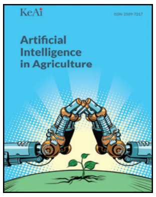
Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Yield performance estimation of corn hybrids using machine learning algorithms

Farnaz Babaie Sarijaloo [a,*], Michele Porta [b], Bijan Taslimi [a], Panos M. Pardalos [a]

[a] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA
[b] Department of Computer Engineering, University of Parma, Parma, Italy

ARTICLE INFO

Article history:
Received 20 November 2020
Received in revised form 5 May 2021
Accepted 5 May 2021
Available online 7 May 2021

Keywords:
Yield prediction
Data analysis
Corn hybrids
Machine learning
Agricultural data

ABSTRACT

Estimation of yield performance for crop products is a topic of interest in agriculture. In breeding programs, we cannot test all possible hybrids created by crossing two parents (inbred and tester) since it would be too time consuming and costly. In this paper, we exploit different machine learning algorithms including decision tree, gradient boosting machine, random forest, adaptive boosting, XGBoost and neural network to predict the yield of corn hybrids using data provided in the 2020 Syngenta Crop Challenge. The participants were asked to predict the yield of missing hybrids which were not tested before. Our results show that the prediction obtained by XGBoost is more accurate than other models with a root mean square error equal to 0.0524. Therefore, we use XGBoost model to estimate the yield performance for untested combinations of inbreds and testers. Using this approach, we identify hybrids with high predicted yield that can be bred to increase corn production.

1. Introduction

Continuous increase in the world population together with the substantial environmental issues, namely undesirable weather conditions caused by climate change, have led to a global concern to secure food resources such as crops (Godfray et al., 2010). The prediction of crop yield has a direct impact on national and international economies and plays a crucial role in food management all around the world. Therefore, it is very important to assist farmers and food industries with accurate models to predict crop productivity. This will help them to assess their current situation in terms of productivity and efficiency, and identify possible enhancements such as improving soil quality or choosing high-performance crop breeds. Furthermore, managers can utilize such prediction models in their decision making process related to agronomy and crop choice policies to minimize losses when undesirable environmental conditions like drought and pest problems occur (Aubry et al., 1998).

With advancements in computational technology and data collection, an essential need of developing novel models for yield prediction has emerged. To this end, researchers have presented various yield prediction models including traditional machine learning algorithms, data mining methods and deep neural network models (Liakos et al., 2018; Chlingaryan et al., 2018; Vlontzos and Pardalos, 2017; Mucherino et al., 2009; Papajorgji et al., 2006). In recent years, deep-learning based models have been widely used in crop yield prediction. These techniques can model complex inputs with complicated interactions. You et al. (2017) introduced a deep Gaussian process framework to predict crop yield based on remote sensing data. In this study, relevant features from raw data were effectively extracted using a novel dimensionality reduction method. Elavarasan and Vincent (2020) proposed a recurrent neural network deep learning algorithm over the Q-learning reinforcement learning algorithm to predict yield. Based on their experiments, this approach can predict yield with an accuracy of 93.7% using environmental, soil, water and crop parameters. In another study, a deep convolutional neural network model was developed to extract key features from normalized difference vegetation index (NDVI) and RGB data to predict yield (Nevavuori et al., 2019).

Corn hybrids and their performance evaluation have also been a topic of interest in this field. During the last few decades corn (also called maize) has become an important product among crops (Burchfield and Schumacher, 2020). Besides the diverse usage of corn in the food industry, it is also used for making animal food, bio-fuel and other industrial products (Chavas and Mitchell, 2018). Due to the increasing demand of corn, farmers are seeking new ways to improve their productivity. Hybrids are new seeds created by crossing two parents together in order to obtain a potentially stronger and more productive seed (Crow, 1998) which might lead to higher corn yield. They are shown to incorporate favorable qualitative traits and be adapted to different conditions. An important feature of hybrids is their higher tolerance to drought compared to natural breeds (Crow, 1998). Identifying the hybrids with maximum predicted yield has always been a challenging problem in agriculture. Several empirical breeding programs were developed to explore the impact of the breeding on crop yield and

[*] Corresponding author.
E-mail address: farnazbs@ufl.edu (F. Babaie Sarijaloo).

https://doi.org/10.1016/j.aiia.2021.05.001

resistance to diseases and environmental factors like temperature and water loss (Duvick et al., 2004). In these programs, outstanding hybrids from inbred lines are selected. Results obtained from breeding programs can provide farmers and managers options to manage crops and pick the best strategy for planting schedules (Cooper et al., 2014). In breeding programs, hybrid performance is evaluated from extensive yield trials that are costly and time consuming (Lanza et al., 1997). While this process traditionally used to be done only by trial and error, breeders are now exploiting data analysis techniques to predict hybrids yield accurately prior to field evaluation. Therefore, there is a remarkable demand for new approaches to predict the performance of hybrids.

There are various models in the literature which tackled hybrid performance prediction problem. They have utilized different methods depending on their objectives and data. Bernardo (1996) proposed to use best linear unbiased prediction model to identify high-yielding single-cross corn hybrids. This method and its extensions were also applied to other crops such as soybean and wheat (Panter and Allen, 1995; Arruda et al., 2015). In another study researchers identified corn hybrids tolerant to drought and heat using an unsupervised approach based on a stress metrics obtained from a deep learning model (Khaki et al., 2019). With recent progress in genome sequencing, phenotypic and DNA marker characterizations provide helpful information to determine the best parents with expected traits in different corn uses like food industry (Somerville and Somerville, 1999). Riedelsheimer et al. (2012) have shown that they can achieve a reliable accuracy in predicting biomass and bioenergy-related traits in inbred and testers crosses using single-nucleotide polymorphisms (SNP) and metabolites. They used ridge regressionâ€"best linear unbiased prediction model to predict polygenic traits in corn hybrids. In another study, researchers compared performance of three models including multi layer perceptron (MLP), support vector machine (SVM) and Bayesian threshold genomic best linear unbiased prediction (TGBLUP) models in predicting ordinal traits in plant breeding (Montesinos-López et al., 2019). Khaki et al. (2020) introduced an ensemble of matrix factorization and neural network models to predict corn hybrid performance. Their proposed model outperformed deep factorization machines, least absolute shrinkage and selection operator (LASSO), and random forest.

