Crop yield prediction is a highly complex trait determined by multiple factors such as genotype, environment, and their interactions. Accurate yield prediction requires fundamental understanding of the functional relationship between yield and these interactive factors, and to reveal such relationship requires both comprehensive datasets and powerful algorithms. In the 2018 Syngenta Crop Challenge, Syngenta released several large datasets that recorded the genotype and yield performances of 2,267 maize hybrids planted in 2,247 locations between 2008 and 2016 and asked participants to predict the yield performance in 2017. As one of the winning teams, we designed a deep neural network (DNN) approach that took advantage of state-of-the-art modeling and solution techniques. Our model was found to have a superior prediction accuracy, with a root-mean-square-error (RMSE) being 12% of the average yield and 50% of the standard deviation for the validation dataset using predicted weather data. With perfect weather data, the RMSE would be reduced to 11% of the average yield and 46% of the standard deviation. Our computational results suggested that this model significantly outperformed other popular methods such as Lasso, shallow neural networks (SNN), and regression tree (RT).

Keywords Yield prediction · Machine Learning · Deep Learning

1 Introduction

Crop yield prediction is of great importance to global food production. Policy makers rely on accurate predictions to make timely import and export decisions to strengthen national food security [1]. Seed companies need to predict the performances of new hybrids in various environments to breed for better varieties [2]. Growers and farmers also benefit from yield prediction to make informed management and financial decisions [1]. However, crop yield prediction is extremely challenging due to numerous complex factors. For example, genotype information is usually represented by high-dimensional marker data, containing many thousands to millions of makers for each plant individual. The effects of the genetic markers need to be estimated, which may be subject to interactions with multiple environmental conditions and field management practices.

Many studies have focused on explaining the phenotype (such as yield) as an explicit function of the genotype (G), environment (E), and their interactions (G×E). One of the straightforward and common methods was to consider only additive effects of G and E and treat their interactions as noise [3,4]. A popular approach to study the G×E effect was to identify the effects and interactions of mega environments rather than more detailed environmental components. Several studies proposed to cluster the environments based on discovered drivers of G×E interactions [5,6]. Crossa et al. [7,8] used the sites regression and the shifted multiplicative models for G×E interaction analysis by dividing environments into similar groups. Burgueño et al. [9] proposed an integrated approach of factor analytic (FA) and linear...
mixed models to cluster environments and genotypes and detect their interactions. They also stated that FA model can improve predictability up to 6% when there were complex G×E patterns in the data [10]. Linear mixed models have also been used to study both additive and interactive effects of individual genes and environments [11][12].

More recently, machine learning techniques have been applied for crop yield prediction, including multivariate regression, decision tree, association rule mining, and artificial neural networks. A salient feature of machine learning models is that they treat the output (crop yield) as an implicit function of the input variables (genes and environmental components), which could be a highly non-linear and complex function. Liu et al. [13] employed a neural network with one hidden layer to predict corn yield using input data on soil, weather, and management. Drummond et al. [14] used stepwise multiple linear regression, projection pursuit regression, and neural networks to predict crop yield, and they found that their neural network model outperformed the other two methods. Marko et al. [15] proposed weighted histograms regression to predict the yield of different soybean varieties, which demonstrated superior performances over conventional regression algorithms. Romero et al. [16] applied decision tree and association rule mining to classify yield components of durum wheat.

In this paper, we use deep neural networks to predict yield, check yield, and yield difference of corn hybrids from genotype and environment data. Deep neural networks belong to the class of representation learning models that can find the underlying representation of data without handcrafted input of features. Deep neural networks have multiple stacked non-linear layers which transform the raw input data into higher and more abstract representation at each stacked layer [17]. Given the right parameters, deep neural networks are known to be universal approximator functions, which means that they can approximate almost any function, although it may be very challenging to find the right parameters [18][19].

