Maize Yield Prediction using Artificial Neural Networks based on a Trial Network Dataset

Paulo Vitor Duarte de Souza
Campus Santa Helena, Universidade Tecnologica Federal do Parana, Brazil
victorduart366@gmail.com

Leiliane Pereira de Rezende
Campus Santa Helena, Universidade Tecnologica Federal do Parana, Brazil
rezende@utfpr.edu.br

Aildson Pereira Duarte
Centro de Graos e Fibras, Instituto Agronomico de Campinas, Brazil
duarteaildson@hotmail.com

Glauco Vieira Miranda
Campus Santa Helena, Universidade Tecnologica Federal do Parana, Brazil
glaucovmiranda@utfpr.edu.br
(corresponding author)

Received: 6 January 2023 | Revised: 17 January 2023 | Accepted: 20 January 2023

ABSTRACT

The prediction of grain yield is important for sowing, cultivar positioning, crop management, and public policy. This study aims to predict maize productivity by applying an artificial neural network and by building models of multilayer perceptrons (MLPs) using public data and maize experimental networks. The dataset included parameters of climate, soil water balance, and agronomic characteristics from maize hybrids of an experimental network of two agricultural years. The climatic and soil balance water parameters were divided according to the maize plant development stages. Six databases were obtained by combining the imputation of missing data with the agronomic characteristics of the maize hybrids, the climatic parameters/soil water balance, and the complete database with both. Hyper parameterization of the models was obtained using GridSearch and k-fold cross-validation. The models with imputation were more accurate than those without it. The model with climate data/soil water balance and the complete model with imputation presented the smallest errors of 71 kg ha^{-1}. In all the models, cultivars, locations, and their interactions were important, and different climatic conditions had the greatest weight in predicting productivity. It was concluded that the MLP models performed adequately and captured the non-linear effects of the interaction between the environment and maize hybrids. Climatic and soil balance water parameters at different stages of maize plant development explain the productivity of maize hybrids more than the agronomic characteristics of the cultivars.

Keywords-artificial neural networks; deep learning; multilayer perceptron; agricultular productivity

I. INTRODUCTION

Interpretation of the genotype x environment interaction is essential to increase productivity with the positioning of the cultivar in the microregion and its predictability of performance according to the sowing date [1]. Traditionally, analysis methods of the genotype x environment interaction have been used for the positioning of cultivars, however, the interpretations obtained after obtaining data and generalizations about the environment do not allow the application or extrapolation of results for all regions [2-3]. However, when the environments are well characterized, conclusions on cultivar performance are extrapolatable, as used nitrogen levels in rice [4] or used water deficit environments in maize [5]. The simulation models of plant growth in response to climate, soil, and cultivars allow identifying the best sowing times, which are made available annually for different cultures and locations, such as the Agricultural Zoning of Climatic Risk, and have helped Brazilian farmers in all grain-producing locations [6-7]. However, even with agricultural risk zoning, weather conditions are unpredictable, and productivity losses are common in different regions, locations, and years when this information is not available in real time. Production losses caused by drought in the Paraná-La Plata basin, Brazil, have reduced soybean and maize production and affected global and South American agricultural markets during the past ten years, with the 2020-2021 drought imposing a 2.6% reduction in the cereal harvest compared to the previous year [8].

The multilayer perceptron (MLP) is an Artificial Neural Network (ANN) that can be used for productivity estimates. It uses the factors that affect productivity as an implicit function of the network input. MLPs have been successful in creating predictive models using plant development parameters, soil characteristics, management, climatic conditions, and geographic positions with a small mean error for yield prediction [9-11]. The predicting maize productivity using ANN can generate information for public policy on the production and supply of this important crop for local and national food security. The prediction can also generate real-time information for managing corn crops to increase productivity and sustainability.

ANNs are being applied with different models or architectures to solve common problems in plant breeding and genotype x environment interactions and other agriculture situations. This can be observed when determining drought tolerance indices for durum wheat and identifying their efficiency in relation to other methods [12], parents for crossings in breeding programs [13] and, selected soybean plants in segregating populations of different maturity groups [14]. A machine learning approach was used in the automatic irrigation system based on humidity and temperature in the soil in [15]. The deep convolutional neural network architecture is very efficient to identify plants in the different stages including seedling of weed plants [16]. In addition, the estimate of maize productivity as a function of leaf chlorophyll content, measured in three stages of plant development, was determined by an ARN-MLP and was adequate, considering that in the stage of development of the plant with six leaves (V6), it explained 50% or more of maize productivity data [17]. In turn, arabica coffee production was predicted to meet the market demand using ANNs with data on productivity, rainfall, relative humidity, and minimum and maximum temperatures. An adequate prediction accuracy was obtained with an R^2 of 0.8524 and RMSE (root-mean-square deviation) of 0.0784 tons, demonstrating the potential of ANNs in determining the yield of the coffee cherry [18].

Therefore, the objective of this study was to predict maize productivity by applying ANNs and building MLP models while using public data and maize experimental networks.

II. MATERIAL AND METHODS

The variables for the analysis in this study were divided into agronomic, climatic, and soil water balance parameters.

The agronomic characteristics of maize hybrids were obtained from the Performance of Maize Hybrids at the Second Season, with 38 hybrids in 2018 and 32 hybrids in 2019 in 9 locations and 4 replications/locations in the Bernardino de Campos, Capão Bonito, Cândido Mota, Cruzália, Ibirarema, Manduri, Maracaí, Palmital, and Pedrinhas Paulista in the middle region of the Paranapanema Valley, state of São Paulo [19-21]. The agronomic characteristics evaluated were plant height, ear height, the ratio between ear height and plant height, the number of lodged plants, number of broken plants, number of days to flowering, plant population, productivity, grain moisture at harvest, and ear index per plant. The daily climate parameters were total precipitation, maximum, average, and minimum temperatures, average atmospheric pressure, average dew point temperature, average and minimum relative humidity, average wind speed, and maximum wind gust. The daily soil water balance variables included water storage, actual evapotranspiration, soil water deficit, soil water surplus, and evapotranspiration. Total precipitation, water storage, and water surplus in the soil were considered at the cumulative sum of each plant development stage. The average for each period was considered for the other climate and water balance variables. Daily climatic and soil water balance data were obtained from automatic stations (Ourinhos A716, Itapeva A714, Avaré 725) at the National Institute of Meteorology (INMET), which provides a variety of daily meteorological data on a platform [22]. As the INMET does not have meteorological stations in some municipalities, it was necessary to consider data from the nearest stations, covering distances from 27 to 95km. Water balance variables were subdivided according to the growth and developmental stages of the maize plant.