In this paper, we aim to utilize machine learning algorithms to predict crop yield for new corn hybrids using the data provided by *2020 Syngenta crop challenge*. In this challenge, participants were provided a dataset containing the yield for 199,476 combinations of inbred and testers which are grown in 280 different locations within years 2016 to 2018. We first perform a data processing on the provided dataset to do feature selection. We then apply different algorithms including *decision tree* (Safavian and Landgrebe, 1991), *gradient boosting machine* (GBM) (Friedman, 2002), *random forest* (Svetnik et al., 2003; Breiman, 2001), *adaptive boosting* (Adaboost) (Freund and Schapire, 1997), *extreme gradient boosting* (XGBoost) (Chen and Guestrin, 2016) and *neural network* (Liu et al., 2017; Sze et al., 2017) on the data and evaluate their performance. We select the best model among them and utilize it to predict yield for new combinations.

2. Materials & methods

2.1. Data source & summary

This study was performed as a part of the 2020 Syngenta Crop Challenge, which focuses on estimating yield performance of the cross between inbred and tester combinations to identify the best parent combinations. The participants were given a dataset containing the observed yield for 199,476 corn hybrids tested across 280 environments between 2016 and 2018. These hybrids are created through crossing of 593 unique inbreds and 496 unique testers.

Due to confidentiality concerns, the yield values are scaled to an internal benchmark to make the average and standard deviation of yields 1.0017 and 0.1047, respectively. These hybrids were tested in 280 unique locations. On average in each location 712 hybrids have been tested during the study period. 21% of the observations were collected in 2016 (with average yield 1.0082 and standard deviation 0.1060), 42% in 2017 (with average yield 0.9990 and standard deviation 0.0980), and 37% in 2018 (with average yield 1.0003 and standard deviation 0.1110).

With 593 inbreds and 496 testers, there are 294,128 possible hybrids, among which only 10,919 (4%) hybrids have been tested. This means around 96% of the hybrids are missing and we need to develop a model to predict yield for these missing combinations.

The inbreds and testers are clustered based on an internal analysis performed by the problem owner. Inbreds and testers are not treated any differently when clustering. Therefore, same cluster number indicates genetic similarity regardless of whether a parent is defined as an inbred or a tester. Therefore, Inbreds and testers with same cluster ID are considered independent. The inbreds and testers were clustered into 14 and 13 groups, respectively. 47.8% of the crossing inbred and tester clusters were not tested in this study (Fig. 1). In the given dataset, 95 unique combinations of inbred and tester clusters are provided for the performance analysis. These clusters were tested on average 2100 times in different locations during three years.

2.2. Input variables and outcome

In the given dataset, each observation consists of six input variables including location, year, inbred, inbred cluster, tester and tester cluster, and an outcome variable. The outcome is defined as the ratio of the hybrid yield to the benchmark in that environment.

2.3. Analytics workflow

To check whether the outcome depends on all the given input variables, we ran a GBM model (with 100 trees and shrinkage rate equal to 0.001) on the dataset. This variable selection helps to reduce overfitting. In GBM model, we can calculate relative importance for variables which indicates how significant each variable was in the construction of the boosted decision trees within the model. Variables with zero relative importance do not have any impact on the outcome. After removing the insignificant variables (the variables with zero relative importance), the mean yield of the observations whose independent variables are identical was calculated to reduce the variability and avoid having multiple records of the same observations because high variability might mislead the model in predicting yield and affect the performance. The final dataset has 10,919 unique hybrids with mean yield calculated for each hybrid. Fig. 2 shows the conceptual diagram of what we have done in this study.

We randomly picked 80% of the data as the training and the left 20% as the testing set. This splitting has been repeated 5 times using 5 different seeds to avoid overfitting. We ran multiple machine learning models including Adaboost, decision tree, random forest, neural network, XGBoost and GBM on the training set and evaluated the models using the testing set. This procedure was performed on 5 different sets of training and testing sets randomly chosen using five different seeds.

After preprocessing the data, we encoded all the data to convert the categorical variables into numbers to have better understanding by machine learning algorithms. In order to find the best encoding method, we performed three methods on the data including one hot encoding, integer encoding and feature hashing encoding (Potdar et al., 2017; Fitkov-Norris et al., 2012; Seger, 2018). In integer encoding, we convert the categorical data into integers. Data encoded using one hot encoding has a binary variable for each category and in feature hashing encoding,

