to 1; meanwhile, regarding the RMSE parameter (root-mean-squared error), which can range from 0 to 1, it indicates a lower error and higher efficiency the closer it is to 0. The high positive estimates of R^2 (above 0.998) and the low magnitude of RMSE for maturity (0.241) indicate good accuracy of the model, a low magnitude of error, and no tendency to over- or under-predict values.

The results obtained corroborate the results from other studies that neural networks, unlike traditional REML/BLUP models, allow the capture of nonlinear relationships from data information, and thus more effectively capture the non-additive effects associated with genetic control of productivity and maturity traits, as for other traits, such as flowering in bean cultivars [29,34]. DNN is applied with a huge dataset to adjust the artificial neural network and several hidden layers [11], such as for convolutional and other neural networks. The popular BP, RBF, GN, GRNN, SVM, and SVR models, as well as MLP, traditionally use numerical data and one or two hidden layers, which are suitable for more specific situations.

Autogamous species, such as soybean, exhibit non-additive or epistatic effects, owing to their high level of homozygosity, which is observed in different species, such as common beans, rice, barley, and sorghum [29]. Therefore, when parametric models are used, the prediction of genetic values for both MATURITY and grain productivity may have low accuracy.

The variation in genotype rankings can be attributed to the neural networks' capacity to comprehend intricate data traits and rely on experiential knowledge for genetic value predictions. This unique feature of neural networks also clarifies the ordering of the F5 genotypes in the northern population, even in the absence of genetic variance, as determined by REML/BLUP analysis.

The lower coincidence percentages observed in the populations during the first year compared to those in the third year for the MATURITY trait can be attributed to greater variability within populations during early generations, in which segregation processes are still ongoing. This heightened variability results in divergent rankings when employing each methodology.

Based on the estimates of the expected gains with selection, it is possible to predict the success of selecting specific populations. The neural network was always superior to the RR-BLUP for traits and all populations.

MLP has been previously applied in different areas, such as weed science [35] or drought tolerance [36]. Soybean productivity has also been estimated using various machine-learning algorithms, such as multilayer perceptron, support vector machine (SVM), and random forest (RF), using spectral reflectance data [37]. The authors concluded that the MLP is efficient for soybean breeding.

Developing increasingly productive and resistant cultivars depends directly on the genetic variability in selected populations [22,38]. The results of this study indicate the presence of variability among the progenies in most of the cases evaluated. However, low estimates of genetic variance may be explained by the narrow genetic base of soybean, as pointed out by several authors [39]. This may imply low variability within the "Subtropical" and "Tropical" populations and the existence of relatedness among parents, especially in the case of populations derived from biparental crosses.

The non-significant genetic variances observed for MATURITY in the "Tropical" population can be attributed to the intricate interactions between the long-juvenile trait in the late parents and the E allelic series. These interactions result in reduced variability when evaluated under our specific conditions [40]. It is worth noting that in individual analyses, genetic variance, heritability, and accuracy parameters may be either underestimated or overestimated if genotype × environment interactions are not considered, especially for quantitative traits. Similar genotype–environment interactions were also highlighted by [1]. In the case of the augmented block design, the absence of genotype repetitions within and between years could potentially account for the variation in parameter estimates and their relatively lower magnitude [41].

The heritability estimates indicate the possibility of selecting for MATURITY with the potential for more significant genetic gains than GY in the "WRMG" population while maintaining the same proportion of selected individuals. The "Subtropical" population allows for the selection of both MATURITY and GY with similar genetic gain possibilities, but this is lower than the "WRMG" population for MATURITY. However, for the "Northern" population, the genetic gains for MATURITY will be very low, and for GY, they could be intermediate to high.

The "WRMG" population with genetic variability for maturity and GY stood out with an efficient selection for reducing maturity and increasing grain yield, demonstrating the potential for generating lines, combining earliness with high productivity.

The "Subtropical" population, originating from crosses of cultivars adapted to higher latitudes, presented genetic variability for maturity and GY, standing out with a low average for maturity and productivity in the F5 generation. However, after an efficient selection for increased maturity combined with increased productivity, it demonstrated the potential to generate earlier maturing lineages than the "WRMG" population but with similar productivity.

The "Tropical" population originated from crosses of cultivars adapted to lower latitudes and presented a longer average cycle length and high productivity. However, after intense selection for cycle reduction, super-early lines were generated within 104 days and a high grain productivity of approximately 35 g per plant. Nevertheless, the intensity of selection prioritized for MATURITY to adapt the lines to the region caused a drastic reduction in GY genetic variability, which was insignificant in F7.