The developmental stages of the maize were plant emergence (VE), 4 leaves with a visible leaf collar (V4), 8 leaves with a visible leaf collar (V8), tasseling (VT), silk and blister (R1 and R2), milky and dough grains (R3 and R4), dent grains (R5), and physiological maturity (R6). The cultivars had the same 120-day cycle, and the stages were considered with the same periods for all cultivars. The days for each stage were VE from the emergence to the 10th day, V4 from the 11th to the 40th day, V8 from the 41st to the 50th day, VT from the 51st to the 60th day, R1 and R2 from the 61st to the 75th day, R3 and R4 from the 76th to the 90th day, R5 from the 91st to the 105th day, and R6 from the 106th to the 120th day.

The database was processed before the effective implementation of the models and comprised up to 2392 data from 17 experiments on the performance of maize hybrids at the second season in 2018 and 2019. The necessary procedures for data preparation were the imputation of missing data, coding of categorical variables, and normalization (Figure 1). The database contained missing data on flowering and grain moisture, representing 6% of the total database. The iterative imputation algorithm was used to impute these data, and the necessary parameterization for the execution of the iterative imputation was the neural network's own MLP with a δ_{min} of 0.1. The parameter δ_{min} defines the number of iterations performed by the iterative imputation and the minimum acceptable error in relation to previous iterations.

The regressor defined for the predictions in the MLP with the hyperparameters for data imputation had three hidden layers, 64 neurons per layer, the ReLU (Rectified Linear Unit) activation function in the hidden layer, 300 epochs, ADAM optimizer, 0.001 rate learning curve, MSE (Mean Squared Error) cost function, and the Glorot normal-weight initialization method. This regressor was chosen based on the ability of ANNs to capture nonlinear relationships between data [23]. The dataset was divided into two groups: one with all variables with complete data and the other with only variables with missing values. Therefore, two MLP models were built to estimate the variables of both sets using the same hyperparameters. For model training, each dataset was partitioned into training and testing sets. The training set had 80% of the examples, and the test set had the 20%. After inputting the data, transformations were applied to the categorical variables, cultivars, and years. The codification adopted transformed them as a function of annual productivity, given the variable of interest:

$$Location_r = \frac{1}{l}\sum_{i=1} p_i - p_{year} \tag{1}$$

where l is the number of observations carried out in the r^{th} *Location*, p_i is the i^{th} productivity of the *Location*, and p_{year} is the average of the observations made.

$$Year_s = \frac{1}{a}\sum_{j=1}^{a} p_j - p_{year} \tag{2}$$

where a is the number of observations carried out in the s^{th} *Year*, and p_j is the j^{th} productivity of the *Year*.

$$C_k = p_k - p_{Location} \tag{3}$$

$$cultivar_e = \frac{1}{e}\sum_{k=1}^{e} c_k \tag{4}$$

where p_k is the k^{th} observation of the productivity of cultivar c_k, $p_{Location}$ is the average productivity of a given location and, e is the number of observations of the cultivar, and c_k is the k^{th} difference between p_k and $p_{Location}$.

In all the above equations, the categorical variables are functions of the observed average productivity, allowing the attribution of an evaluation to these variables. In the normalization procedure, the network input dataset was normalized by MinMax scaling. The dataset was separated into 3 distinct bases: agronomic characteristics, climatic/water balance parameters, and complete with all variables with or without missing data imputations. In addition, 6 network models were implemented for the same purpose, with the task of predicting productivity as the model output. The necessary hyper parameterization in the 6 models was defined using Grid Search and k-fold cross-validation to find a set of hyperparameters with the smallest error in the cost function. For this purpose, a set of hyperparameters was previously considered to be tested by Grid Search and evaluated by k-fold cross-validation. $k = 5$ was considered for k-fold cross-validation, which is one of the values commonly used in the literature [24].

The input hyperparameters for GridSearch were 3 hidden layers with 32, 64, and 128 neurons per layer, the ReLu activation function in the hidden layer, 300, 600, and 900 epochs, the ADAM optimizer, learning rates of 0.005, 0.003, and 0.001, the MSE cost function, and the Glorot normal weight initialization method.

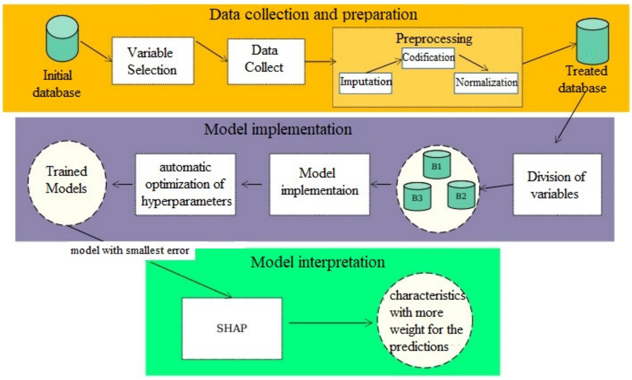

Fig. 1. The proposed methodology.

The MLP experiments were performed on the services provided by Google Research, Google Collaboratory, or Colab, which is a host of services based on the Jupyter notebook, where it is possible to interactively execute codes in the Python language, in addition to having a set of tools for the development of deep learning models [25-26]. Six models were constructed using TensorFlow and Keras [27-28]. The k-fold cross-validation used in model validation matched that of the machine learning library scikit-learn, which has a set of tools for developing machine learning models [23]. The SHapley additive exPlanations (SHAP) method was applied so that the model with the best performance could be interpreted agronomically [29].