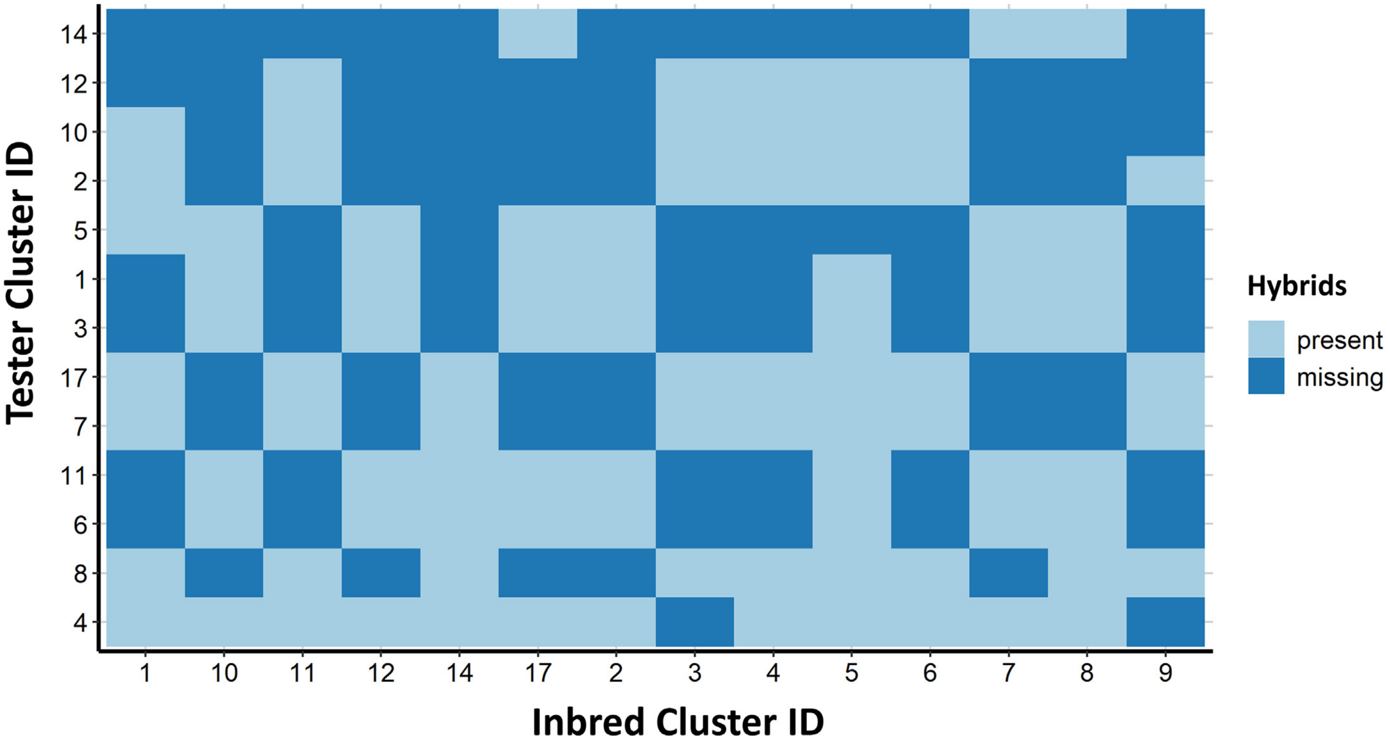


Fig. 1. Missing and present hybrids in the provided dataset created through crossing inbred and tester clusters. Missing rate is 47.8%. The goal is to develop a model using the available tested hybrids and estimate the performance of missing inbreds and testers.

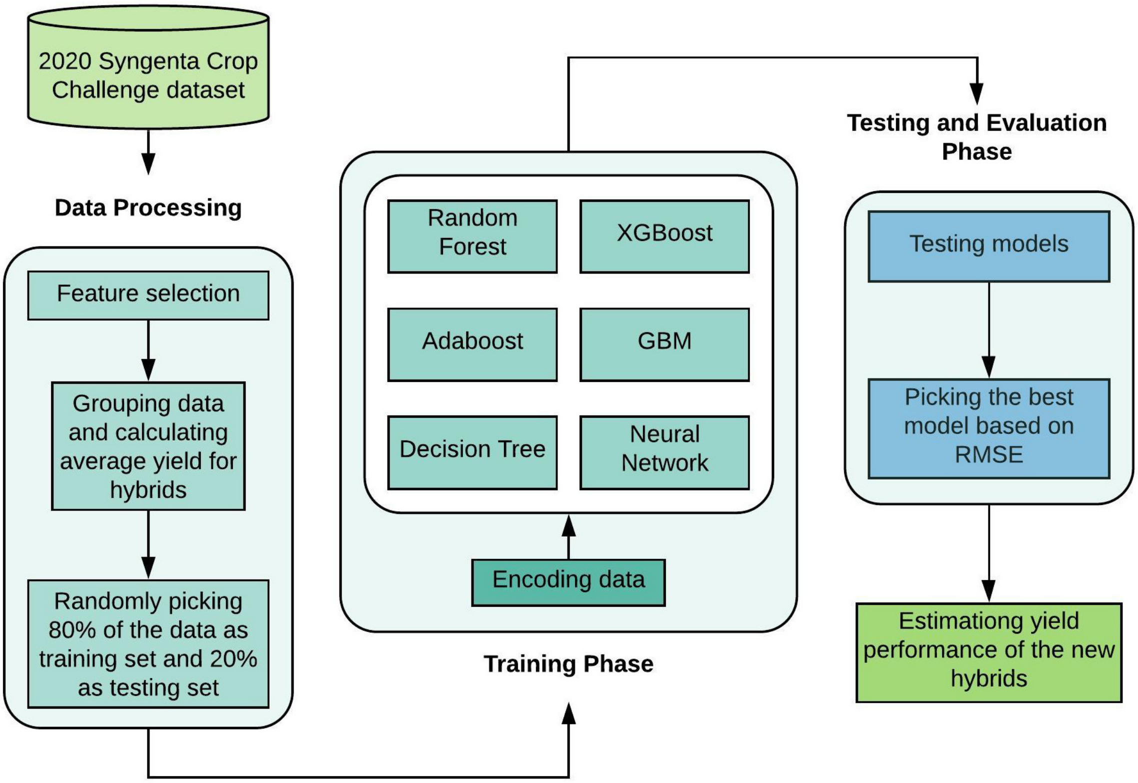

2020 Syngenta Crop Challenge dataset
Data Processing
Feature selection
Grouping data and calculating average yield for hybrids
Randomly picking 80% of the data as training set and 20% as testing set
Random Forest
XGBoost
Adaboost
GBM
Decision Tree
Neural Network
Encoding data
Training Phase
Testing and Evaluation Phase
Testing models
Picking the best model based on RMSE
Estimatiog yield performance of the new hybrids

Fig. 2. Flow diagram of the 2020 Syngenta crop challenge analysis; Predicting yield performance of the cross between inbred and tester combinations. This diagram shows the sequence of steps from data processing, model training and testing leading to new hybrids yield prediction.

a hash function is applied on the data to map them into small finite set of values. The encoding methods were assessed using root mean square error (RMSE) obtained in the testing phase. Yield performance is unitless based on its definition. Therefore, RMSE of the models is unitless, as well. Then, we ran the models on the testing data and picked the one that gave us the lowest RMSE. In the training phase, parameters for the models were tuned to minimize RMSE. The final model was picked based on the lowest RMSE obtained in the testing phase.

The parameters of the models were tuned as follows:

- Adaboost (with maximum depth 3, 1000 estimators, learning rate 1 and a linear loss function),
- decision tree (with minimum number of samples required to split an internal node equal to 2 and maximum depth is set as default which means nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples),
- random forest (with 10 estimators and maximum number of variables in each split equal to total number of variables)
- neural network (epoch 50, batch size 32, and a dropout layer has been added to every hidden layer [3 layers in total] with a dropout rate of 0.5),
- XGBoost (with 1000 estimators, maximum depth 5, $\alpha = 10$, and sub sample ratio of columns when constructing each tree is 0.5) and,
- GBM (with 1000 trees and shrinkage rate equal to 0.05)

The analyses were performed using Python 3.7.5.