The three populations produced highly productive soybean lines with variable and appropriate MATURITY for the region and strategic crop management for early and

late sowing. Thus, the populations showed both genetic variability and high means of lines suitable for MATURITY and GY, proving that populations with broad and restricted latitude adaptations are ideal for extracting new lines in intermediate geographic regions of latitudes.

For GY, the gains obtained by the progenies ordered according to the neural network prediction were higher than those obtained using BLUP ordering in 83.4% of the northern population, 59.3% of the southern population, and 47.7% the Brazilian population.

For MATURITY, gains were higher when considering the ordering performed by the neural network in proportions of 53.3% for the "WRMG" population, 36.2% for the "Tropical" population, and 25.1% for the "Subtropical" population regarding the mixed models. Although the percentage gain was higher for the population with wider crossings for the trait, it can be considered that in this situation, there was a more significant reduction in variability due to the selection directed to the evaluation site of ideal GMR between six and eight. Although there may be an overestimation by the neural network in predicting genetic values, the estimates obtained for both MATURITY and GY agree with those obtained by [42] for soybean crops by applying the same selection intensity.

Both prediction methodologies have demonstrated the viability of successful selection based on gain estimates. The selection of genotypes for Brazil, specifically between the northern and southern regions, was grounded in MATURITY, GY, or a combination of both traits. This choice was made due to the high and relatively similar genetic means and variances observed. Opting for selection within a single environment can effectively pre-screen soybean genotypes, enhancing their potential adaptability across a broader experimental network. Moreover, it offers the advantage of significantly reducing costs within breeding programs.

5. Conclusions

Artificial neural networks demonstrated their efficiency in predicting genetic values when applied to soybean populations originating from both broad and restricted crossing populations. These networks yielded genetic gain estimates with the selection of superior progenies that outperformed those obtained through the REML/BLUP methodology. As a result, local selection at intermediate latitudes is deemed suitable for advancing generations and developing lines adaptable to regions at both higher and lower latitudes.

Author Contributions: Conceptualization, L.d.O.A. and G.V.M.; Formal analysis, G.V.M., R.S.R.d.A. and S.H.U.-T.; Investigation, L.d.O.A., A.C.R.M., D.S.C. and H.K.d.S.; Methodology, G.V.M. and S.H.U.-T.; Project administration, S.H.U.-T.; Resources, G.V.M., J.d.S.S. and S.H.U.-T.; Supervision, S.H.U.-T.; Validation, G.V.M. and S.H.U.-T.; Visualization, J.d.S.S.; Writing—original draft, L.d.O.A., G.V.M., J.d.S.S. and S.H.U.-T.; Writing—review and editing, L.d.O.A., G.V.M., L.F.A. and S.H.U.-T. All authors have read and agreed to the published version of the manuscript.

Funding: Project 404306/2021-7 National Council for Scientific and Technological Development (CNPq, Brazil).

Data Availability Statement: https://www.frontiersin.org/articles/10.3389/fpls.2022.814046/full#supplementary-material, accessed on 15 May 2023.

Acknowledgments: We thank the São Paulo State University—UNESP/FCAV, Jaboticabal, São Paulo, Brazil, the Graduate Agronomy Program (Genetics and Plant Breeding) and the Coordination for the Improvement of Higher Education Personnel (CAPES).

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

References

1. Alliprandini, L.F.; Abatti, C.; Bertagnolli, P.F.; Cavassim, J.E.; Gabe, H.L.; Kurek, A.; Matsumoto, M.N.; de Oliveira, M.A.R.; Pitol, C.; Prado, L.C.; et al. Understanding Soybean Maturity Groups in Brazil: Environment, Cultivar Classi-541 fication, and Stability. *Crop Sci.* **2009**, *49*, 801–808. [CrossRef]