III. RESULTS

The hyperparameters of the models without imputation, with the lowest RMSE among the 142 models trained using Grid Search and k-fold cross-validation, are listed in Table I. The AGR_M model required a greater number of layers and a larger batch size than the CLI_M and COP_M models, demonstrating a greater requirement of hyperparameters for model adjustment. A hidden search is computationally expensive, which makes it impossible to apply the method to a large number of hyperparameters [30]. Therefore, there is a possibility that the hyperparameters are not the global optimum but local optima due to the small set of hyperparameters. However, the use of the grid search is justified by its computational cost [30].

The learning curves of the AGR_M, CLI_M, and COP_M models with MSE show concise training without much variation between training and validation and without overfitting to training data, with training time up to 1000 epochs (Figure 2). In the CLI_M model, closer values were obtained between the training and validation sets, as can be observed in the distance between the training and validation curves. In addition, there was variation at the beginning of the adjustment of the models, which is evidenced by the epochs where the adjustments of the functions began, remaining below 100 epochs for AGR_M and COP_M, and close to 250 for CLI_M. The possible causes for this are the hyperparameters used in the models, such as the batch size, number of layers, and database. The COP_M model presented an intermediate distance between the training and validation cost functions.

This occurred when the data of agronomic characteristics were included, bringing data variation to the adjustment of the functions. However, they lose the advantage of the greater adjustment of the climatic data of the CLI_M model.

In the model validation dataset, the AGR_M obtained the highest RMSE (201 kg ha^{-1}) and the CLI_M model had the lowest RMSE (191 kg ha $^{-1}$), demonstrating the best fit in relation to the first (Table II). Model-associated errors (MSE) were relatively similar, with a tendency to be greater in the AGR_M model than in the others, demonstrating the lowest fit and highest hyper parameterization of the ANN. Based on Pearson's correlation statistics, all models obtained similar values close to 1. In the training dataset, the AGR_M model presented the lowest RMSE and MSE values, characterizing its better performance in the adjustment in relation to the other two models that were very similar, with the smallest variation of the dataset. In turn, this situation is reversed, in validation, when the AGR_M model presented the highest RMSE and MSE values.

TABLE I. HYPERPARAMETERS OF MODELS WITH LOWER ERRORS BY GRID SEARCH AND K-FOLD CROSS-VALIDATION WITH (I) AND WITHOUT (M) IMPUTATION FOR PREDICTING GRAIN YIELD (KG.HA–1)

Hyper parameters	**AGR _M**	**CLI _M**	**COP _M**	**AGRI _I**	**CLI _I**	**COP _I**
Hidden layers	2	1	1	2	1	1
Neurons	64	64	64	128	128	64
Epochs	1000	1000	1000	1000	600	1000
Batch size	128	32	64	128	32	64
Learning rate	0.001	0.001	0.001	0.005	0.001	0.001

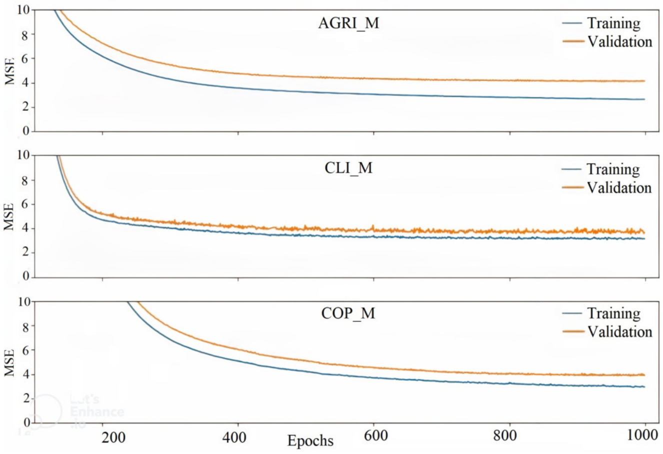

Fig. 2. Cost function (MSE) training and k-fold cross-validation without data imputation for AGR_I, CLI_I and COP_I models.

When imputing the missing data in the base, another set of hyperparameters was needed to reach the local optimum when submitted to grid search (Table II). This indicates that the missing data complement new standards for the adjustment of the models and that each one of them was changed differently with a greater number of hidden layers, neurons/layer, epochs, and batch size for AGR_I, a smaller number of layers, batch size, epochs, and larger neurons/layer for CLI_I, and fewer hidden layers and neurons/layer for COP_I among the 142 models trained by grid search and k-fold validation (Table I). The changes between the non-imputed and imputed models were in the number of neurons, epochs, batch size, and learning rate, requiring a new ANN architecture. Considering that each neuron corresponds to a hyperplane, a greater number of 128 neurons for the AGR_I and COP_I models is necessary to properly separate the instance space. Only COP_I remained with the same number of neurons as the models without data imputation, characterizing the explanation of model adjustment using climate variables.

To predict the behavior of parents for breeding programs, the architecture of the ANN was optimized using 3 layers: the first with 64 neurons, the second with 32 neurons, and the third with 16 neurons, with the ReLU activating function having the highest accuracy of prediction [13]. The learning rate of AGR_I was 0.005, demonstrating that the model was "stuck" in poor locations with a lower learning rate. Only CLI_I managed to converge in a smaller number of epochs, and 600 epochs reinforced the quality of the climate data for predicting productivity with the imputed data. The impact of data imputation on model training showed that the MSE in 1000 epochs was lower than that of the models without data imputation for AGR_M and COP_M (Table I). However, the training was less consistent, as can be seen by the requirement for a greater number of epochs to start the adjustment of the imputed models AGR_I, CLI_I, and COP_I compared to the non-imputed models AGR_M, CLI_M, and COP_M (Figures 2 and 3). One possible reason for this is that imputation increases data variability.