3. Results

3.1. Variable selection

To evaluate the effect of given variables on the yield, we utilized a GBM model. The results indicated that the only variables with nonzero relative importance were inbred, tester and inbred cluster and tester cluster. As shown in Fig. 3, variable tester has the highest relative importance (63.5) among all of the variables and, variables year and location have appeared in the model with zero relative importance. Therefore, we concluded that the environmental effects across the locations during the 3-year time period for all the hybrids does not have a significant impact on the crop yield. Thus, we removed these two variables from the dataset. The final dataset contains 10,919 unique hybrids (with average yield 0.9960 and standard deviation 0.0586).

3.2. Model development and performance analysis

Five different experiments were conducted using different seeds to test the models. The data were encoded using three methods. The best encoding method was one hot encoding by which we obtained the minimum RMSE among all the models. Fig. 4 demonstrates the performance of the models developed using data encoded with three encoding methods.

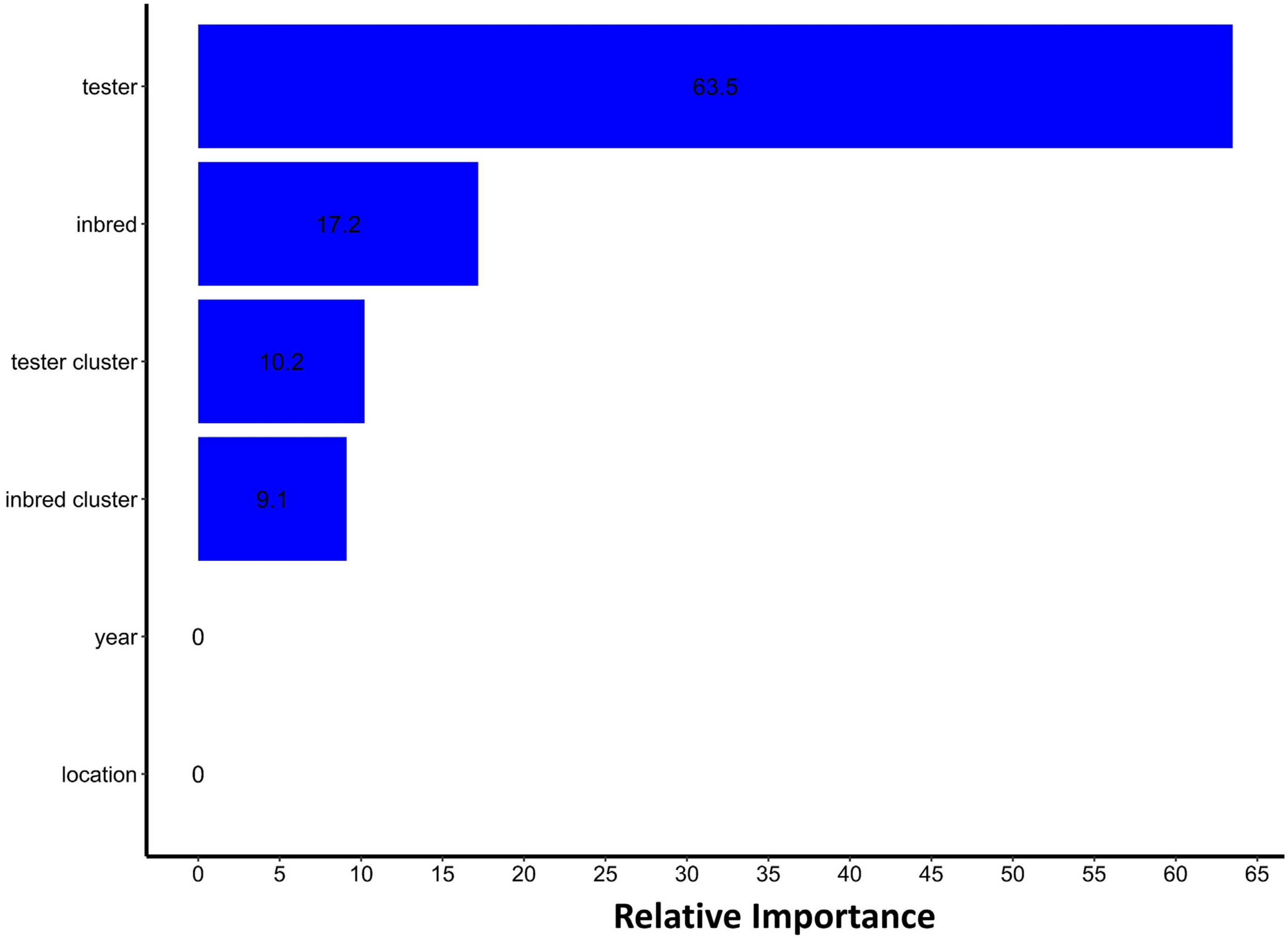


Fig. 3. Relative importance of the variables in the feature selection process. A GBM model was utilized to identify the significant variables for further inclusion in the training procedure. Variables tester, inbred, tester cluster and inbred cluster have non-zero relative importance in this model.