2. Samanfar, B.; Molnar, S.J.; Charette, M.; Schoenrock, A.; Dehne, F.; Golshani, A.; Belzile, F.; Cober, E.R. Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. *Theor. Appl. Genet.* **2017**, *130*, 377–390. [CrossRef] [PubMed]
3. Wang, Y.; Wang, X.; Paterson, A.H.; Liu, R. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Plant Commun.* **2020**, *1*, 100050.
4. Valliyodan, B.; Qiu, D.; Patil, G.; Zeng, P.; Huang, J.; Dai, L.; Chen, C.; Li, Y.; Joshi, T.; Song, L.; et al. Landscape of genomic diversity and trait discovery in soybean. *Sci. Rep.* **2016**, *6*, 23598. [CrossRef] [PubMed]
5. Zhou, X.; Wang, D.; Mao, Y.; Zhou, Y.; Zhao, L.; Zhang, C.; Liu, Y.; Chen, J. The Organ Size and Morphological Change During the Domestication Process of Soybean. *Front. Plant Sci.* **2022**, *13*, 913238. [CrossRef]
6. Lin, J.; Arief, V.; Jahufer, Z.; Osorno, J.; McClean, P.; Jarquin, D.; Hoyos-Villegas, V. Simulations of rate of genetic gain in dry bean breeding programs. *Theor. Appl. Genet.* **2023**, *136*, 1–22. [CrossRef]
7. Sandhu, K.S.; Aoun, M.; Morris, C.F.; Carter, A.H. Genomic Selection for End-Use Quality and Processing Traits in Soft White Winter Wheat Breeding Program with Machine and Deep Learning Models. *Biology* **2021**, *10*, 689. [CrossRef]
8. Silva, C.M.; Mezzomo, H.C.; Alves, R.S.; de Resende, M.D.V.; Nardino, M. Optimizing selection of wheat genotypes through simulated individual BLUP and modified simulated individual BLUP. *Agron. J.* **2023**, *115*, 1237–1247. [CrossRef]
9. Amaral, L.d.O.; Miranda, G.V.; Val, B.H.P.; Silva, A.P.; Moitinho, A.C.R.; Unêda-Trevisoli, S.H. Artificial Neural Network for Discrimination and Classification of Tropical Soybean Genotypes of Different Relative Maturity Groups. *Front. Plant Sci.* **2022**, *13*, 814046. [CrossRef]
10. Souza, P.V.D.; de Rezende, L.P.; Duarte, A.P.; Miranda, G.V. Maize Yield Prediction using Artificial Neural Networks based on a Trial Network Dataset. *Eng. Technol. Appl. Sci. Res.* **2023**, *13*, 10338–10346. [CrossRef]
11. Alves, T.S.; Pinto, M.A.; Ventura, P.; Neves, C.J.; Biron, D.G.; Junior, A.C.; Filho, P.L.D.P.; Rodrigues, P.J. Automatic detection and classification of honey bee comb cells using deep learning. *Comput. Electron. Agric.* **2020**, *170*, 105244. [CrossRef]
12. Silva, G.O.; Schimiguel, J. Machine learning approach for crop yield prediction. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 1–5.
13. Coutinho, A.E.; Neder, D.G.; Silva, M.C.D.; Arcelino, E.C.; Brito, S.G.D.; Carvalho Filho, J.L.S.D. Phenotypic and genotypic value prediction via RR-BLUP/GWS and neural networks. *Rev. Caatinga* **2018**, *31*, 532–540. [CrossRef]
14. Sant'anna, I.C. Artificial Neural Networks in the Discrimination of Backcross Populations with Different Degrees of Similarity (Redes Neurais Artificiais na Discriminação de Populações de Retrocruzamento com Diferentes graus de Similaridade). Master's Thesis, Universidade Federal de Viçosa, Viçosa, Brazil, 2014.
15. Silva, W.D.M. Artificial Neural Networks as a Tool for Growth Prognosis and Forest Genetic Improvement. (Redes Neurais Artificiais como Ferramenta para Prognose de Crescimento e Melhoramento Genético Florestal). Ph.D. Thesis, Universidade Estadual Paulista, São Paulo, Brazil, 2019.
16. Souza, J.S.; Pedrosa, L.M.; Moreira, B.R.A.; Rêgo, E.R.D.; Unêda-Trevisoli, S.H. The More Fractal the Architecture the More Intensive the Color of Flower: A Superpixel-Wise Analysis towards High-Throughput Phenotyping. *Agronomy* **2022**, *12*, 1342. [CrossRef]
17. Federer, W.T. Augmented (or hoonuiaku) designs. *Hawaii. Plant. Rec.* **1956**, *55*, 191–208.
18. EMBRAPA—Empresa Brasileira de Pesquisa Agropecuária. *Tecnologias de Produção de Soja—Região Central do Brasil 2014*; Embrapa Soja: Londrina, Brazil, 2013; 265p.
19. Fehr, W.R.; Caviness, C.E. *Stages of Soybean Development*; (Special Report, 80); Iowa State University: Ames, IA, USA, 1977; 12p.
20. R Version 4.0.2. "Taking Off Again". The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64 (64-bit). 2020. Available online: https://www.r-project.org/ (accessed on 15 May 2023).
21. Scott, A.J.; Milliken, G.A. Monte Carlo estimation of variance components in unbalanced mixed linear models with applications to breeding trials. *Biometrics* **1993**, *49*, 97–110.
22. Bernardo, R. *Breeding for Quantitative Traits in Plants*, 2nd ed.; Stemma Press: Viçosa, Brasil, 2010.
23. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
24. Armstrong, J.S.; Collopy, F. Error measures for generalizing about forecasting methods: Empirical comparisons. *Int. J. Forecast.* **1992**, *8*, 69–80. [CrossRef]
25. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; pp. 180–184. ISBN 978-0262035613.
26. Dreyfus, S.E. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *J. Guid. Control Dyn.* **1990**, *13*, 926–928. [CrossRef]
27. Bottou, L.; Bousquet, O. The tradeoffs of large scale learning. In *Optimization for Machine Learning*; Sra, S., Nowozin, S., Wright, S.J., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 351–368. ISBN 978-0-262-01646-9.
28. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
29. Rosado, R.D.S.; Cruz, C.D.; Barili, L.D.; Carneiro, J.E.d.S.; Carneiro, P.C.S.; Carneiro, V.Q.; da Silva, J.T.; Nascimento, M. Artificial Neural Networks in the Prediction of Genetic Merit to Flowering Traits in Bean Cultivars. *Agriculture* **2020**, *10*, 638. [CrossRef]
30. González-Camacho, J.M.; de Los Campos, G.; Pérez, P.; Gianola, D.; Cairns, J.E.; Mahuku, G.; Babu, R.; Crossa, J. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* **2012**, *125*, 759–771. [CrossRef]