TABLE II. METRICS OF MODELS WITH AND WITHOUT IMPUTATION FOR PREDICTING GRAIN YIELD (KG.HA^{-1})

Metrics	**AGR _M**	**CLI _M**	**COP _M**	**AGR _I**	**CLI _I**	**COP _I**
RMSE Validation	201	191	197	80	71	74
MSE Validation	414	365	390	87	60	60
Correlation Validation	0.99	0.99	0.99	1	1	1
R^2 (%) Validation	0.9875	0.9819	0.9842	0.9978	0.9975	0.9983
RMSE Training	162	177	172	49	75	69
MSE Training	264	315	295	25	56	48
Correlation Training	1	0.99	0.99	1	1	1
R^2 (%) Training	0.9912	0.9890	0.9907	0.9994	0.9986	0.9987

The curves of the models with imputation show that CLI_I training and validation are closer than those of AGR_I and COP_I (Figure 3). In addition, the number of epochs needed to converge was smaller, so, it can be inferred from this model, with or without imputation of data, that climate explains the behavior of productivity. The imputed model that obtained the lowest RMSE was CLI_I with 71 kg.ha^{-1}, demonstrating an adequate fit with an error of just over one bag of maize per hectare (Table II). It is also possible to notice that the training and validation errors are similar, demonstrating that the model can generalize productivity prediction. However, AGR_I obtained the highest validation RMSE (80 kg ha $^{-1}$), which allows adjusting the model to the training examples. The COP_I model obtained a validation RMSE of 74 kg ha $^{-1}$. However, it tended to increase the error (MSE) during training. In all models, there was a high correlation, with no significant differences with the models without data imputation. The AGR_I model, when compared to the AGR_M model, showed more overfitting to the training data (Figures 1 and 2). Like AGR_M, the AGR I model presents the same situation as the agronomic characteristics, which are not sufficient to explain productivity. This suggests that more agronomic traits are needed to increase grain yield prediction and that climate variables are fundamental for hybrid yield estimation. This can be confirmed by the COP_I and COP_M models, which consider all variables in the database; therefore, the training and validation curves are closer to each other.

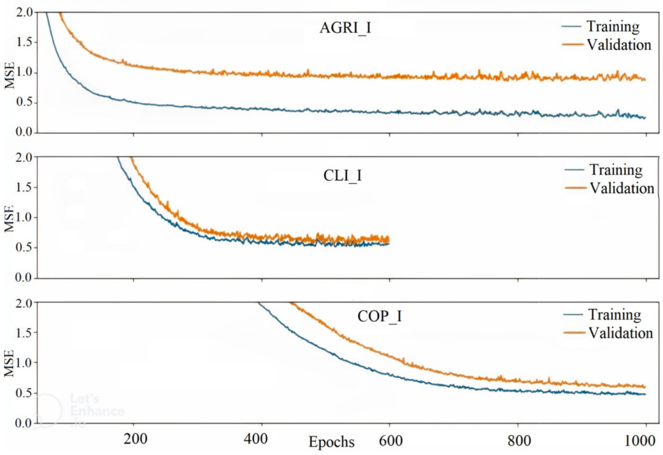

Fig. 3. Cost function (MSE) training and k-fold cross-validation validation with data imputation for AGR_I, CLI_I and COP_I models.

Based on the metrics of the models, the imputation of data obtained superior performance compared to the base with missing data. The models with imputation presented RMSE up to 3 times lower than those without, which suggests that the problem in question requires a reasonable amount of data so that the performance of the ANN is increasing, but due to database reduction and imputation, the MLP models are adequate. The models with and without imputation obtained an RMSE very close to 1% and 3% of grain yield which is considered very good. In the Syngenta Crop Challenge 2018, with a productivity data set of 2,267 maize hybrids in 2,247 locations between 2008 and 2016, the of [10] was considered one of the best, with an RMSE of 12% of the average yield using forecast weather data.

Using another strategy, maize yield prediction was performed with machine learning with different models, and the best of them, Stacked LASSO, presented a Mean Bias Error (MBE) of 53 kg/ha and a Relative RMSE (RRMSE) of 9.5% [32]. For the soybean crop, annual productivity was predicted in the region of the Brazilian states of Maranhão, Tocantins, Piauí, and Bahia as a function of monthly climate variables (air temperature, precipitation, and global radiation) and water balance components (cultivation evapotranspiration, storage, actual cultivation evapotranspiration, water deficiency, and surplus) during plant development through a deep artificial neural network [33] with a dataset size of 920 examples, the obtained RMSE was 167.85 kg.ha^{-1}. The unprecedented differential in the strategy of using climatic variables subdivided into maize plant development stages that are differentially sensitive to water stress was assertive. This can be the main reason why, even with a smaller database, the RMSE of the models was smaller than that found in the literature.

Despite the generalization of the positioning of cultivars by different methodologies, either the traditional methods of adaptability and stability or ANN models, the germplasm of the species must have a genetic response to the common stress in the region, as verified by the authors in [34, 35], who identified maize cultivars with different responses to nitrogen. Authors in [36] identified the response of maize cultivars as a function of organic fertilization, and in [37] identified different responses of maize cultivars to irrigation with saline water. Therefore, without genetic control of tolerance or resistance to stress, the differences between cultivars do not affect productivity, and it is simply dependent on the environment, being high when this is favorable. However, when tolerance or resistance to stress has genetic control, the environment with or without the presence of stress does not affect the productivity of the cultivar, as it is due to its genes. Due to the importance of the cultivar in the models, it can be stated that there are differences between the cultivars how they interact with the environment.