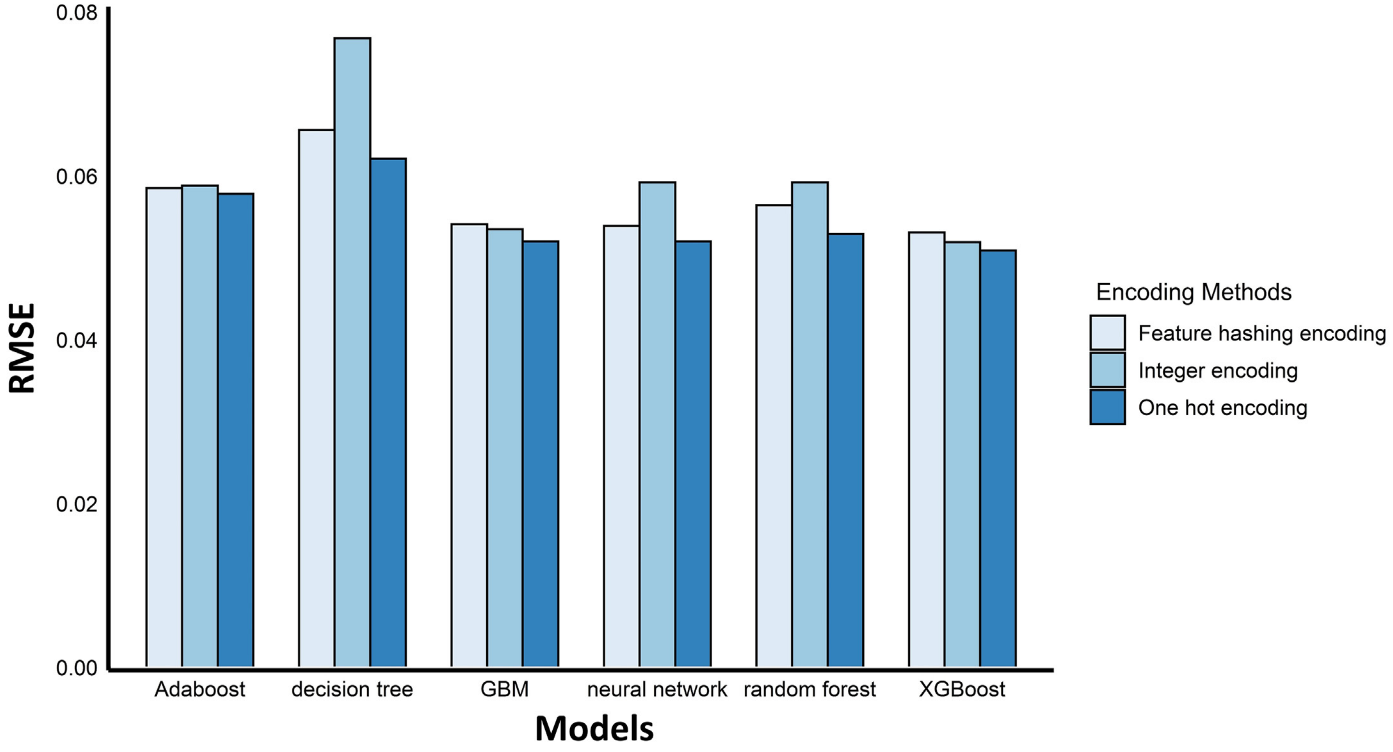


Fig. 4. Comparison of the model performance using different encoding methods. The performance of the models was assessed using RMSE. The models with one hot encoding have been observed to have the lowest RMSE (on average 0.0546).

Table 1
RMSE of the machine learning models in each experiment using data encoded with one-hot encoding method.

Experiment	Decision tree	Adaboost	GBM	Random forest	Neural network	XGBoost
1	0.0644	0.0596	0.0548	0.0559	0.0539	0.0537
2	0.0655	0.0585	0.0534	0.0553	0.0520	0.0524
3	0.0667	0.0585	0.0540	0.0550	0.0510	0.0526
4	0.0621	0.0578	0.0520	0.0529	0.0520	0.0509
5	0.0646	0.0591	0.0539	0.0543	0.0539	0.0525

Table 1 displays the performance of the machine learning models with 5 different seeds. The data fed into these models was encoded using one-hot encoding method. As shown in Table 1, XGBoost achieved the minimum RMSE in 3 of the experiments and neural network has a slightly better performance in 2 of them. However, on average XGBoost outperformed the other machine learning models.

The RMSE achieved in the testing phase was 0.0561±0.0043 (average ± 95% confidence interval). Among them XGBoost and neural network model had the minimum RMSE. According to the results (Fig. 4), we decided to encode the data using one hot encoding method and predict the yield for the hybrids using the XGBoost model which provided the lowest RMSE (0.0524) in the testing phase.

3.3. Comparing implemented models

As observed in Fig. 4 and Table 1, XGBoost outperformed all the models due to using advanced regularization in loss function. On the contrary, decision tree algorithm achieved the worst RMSE among the tested models. This poor performance mostly happens because decision tree models are week learners and prone to overfitting. The best RMSE in each of the models is obtained when the data is encoded using one hot encoding. The performance of XGBoost algorithm with integer encoding is slightly better than feature hash encoding. The same pattern holds true for Adaboost and GBM models. Random forest is observed to have better performance compared to Adaboost and decision tree using data encoded by one hot encoding and feature hashing encoding methods. Neural network model outperformed Adaboost and random forest in every encoding methods.

3.4. Decision making by predicting yield performance of new hybrids

We applied our selected model on 1200 combinations of inbreds and testers which were not tested before. The predicted yield performance had an average equal to 0.9963 with standard deviation 0.0212 (Fig. 5). Among these combinations, the maximum and minimum yield performance are 1.0610 and 0.8798, respectively. Moreover, we observed that 38% of the combinations have a predicted yield performance higher than 1 which means their predicted yield is higher than the benchmark in the corresponding environment. Decision makers can further use these predictions in their breeding programs to obtain a superior level of yield performance. The combinations were grouped by inbred cluster and tester cluster and the average of predicted yields was calculated for each hybrid. The heat map of this data is shown in Fig. 6. Using this figure, we can identify the combinations with high performance.

3.5. Determining the role of a cluster in a hybrid

In the provided data set, there are 14 clusters defined by the problem owner. In breeding programs, depending on the role of a parent in a hybrid, whether it is an inbred or a tester, the same yield performance might not be obtained. Fig. 7 is provided to compare the distribution of hybrids yield performance when we use these clusters in two main roles, inbred and tester. As depicted in Fig. 7, clusters 1, 4, 5, 7 and 10 appeared to have yield distributions almost the same when these clusters were used in a hybrid as inbred or tester. On the contrary, we observe