31. Alves, G.R.; Teixeira, I.R.; Melo, F.R.; Souza, R.T.G.; Silva, A.G. Estimating soybean yields with artificial neural networks. *Acta Sci. Agron.* **2018**, *40*, e35250. [CrossRef]
32. Najafabadi, M.Y.; Hesami, M.; Eskandari, M. Machine Learning-Assisted Approaches in Modernized Plant Breeding Programs. *Genes* **2023**, *14*, 777. [CrossRef]
33. Kaul, M.; Garg, M.K.; Choudhary, A.K. Artificial neural networks for crop yield prediction. *Agric. Syst.* **2005**, *85*, 1–18.
34. Nayak, N.J.; Maurya, P.K.; Maji, A.; Mandal, A.R.; Chattopadhyay, A. Combining Ability and Genetic Control of Pod Yield and Component Traits in Dolichos Bean. *Int. J. Veg. Sci.* **2018**, *24*, 390–403. [CrossRef]
35. Tamouridou, A.A.; Alexandridis, T.K.; Pantazi, X.E.; Lagopodi, A.L.; Kashefi, J.; Kasampalis, D.; Kontouris, G.; Moshou, D. Application of Multilayer Perceptron with Automatic Relevance Determination on Weed Mapping Using UAV Multispectral Imagery. *Sensors* **2017**, *17*, 2307. [CrossRef]
36. Etminan, A.; Pour-Aboughadareh, A.; Mohammadi, R.; Shooshtari, L.; Yousefiazarkhanian, M.; Moradkhani, H. Determining the best drought tolerance indices using Artificial Neural Network (ANN): Insight into application of intelligent agriculture in agronomy and plant breeding. *Cereal Res. Commun.* **2019**, *47*, 170–181. [CrossRef]
37. Yoosefzadeh-Najafabadi, M.; Mobasheri, M.R.; Mahdian, M.; Moghimi, E. Predicting soybean yield using different machine learning algorithms based on spectral reflectance data. *J. Appl. Remote Sens.* **2021**, *15*, 014512.
38. Ramalho, M.A.P.; Abreu, A.D.F.; Santos, J.D.; Nunes, J.A.R. *Applications of Quantitative Genetics in Breeding of Self-Pollinated Plants*; UFLA: Lavras, Brazil, 2012.
39. Hiromoto, D.M.; Vello, N.A. Genetic diversity in soybean cultivars released in Brazil before 1980. *Crop Sci.* **1986**, *26*, 1149–1153.
40. Yang, W.; Wu, T.; Zhang, X.; Song, W.; Xu, C.; Sun, S.; Hou, W.; Jiang, B.; Han, T.; Wu, C. Critical Photoperiod Measurement of Soybean Genotypes in Different Maturity Groups. *Crop. Sci.* **2019**, *59*, 2055–2061. [CrossRef]
41. Rocha, M.D.M.; Vello, N.A. Genotype-location interaction for seed yield in soybean lines with different maturity cycles. *Bragantia* **1999**, *58*, 69–81. [CrossRef]
42. Amaral, L.O.; Bruzi, A.T.; de Resende, P.M.; Silva, K.B. Pure line selection in a heterogeneous soybean cultivar. *Crop. Breed. Appl. Biotechnol.* **2019**, *19*, 277–284. [CrossRef]