MLP models with different precisions do not allow us to identify which agronomic or climatic variables are the most important for predicting productivity. This can be achieved with SHAP analysis applied to the COP_I model, identifying the variables with the greatest impact on the model (Figure 4). The most important variables were the cultivar, the average productivity of the cultivars/site, and the site itself (edaphic conditions), which is consistent with the agronomic explanation of the productivity of the cultivar, the physical and climatic environment, and the interaction of the cultivar with the environment [38]. The cultivar defines the productive potential of the environment and its limitations in certain environments. The adaptation of cultivars to the environment is related to their genetic adaptation to the soil, water regime, plant development cycle, absorption and use of nutrients, and tolerance to insects, pests, and diseases [39]. Cultivar, productivity/location (prod_local), and location had the greatest impact on the output of the variable in the COP_I model. The prod_local defines the average productivity that can be obtained in a given location, showing the effect of the climate× soil×cultivar interaction. The local variable defines the edaphic conditions, latitude, and altitude of the environment. This means that the model was able to capture edaphic effects on productivity and the interaction between the cultivar and the environment. The environment, represented by the effect of climate and local conditions, is an important factor that can vary over time, and edaphic conditions are fixed between harvests [38]. EPH (ear/plant height) ratio is an important characteristic of grain yield and is directly related to plant lodging, which directly interferes with harvesting when performed mechanically. However, it should be considered that the agronomic characteristics present genetic correlations, which are determinants in the expression of the phenotype when evaluated under a certain stress, which is also a cause of productivity [40]. In the results found, the evaluated agronomic characteristics and the cultivars were presented in a similar way, not being related to the climatic variables that explain productivity. relative humidity (RH), excess precipitation in the soil (EP), EPH ratio, total daily precipitation (TP), minimum daily temperature (MT), soil water storage (WS), deficit precipitation in the soil (DP), wind with maximum daily gust (WG), evapotranspiration (EV), emergence (VE), forth leaf (V4), eighth leaf (V8), tassel (VT), silk and blister (R1_R2), physiological maturity (R6).

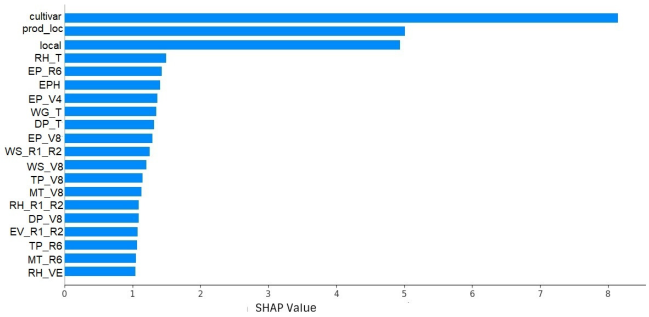

Fig. 4. Average impact of the variables on the output of the COP_I model.

In the emergency stage, only the relative air humidity (HR_VE) is important. The emergence of seedlings and their establishment provide the formation of a stand, which is the main component that explains productivity and is affected by HR_VE. Relative air humidity plays an important role in plant transpiration and soil evaporation, making plants more sensitive to a lack of precipitation [41].

In the vegetative phase, during the fourth-leaf stage, only excess precipitation (PS_V4) was highlighted, and in the eighth-leaf stage, the total daily precipitation (PT_V8), minimum daily temperature (MT_V8), water storage (WS_V8), and precipitation deficit (PD_V8) were highlighted. MT_V8 and PS_V4 act together in plant growth and development, where the number of plants per hectare and plant height is defined. Relative air humidity plays an important role in plant transpiration and soil evaporation, rendering the lack of precipitation more sensitive [41]. The minimum daily temperature is an important factor for the metabolism of corn and the entire C4 plant, inhibiting the rubisco enzyme that affects carbohydrate production, which is closely related to productivity [42]. The variables PS_V4, PT_V8, WS-V8, and PD_V8 related to precipitation are fundamental for the development of the plant and the prediction of productivity, requiring 3–4 mm in the initial stages of the corn plant and reaching the need close to 6-8 mm being the main parameters associated with the limitation of grain yield [43].

In tasseling, relative air humidity (AH_T), wind with maximum daily gust (WG_T), and precipitation deficit (PD_T) were important. The variable AH_T during this phase can affect grain production by reducing plant fertility in dry environments typical of the second crop sowing in February in São Paulo State, Brazil [44-45]. Wind gusts can damage maize plant leaves, break plant culms, or bed down plants, causing a direct loss of harvest [46]. These results are consistent in predicting the performance of maize productivity and temperature in the flowering phase, and the frequency and the amount of water received during the vegetative phase, and grain filling were identified as the main environmental factors [47]. The lack of precipitation is considered the most important environmental parameter in bolting, delaying the coincidence of receptivity between male and female flowers and preventing fertilization [44, 48].

In the reproductive phase, in the silk and blister stage, relative humidity (RH_R1_R22), water storage (WS_R1_R22), and evapotranspiration (EV_R1_R22) were important. At the physiological maturity stage, total daily precipitation (TP_R6), minimum daily temperature (MT_R6), and excess precipitation (EP_R6) were important for predicting the model in the final stage of maturation and were related to plant physiology at this stage. This is justified by the intense water, thermal, and relative humidity requirements of the air at the beginning of filling and grain formation of the corn plant. Note that the 30 days referring to the reproductive stages R3 to R5 did not influence productivity prediction.

Different climatic and water balance variables had different impacts on the stages of maize plant development, indicating that the partitioning of these variables must be carried out to increase the accuracy of the models.

IV. DISCUSSION

The different ANN architectures of the models demonstrate a greater need for data or hyper parameterization for the AGR_M model than for the other models due to the greater number of hidden layers and batch size. The CLI_M model had the lowest data requirement for productivity prediction, characterizing the importance of environmental data in relation to hybrids when they were evaluated in groups with the same cycle. However, all three models without imputation showed good error estimates with values around 200 kg.ha^{-1}, which is considered small in yields that reach at least 6000 kg.ha^{-1}, representing only 3%.

The AGR_M, CLI_M, and COP_M models with 3 hidden layers did not obtain optimal hyperparameters defined by the grid search and k-fold cross-validation in the databases without data imputation. Only the models with 1 or 2 hidden layers obtained optimal hyperparameters, and therefore, the models with 3 layers were discarded. This may have occurred because the model tended to overfit the training data.