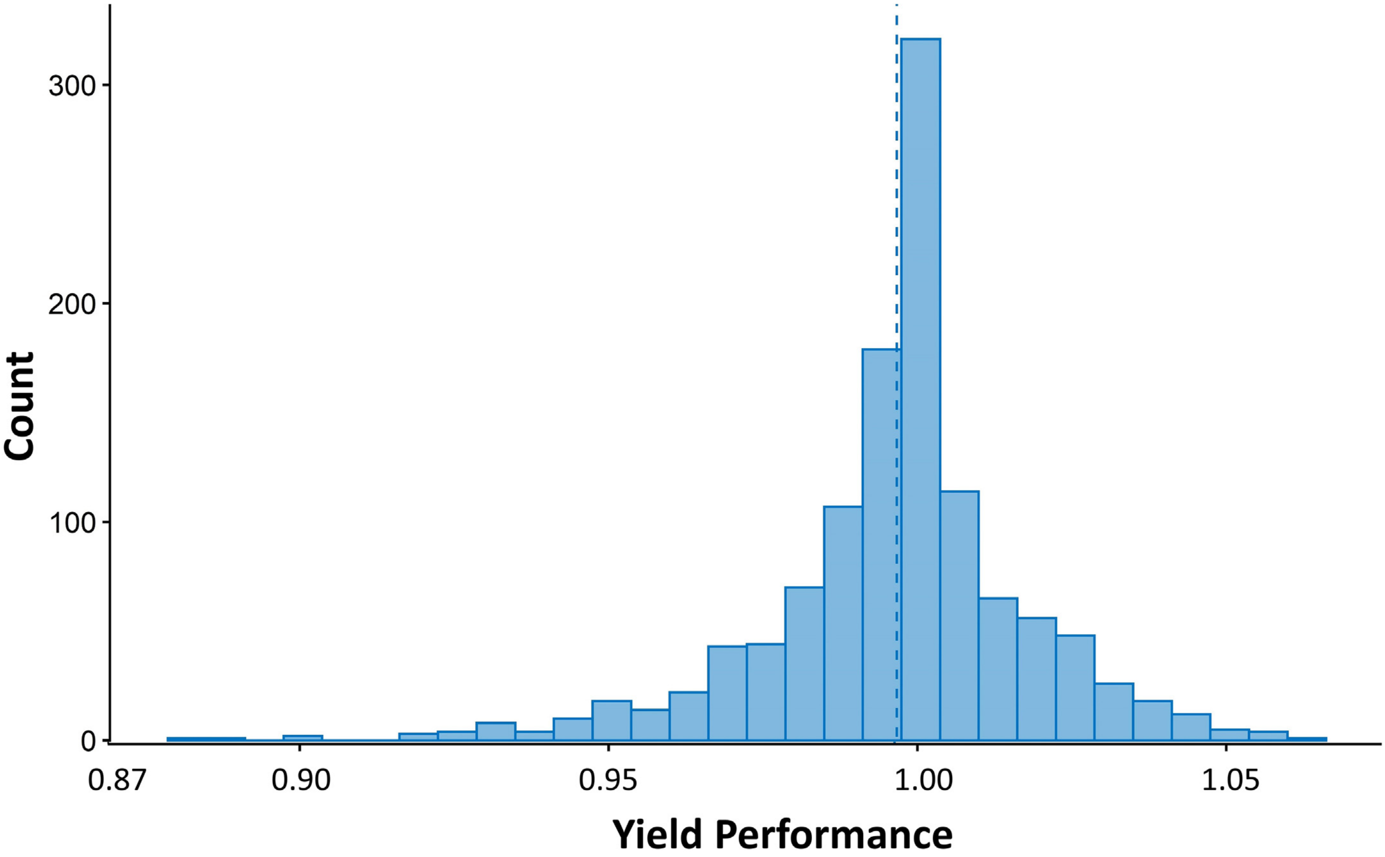


Fig. 5. Distribution of the predicted yield performance for the untested hybrids by the XGBoost model. Yield performance is defined as the ratio of the hybrid performance to the benchmark in that environment. 38% of the hybrids were predicted to have yield performance greater than 1. The dashed line indicates the average (0.9963).

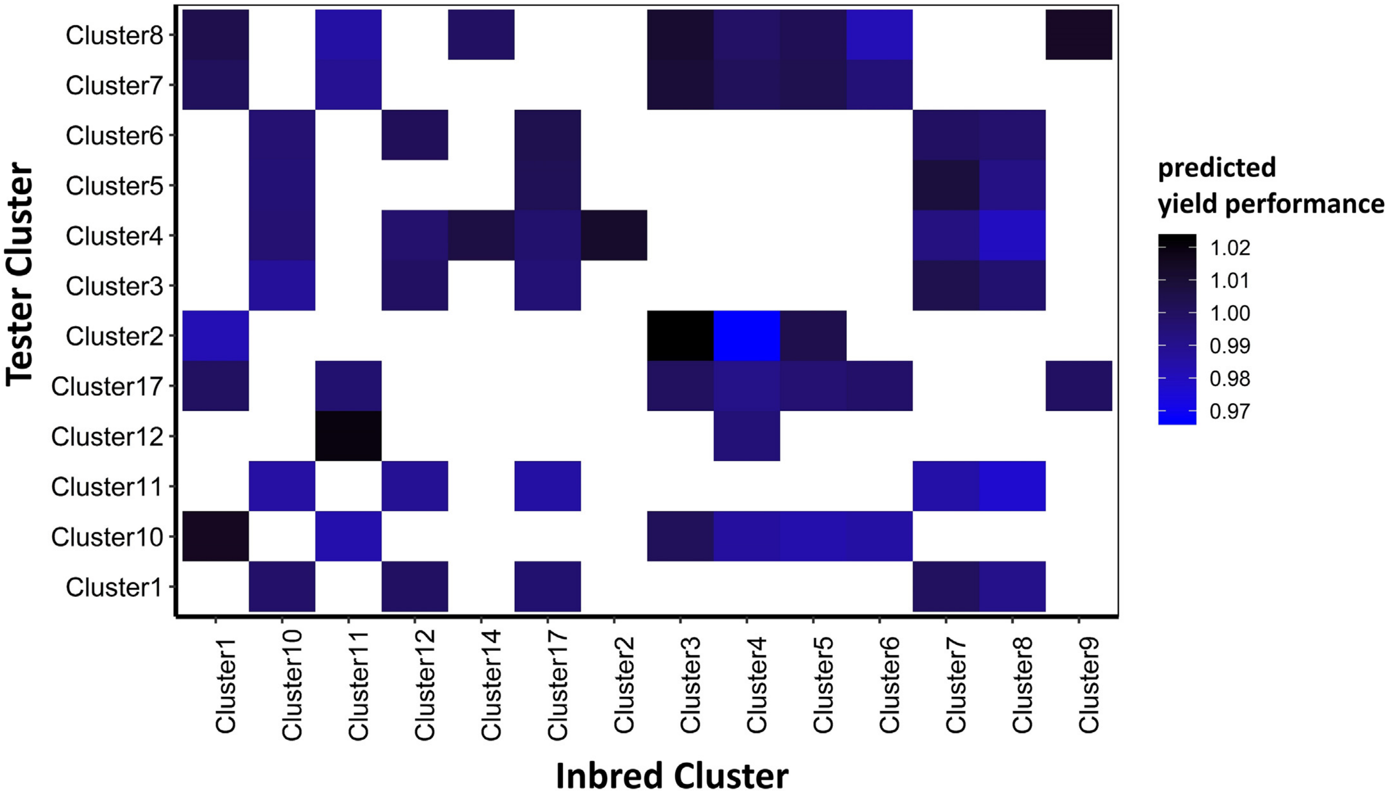


Fig. 6. Heat map of the average predicted yield performance for untested combinations of inbred and tester clusters.