The AGR_M model required a larger batch size than the other two models for more consistent training, because the agronomic characteristics of the hybrids, such as the same cycle, height of plants and spikes, and grain productivity, are similar or very close, because Brazilian legislation requires that the new cultivars must be at least equal to those on the market in terms of productivity, even though the germplasms that originated from different hybrids have some similarity between the companies. Therefore, the agronomic characteristics of the hybrids do not concisely explain the different yields between the environments since none of the hybrids capitalized on the genotype×environment interaction between the evaluated microregions. This similarity between the hybrids required different batch sizes of the CLI_M and COP_M models for adjustment.

The MLP-ANN models with imputation (I) showed lower errors than the models without imputation (M), lower ANN architecture requirements, and even fewer epochs. The set of routines for maize hybrid evaluation is more suitable for use in MLP-ANN models when data imputation is applied. Different architectures of the MLP-ANN models with smaller batch sizes, layers, or epochs occur for the CLI_M and CLI-I models. They also present the smallest errors and converge in models more assertively even with climate data from distant meteorological stations, from 27 to 100 km. This characterizes the robustness of the MLP-RNA models and the importance of climate data in predicting productivity in relation to the agronomic data of hybrids that present significant differences in their productivity [17, 18].

The AGR_M and AGR_I models that are based only on the agronomic data of corn hybrids require a greater process capacity and more complex models with larger batch sizes, layers, and times. The AGR_I model had the lowest precision (0.005) among all models. The accuracy of the models characterizes their adequacy. It was also found that, in several scientific publications in experiments evaluating the productivity of maize hybrids, the coefficients of variation were equal to or less than 12%, which can characterize minimal significant differences between the hybrids than 600 kg.ha^{-1}, considering yields from 5 ton.ha^{-1}. Significant differences between corn hybrids in experiments as to their yields were detected when their differences were greater than 12%, as verified in 143 experiments in 15 years of publications with the maize crop for the grain yield, in which the average coefficient of variation was 11.87% [49].

When dealing with both agronomic and climatic characteristics, the COP_M and COP_I models bring to their iterations the difficulties encountered with the AGR models. However, with data imputation, the AGR_I model presents an MSE close to the COP_I's but requires a greater number of epochs. The low RMSE values can be explained by the methodology used to evaluate the climatic data in the different stages of plant development, which were strategically correct and possibly the cause of the success of the ARN used. This was confirmed by the importance of different climatic parameters and water balance as variables of importance in the construction of the model. This also allows the model to be used in advance according to the stages of development of the plant, allowing for more assertive decision-making in different scenarios and moments of the crop. Therefore, those in charge of the crop can obtain new information at the time of sowing and at the time of cultural practices and fertilization during the vegetative phase, where it is still possible to manage the crops when they are not irrigated. When crops are irrigated, it is possible to manage the application throughout the plant's development, especially at the most critical times, such as grain filling.

The models can detect the non-linear effects between climate, soils, and genotypes, especially when climate information is included in the model or when they are sufficient to predict grain yield, even with several hybrids that have the same cycle.

The results of the models are adequate, even considering the limited amount of data analyzed on the ANNs and their application in the processes of the improvement programs without the need for modifications in the processes. The ANN models can identify important variables during the development stage of the plant, and it is observed that climatic and water balance variables are important. The water balance variables are important at all the stages of maize plant development and demonstrate the availability of water that is stored in the soil and that can be used by the plant, being the main environmental parameter limiting productivity [48].

New environmental parameters can be incorporated into the ANN models, further increasing their accuracy. An ANN based on a graph-based framework developed by Graph Neural Network-Recurrent Neural Network (GNN-RNN), using geospatial and temporal information to predict crop productivity was proposed in [11].

V. CONCLUSION

- The implemented ANNs managed to extract a pattern in the data even if the locations of the experiments and cultivars were different, which shows the capacity of the ANNs in the generalization of productivity prediction.
- The ANN models are well suited for predicting corn yield and present concise training, even with a limitation in the amount of data, when exclusively using the agronomic characteristics of the hybrids. This limitation is overcome when climatic data are incorporated into the model because most of the variation in productivity is due to these parameters.
- The imputed climate model was the most homogeneous in its performance and was the most suitable for productivity prediction.
- The SHAP analysis is important for the agronomic understanding of the results and is completely consistent with the knowledge of the ecophysiology of corn plants under stress.
- Multilayer perceptron models present adequate performance and capture the non-linear effects of the interaction between the environment and maize cultivars.
- The climatic parameters of each stage of maize plant development explain the productivity of the cultivars more than their agronomic characteristics.

REFERENCES

[1] A. Singamsetti *et al.*, "Genotype × environment interaction and selection of maize (Zea mays L.) hybrids across moisture regimes," *Field Crops Research*, vol. 270, Aug. 2021, Art. no. 108224, https://doi.org/10.1016/j.fcr.2021.108224.

[2] N. Anuradha *et al.*, "Comparative Study of AMMI- and BLUP-Based Simultaneous Selection for Grain Yield and Stability of Finger Millet

[*Eleusine coracana* (L.) Gaertn.] Genotypes," *Frontiers in plant science*, vol. 12, Jan. 2021, Art. no. 786839, https://doi.org/10.3389/fpls.2021.786839.

[3] M. Balderacchi *et al.*, "Genotype by Environment Interaction on Tropical Maize Hybrids Under Normal Irrigation and Waterlogging Conditions," *Frontiers in Sustainable Food Systems*, vol. 6, Jun. 2022, Art. no. 913211, https://doi.org/10.3389/fsufs.2022.913211.

[4] M. Abdelrahman *et al.*, "Detection of Superior Rice Genotypes and Yield Stability under Different Nitrogen Levels Using AMMI Model and Stability Statistics," *Plants*, vol. 11, no. 20, Jan. 2022, Art. no. 2775, https://doi.org/10.3390/plants11202775.

[5] L. V. de Souza, G. V. Miranda, J. C. C. Galvao, L. J. M. Guimaraes, and I. C. dos Santos, "Combining ability of maize grain yield under different levels of environmental stress," *Pesquisa Agropecuária Brasileira*, vol. 44, pp. 1297–1303, Oct. 2009, https://doi.org/10.1590/S0100-204X2009001000013.

[6] N. C. Eli-Chukwu, "Applications of Artificial Intelligence in Agriculture: A Review," *Engineering, Technology & Applied Science Research*, vol. 9, no. 4, pp. 4377–4383, Aug. 2019, https://doi.org/10.48084/etasr.2756.

[7] "Zoneamento Agricola," *Ministerio da Agricultura e Pecuaria*. https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/programa-nacional-de-zoneamento-agricola-de-risco-climatico/zoneamento-agricola.

[8] *State of the Climate in Latin America and the Caribbean 2021 (WMO-No. 1295)*. Geneva, Switzerland: WMO, 2022.

[9] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1–18, Jul. 2005, https://doi.org/10.1016/j.agsy.2004.07.009.

[10] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, May 2019, Art. no. 621, https://doi.org/10.3389/fpls.2019.00621.

[11] J. Fan, J. Bai, Z. Li, A. Ortiz-Bobea, and C. P. Gomes, "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 11873–11881, Jun. 2022, https://doi.org/10.1609/aaai.v36i11.21444.

[12] A. Etminan, A. Pour-Aboughadareh, R. Mohammadi, L. Shooshtari, M. Yousefiazarkhanian, and H. Moradkhani, "Determining the best drought tolerance indices using Artificial Neural Network (ANN): Insight into application of intelligent agriculture in agronomy and plant breeding," *Cereal Research Communications*, vol. 47, no. 1, pp. 170–181, Mar. 2019, https://doi.org/10.1556/0806.46.2018.057.

[13] S. Khaki, Z. Khalilzadeh, and L. Wang, "Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach," *PLOS ONE*, vol. 15, no. 5, 2020, Art. no. e0233382, https://doi.org/10.1371/journal.pone.0233382.

[14] L. de O. Amaral, G. V. Miranda, B. H. P. Val, A. P. Silva, A. C. R. Moitinho, and S. H. Uneda-Trevisoli, "Artificial Neural Network for Discrimination and Classification of Tropical Soybean Genotypes of Different Relative Maturity Groups," *Frontiers in Plant Science*, vol. 13, Jul. 2022, Art. no. 814046, https://doi.org/10.3389/fpls.2022.814046.

[15] A. H. Blasi, M. A. Abbadi, and R. Al-Huweimel, "Machine Learning Approach for an Automatic Irrigation System in Southern Jordan Valley," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6609–6613, Feb. 2021, https://doi.org/10.48084/etasr.3944.

[16] N. C. Kundur and P. B. Mallikarjuna, "Deep Convolutional Neural Network Architecture for Plant Seedling Classification," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9464–9470, Dec. 2022, https://doi.org/10.48084/etasr.5282.

[17] G. K. Michelon, P. L. de Menezes, C. L. Bazzi, E. P. Jasse, P. S. G. Magalhaes, and L. F. Borges, "Artificial neural networks to estimate the productivity of soybeans and corn by chlorophyll readings," *Journal of Plant Nutrition*, vol. 41, no. 10, pp. 1285–1292, Jun. 2018, https://doi.org/10.1080/01904167.2018.1447579.

[18] Y. Kittichotsatsawat, N. Tippayawong, and K. Y. Tippayawong, "Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques," *Scientific Reports*, vol. 12, no. 1, Aug. 2022, Art. no. 14488, https://doi.org/10.1038/s41598-022-18635-5.

[19] A. P. Duarte and E. Sawazaki, *Avaliação regional de cultivares de milho safrinha Resultados 2018*, 1st ed. Assis, Brazil: IAC/APTA, 2018.

[20] A. P. Duarte and E. Sawazaki, *Avaliação regional de cultivares de milho safrinha Resultados 2019*, 1st ed. Assis, Brazil: IAC/APTA, 2019.

[21] P. V. D. de Souza, "Rede neural artificial para predicao da produtividade da cultura do milho," Ph.D. dissertation, Federal University of Technology-Parana, Santa Helena, Brazil, 2021.

[22] "BDMEP." https://bdmep.inmet.gov.br/.

[23] M. B. Richman, T. B. Trafalis, and I. Adrianto, "Missing Data Imputation Through Machine Learning Algorithms," in *Artificial Intelligence Methods in the Environmental Sciences*, S. E. Haupt, A. Pasini, and C. Marzban, Eds. Dordrecht, Netherlands: Springer, 2009, pp. 153–169.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2826–2830, Oct. 2011, https://doi.org/10.5555/1953048.2078195.

[25] E. Bisong, "More supervised machine learning techniques with Scikit-learn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Ottawa, ON, Canada: Apress, 2019, pp. 287–308.

[26] "Google Colab." https://research.google.com/colaboratory/faq.html.

[27] "Keras: Deep Learning for humans." Keras, Feb. 02, 2023, Accessed: Feb. 02, 2023. [Online]. Available: https://github.com/keras-team/keras.

[28] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." arXiv, Mar. 16, 2016, https://doi.org/10.48550/arXiv.1603.04467.

[29] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, Dec. 2017, pp. 1–10.

[30] B. Li, "Random Search Plus: A more effective random search for machine learning hyperparameters optimization," M.S. thesis, University of Tennessee, Knoxville, TN, United States, 2020.

[31] E. Ndiaye, T. Le, O. Fercoq, J. Salmon, and I. Takeuchi, "Safe Grid Search with Optimal Complexity," in *36th International Conference on Machine Learning*, Long Beach, CA, USA, Jun. 2019, pp. 4771–4780.