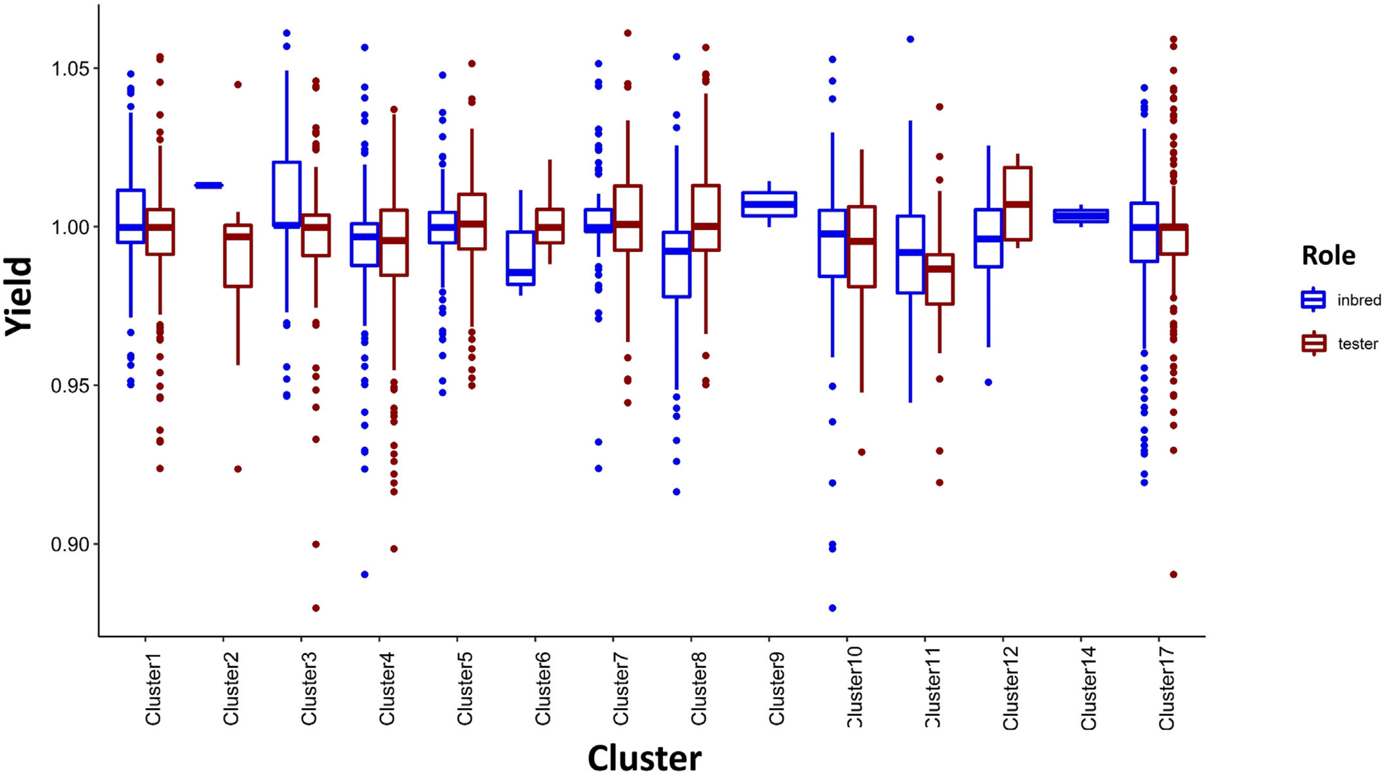


Fig. 7. Distribution of yield performance of hybrids with clusters as inbred or tester. In each hybrid, there are two roles that a cluster can play, tester or inbred. The hybrids might show different performance when a specific cluster is used as a tester or an inbred.

that the same cluster would produce higher yield when it is the inbred in a hybrid like clusters 2, 3 and 11 compared to the hybrids where we have these clusters as tester. Clusters 9 and 14 were only used as inbreds in this dataset. This information can give decision makers a rough idea on how to maximize yield by choosing which clusters to play the role of inbred and which cluster to be selected as tester in a hybrid.

4. Discussion

In this study, advanced machine learning techniques were applied to create a yield estimation model for corn hybrids. Our proposed approach is unique from other previous prediction models in that encoding step before training the model. Depending on what encoding method is applied, the performance of models can be improved. As shown in the results section, XGBoost model with one-hot encoded data outperformed the other models.

In our study, XGBoost algorithm has the best performance among the tested machine learning models due to the following reasons: 1) XGBoost can handle sparsity in the input variables more efficiently compared to other tested models (Chen and Guestrin, 2016) and, 2) due to the regularization used in XGBoost model, this algorithm can control overfitting which leads to higher performance (Nielsen, 2016). This model has been widely used in different fields such as agriculture, healthcare and finance (Ogunleye and Wang, 2019; Torlay et al., 2017; Nobre and Neves, 2019). In our study XGBoost has shown to have the best performance compared to other models. Same result was obtained in several studies. Qin et al. (2018) proposed an XGBoost model to predict corn economic optimal nitrogen rate and this model outperformed ridge regression and LASSO in specific time windows of their study. In another study, XGBoost algorithm exhibited higher accuracy compared to random forest in predicting soybean yield using data acquired by unnamed aircraft system (UAS) (Herrero-Huerta et al., 2020).

The proposed model can be used by managers or farmers to maximize yield by selecting the best hybrid. This machine learning tool can be beneficial to the decision makers in two ways. First, since we only used 4% of the possible hybrids to develop the model, it will help them to lower cost by decreasing the number of hybrid breeding trials. Second, the whole breeding process is not time-consuming anymore because there is no need to wait for the results from field trials. This information can be obtained from the model in seconds.

For the future work, we firstly suggest to define a new criterion for inbreds and testers clustering. In the current study, inbreds and testers were merely clustered based on the genetic features. Thus, further research could address novel clustering approaches that might be more informative for yield prediction. Secondly, to attain a more accurate prediction model, we could include new features in the dataset. Crop production environment consists of numerous factors such as pests, diseases and human activities which can cause substantial variations in the crop yield. Although we showed that the variables year and location do not have any significant impact on the yield performance, this conclusion might have been obtained because of the counteraction between some factors which are not provided in the dataset. Therefore, including further factors to the study could improve the performance of the models.