[32] M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting Corn Yield With Machine Learning Ensembles," *Frontiers in Plant Science*, vol. 11, Jul. 2020, Art. no. 1120, https://doi.org/10.3389/fpls.2020.01120.

[33] V. Barbosa dos Santos, A. M. F. dos Santos, and G. de S. Rolim, "Estimation and forecasting of soybean yield using artificial neural networks," *Agronomy Journal*, vol. 113, no. 4, pp. 3193–3209, 2021, https://doi.org/10.1002/agj2.20729.

[34] R. R. Fidelis, G. V. Miranda, I. C. dos Santos, J. C. C. Galvao, J. M. Peluzio, and S. de O. Lima, "Fontes de germoplasma de milho para estresse de baixo nitrogenio," *Pesquisa Agropecuaria Tropical*, vol. 37, no. 3, pp. 147–153, Oct. 2007.

[35] M. O. Soares, G. V. Miranda, L. J. M. Guimaraes, I. E. Marriel, and C. T. Guimaraes, "Parametros geneticos de uma populacao de milho em niveis contrastantes de nitrogenio," *Revista Ciencia Agronomica*, vol. 42, pp. 168–174, Mar. 2011, https://doi.org/10.1590/S1806-66902011000100021.

[36] I. C. D. Santos *et al.*, "Comportamento de cultivares de milho produzidos organicamente e correlacoes entre caracteristicas das espigas colhidas no estadio verde," *Revista Brasileira de Milho e Sorgo*, vol. 4, no. 1, pp. 45–53, 2005, https://doi.org/10.18512/1980-6477/rbms.v4n01p%p.

[37] G. de O. Garcia, P. A. Ferreira, G. V. Miranda, F. G. de Oliveira, and D. B. dos Santos, "Indices fisiologicos, crescimento e producao do milho irrigado com agua salina," *Irriga*, vol. 12, no. 3, pp. 307–325, Sep. 2007, https://doi.org/10.15809/irriga.2007v12n3p307-325.

[38] A. Borem, G. V. Miranda, and R. Fritsche-Neto, *Melhoramento de plantas*. Brazil, USA: Oficina de Textos, 2021.

[39] G. V. Miranda, E. M. W. Braun, M. E. V. B. Alves, P. Machado, and A. de M. Ramos, "Desempenho de hibridos de milho em diferentes epocas

de semeadura na segunda safra em baixa altitude no extremo Oeste do Estado do Parana," *Brazilian Journal of Development*, vol. 7, no. 4, pp. 34794–34810, Apr. 2021, https://doi.org/10.34117/bjdv7n4-100.

[40] F. R. Pires, C. M. Souza, D. M. Queiroz, G. V. Miranda, and J. C. C. Galvao, "Alteracao de atributos quimicos do solo e estado nutricional e caracteristicas agronomicas de plantas de milho, considerando as modalidades de calagem em plantio direto," *Revista Brasileira de Ciencia do Solo*, vol. 27, pp. 121–131, Feb. 2003, https://doi.org/10.1590/S0100-06832003000100013.

[41] S. Chakraborty, A. R. Belekar, A. Datye, and N. Sinha, "Isotopic study of intraseasonal variations of plant transpiration: an alternative means to characterise the dry phases of monsoon," *Scientific Reports*, vol. 8, no. 1, Jun. 2018, Art. no. 8647, https://doi.org/10.1038/s41598-018-26965-6.

[42] F. Morales *et al.*, "Photosynthetic Metabolism under Stressful Growth Conditions as a Bases for Crop Breeding and Yield Improvement," *Plants*, vol. 9, no. 1, Jan. 2020, Art. no. 88, https://doi.org/10.3390/plants9010088.

[43] Y. Li, H. Tao, B. Zhang, S. Huang, and P. Wang, "Timing of Water Deficit Limits Maize Kernel Setting in Association With Changes in the Source-Flow-Sink Relationship," *Frontiers in Plant Science*, vol. 9, 2018, Art. no. 1326, https://doi.org/10.3389/fpls.2018.01326.

[44] J. L. Hatfield and J. H. Prueger, "Temperature extremes: Effect on plant growth and development," *Weather and Climate Extremes*, vol. 10, pp. 4–10, Dec. 2015, https://doi.org/10.1016/j.wace.2015.08.001.

[45] E. A. Minato, B. M. A. R. Cassim, M. R. Besen, F. L. Mazzi, T. T. Inoue, and M. A. Batista, "Controlled-release nitrogen fertilizers: characterization, ammonia volatilization, and effects on second-season corn," *Revista Brasileira de Ciencia do Solo*, vol. 44, May 2020, Art. no. e0190108, https://doi.org/10.36783/18069657rbcs20190108.

[46] H. A. Cleugh, J. M. Miller, and M. Bohm, "Direct mechanical effects of wind on crops," *Agroforestry Systems*, vol. 41, no. 1, pp. 85–112, Apr. 1998, https://doi.org/10.1023/A:1006067721039.

[47] C. C. Westhues *et al.*, "Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks," *Frontiers in Plant Science*, vol. 12, 2021, Art. no. 699589, https://doi.org/10.3389/fpls.2021.699589.

[48] A. A. Chassaigne-Ricciulli, L. E. Mendoza-Onofre, L. Cordova-Tellez, A. Carballo-Carballo, F. M. San Vicente-Garcia, and T. Dhliwayo, "Effective Seed Yield and Flowering Synchrony of Parents of CIMMYT Three-Way-Cross Tropical Maize Hybrids," *Agriculture*, vol. 11, no. 2, Feb. 2021, Art. no. 161, https://doi.org/10.3390/agriculture11020161.

[49] R. Fritsche-Neto, R. A. Vieira, C. A. Scapim, G. V. Miranda, and L. M. Rezende, "Updating the ranking of the coefficients of variation from maize experiments," *Acta Scientiarum. Agronomy*, vol. 34, pp. 99–101, Mar. 2012, https://doi.org/10.4025/actasciagron.v34i1.13115.