5. Conclusion

This paper presents an application of machine learning algorithms in agricultural data analysis. We implemented various methods on the dataset provided in the 2020 Syngenta crop challenge in order to estimate the performance of new corn hybrids using available tested hybrids. The results in the testing phase implied that XGBoost algorithm has a superior performance among all implemented techniques. We next predicted the yield performance of the new combinations of inbreds and testers using our proposed XGBoost model and presented the summary of results. The results demonstrated that more than one third of the new hybrids had indeed a higher yield performance compared to the benchmark in the corresponding environment. These hybrids are identified as potentially more productive seeds. Therefor, our proposed approach can be utilized by decision makers in breeding programs to select outstanding hybrids from inbred lines only by using a sample of tested hybrids. Using machine learning tools such as the model developed in this study can lead to higher yield and lower cost in agricultural industry.

Conflict of interest statement

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

Data availability statement

The dataset for this study can be found in the 2020 Syngenta crop challenge website. The data was publicly available for the researchers who registered to participate in the challenge during September 2019 to January 2020.

Acknowledgments

We would like to appreciate Syngenta and the Analytics Society of INFORMS for organizing the 2020 Syngenta crop challenge.

References

Arruda, M.P., Brown, P.J., Lipka, A.E., Krill, A.M., Thurber, C., Kolb, F.L., 2015. Genomic selection for predicting fusarium head blight resistance in a wheat breeding program. Plant Genome 8, 1–12.

Aubry, C., Papy, F., Capillon, A., 1998. Modelling decision-making processes for annual crop management. Agric. Syst. 56, 45–65.

Bernardo, R., 1996. Best linear unbiased prediction of maize single-cross performance. Crop Sci. 36, 50–56.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Burchfield, E.K., Schumacher, B.L., 2020. Bright spots in us corn production. Environ. Res. Lett. 15, 104019.

Chavas, J.P., Mitchell, P.D., 2018. Corn Productivity: The Role of Management and Biotechnology. Production and Human Health in Changing Climate, Corn, p. 13.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69.

Cooper, M., Messina, C.D., Podlich, D., Totir, L.R., Baumgarten, A., Hausmann, N.J., Wright, D., Graham, G., 2014. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci. 65, 311–336.

Crow, J.F., 1998. 90 years ago: the beginning of hybrid maize. Genetics 148, 923–928.

Duvick, D., Smith, J., Cooper, M., et al., 2004. Long-term selection in a commercial hybrid maize breeding program. Plant Breed. Rev. 24, 109–152.

Elavarasan, D., Vincent, P.D., 2020. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access 8, 86886–86901.

Fitkov-Norris, E., Vahid, S., Hand, C., 2012. Evaluating the impact of categorical data encoding and scaling on neural network classification performance: the case of repeat consumption of identical cultural goods. International Conference on Engineering Applications of Neural Networks. Springer, pp. 343–352.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. Science 327, 812–818.

Herrero-Huerta, M., Rodriguez-Gonzalvez, P., Rainey, K.M., 2020. Yield prediction by machine learning from uas-based mulit-sensor data fusion in soybean. Plant Methods 16, 1–16.

Khaki, S., Khalilzadeh, Z., Wang, L., 2019. Classification of crop tolerance to heat and drought: a deep convolutional neural networks approach. Agronomy 9, 833.

Khaki, S., Khalilzadeh, Z., Wang, L., 2020. Predicting yield performance of parents in plant breeding: a neural collaborative filtering approach. PLoS One 15, e0233382.

Lanza, L., de Souza Jr, C., Ottoboni, L., Vieira, M.L.C., De Souza, A., 1997. Genetic distance of inbred lines and prediction of maize single-cross performance using rapd markers. Theor. Appl. Genet. 94, 1023–1030.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors 18, 2674.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26.

Montesinos-López, O.A., Martn-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C.M., Montesinos-López, A., Juliana, P., Singh, R., 2019. A benchmarking between deep learning, support vector machine and bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. G3 9, 601–618.

Mucherino, A., Papajorgji, P., Pardalos, P.M., 2009. A survey of data mining techniques applied to agriculture. Oper. Res. 9, 121–140.

Nevavuori, P., Narra, N., Lipping, T., 2019. Crop yield prediction with deep convolutional neural networks. Comput. Electron. Agric. 163, 104859.

Nielsen, D., 2016. Tree Boosting with Xgboost. NTNU Norwegian University of Science and Technology.

Nobre, J., Neves, R.F., 2019. Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets. Expert Syst. Appl. 125, 181–194.

Ogunleye, A., Wang, Q.G., 2019. Xgboost model for chronic kidney disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 17, pp. 2131–2140.

Panter, D., Allen, F., 1995. Using best linear unbiased predictions to enhance breeding for yield in soybean: ii. Selection of superior crosses from a limited number of yield trials. Crop Sci. 35, 405–410.

Papajorgji, P.J., Pardalos, P.M., et al., 2006. Software Engineering Techniques Applied to Agricultural Systems. Springer.

Potdar, K., Pardawala, T.S., Pai, C.D., 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. 175, 7–9.

Qin, Z., Myers, D.B., Ransom, C.J., Kitchen, N.R., Liang, S.Z., Camberato, J.J., Carter, P.R., Ferguson, R.B., Fernandez, F.G., Franzen, D.W., et al., 2018. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. Agron. J. 110, 2596–2607.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., Melchinger, A.E., 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44, 217.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. 21, 660–674.

Seger, C., 2018. An Investigation of Categorical Variable Encoding Techniques in Machine Learning: Binary Versus One-Hot and Feature Hashing.

Somerville, C., Somerville, S., 1999. Plant functional genomics. Science 285, 380–383.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and qsar modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Sze, V., Chen, Y.H., Yang, T.J., Emer, J.S., 2017. Efficient processing of deep neural networks: a tutorial and survey. Proc. IEEE 105, 2295–2329.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M., 2017. Machine learning–xgboost analysis of language networks to classify patients with epilepsy. Brain Inform. 4, 159–169.

Vlontzos, G., Pardalos, P.M., 2017. Data mining and optimisation issues in the food industry. Int. J. Sustain. Agric. Manag. Inform. 3, 44–64.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. Thirty-First AAAI Conference on Artificial Intelligence.