Compared with the aforementioned neural network models in the literature, which were shallow networks with a single hidden layer, deep neural networks with multiple hidden layers are more powerful to reveal the fundamental non-linear relationship between input and response variables [17], but they also require more advanced hardware and optimization techniques to train. For example, the neural network’s depth (number of hidden layers) has significant impact on its performance. Increasing the number of hidden layers may reduce the classification or regression errors, but it may also cause the vanishing/exploding gradients problem that prevents the convergence of the neural networks [20][21][22]. Moreover, the loss function of the deep neural networks is highly non-convex due to having numerous non-linear activation functions in the network. As a result, there is no guarantee on the convergence of any gradient based optimization algorithm applied on neural networks [18]. There have been many attempts to solve the gradient vanishing problem, including normalization of the input data, batch normalization technique in intermediate layers, stochastic gradient descent (SGD) [23][24], and using multiple loss functions for intermediate layers [25]. However, none of these approaches would be effective for very deep networks. He et al. [20] argued that the biggest challenge with deep neural networks was not overfitting, which can be addressed by adding regularization or dropout to the network [26], but it was the structure of the network. They proposed a new structure for deep neural networks using identity blocks or residual shortcuts to make the optimization of deeper networks easier [20]. These residual shortcuts act like a gradient highway throughout the network and prevent vanishing gradient problem.

Deep learning models have recently been used for crop yield prediction. You et al. [27] used deep learning techniques such as convolutional neural networks and recurrent neural networks to predict soybean yield in the United States based on a sequence of remotely sensed images taken before the harvest. Their model outperformed traditional remote-sensing based methods by 15% in terms of Mean Absolute Percentage Error (MAPE). Russello et al. [28] used convolutional neural networks for crop yield prediction based on satellite images. Their model used 3-dimensional convolution to include spatiotemporal features, and outperformed other machine learning methods.

The remainder of this paper is organized as follows. Section 2 describes the data used in this research. Section 3 provides a detailed description of our deep neural networks for yield prediction. Section 4 presents the results of our model. Finally, we conclude the paper in section 5.

2 Data

In the 2018 Syngenta Crop Challenge [2], participants were asked to use real-world data to predict the performance of corn hybrids in 2017 in different locations. The dataset included 2,267 experimental hybrids planted in 2,247 of locations between 2008 and 2016. This was one of the largest and most comprehensive datasets that were publicly available for research in yield prediction, which enabled the deployment and validation of the proposed deep neural network model.

The training data included three sets: crop genotype, yield performance, and environment (weather and soil). The genotype dataset contained genetic information for all experimental hybrids, each having 19,465 genetic markers. The
yield performance dataset contained the observed yield, check yield (average yield across all hybrids of the same location), and yield difference of 148,452 samples for different hybrids planted in different years and locations. Yield difference is the difference between yield and check yield, and indicates the relative performance of a hybrid against other hybrids at the same location [29]. The environment dataset contained 8 soil variables and 72 weather variables (6 weather variables measured for 12 months of each year). The soil variables included percentage of clay, silt and sand, available water capacity, soil pH, organic matter, cation-exchange capacity, and soil saturated hydraulic conductivity. Weather variables included precipitation, solar radiation, snow water equivalent, maximum temperature, minimum temperature, and vapor pressure. Part of the challenge was to predict the 2017 weather variables and use them for yield prediction of the same year.

The goal of the 2018 Syngenta Crop Challenge was to predict the performance of corns in 2017, but the ground truth response variables for 2017 were not released after the competition. In this paper, we used the 2001 to 2015 data and part of the 2016 data as the training dataset (containing 142,952 samples) and the remaining part of the 2016 data as the validation dataset (containing 5500 samples). All validation samples were unique combinations of hybrids and locations, which did not have any overlap with training data.

3 Methodology

3.1 Data Preprocessing

Approximately 37% of the genotype data had missing values. To address this issue, we used a two-step approach to preprocess the genotype data before they can be used by the neural network model. First, we used a 97% call rate to discard genetic markers whose non-missing values were below this call rate. Then we also discarded genetic markers whose lowest frequent allele’s frequency were below 1%, since these markers were less heterozygous and therefore less informative. As a result, we reduced the number of genetic markers from 19,465 to 627. To impute the missing data in the remaining part of the genotype data, we tried multiple imputation techniques, including mean, median, and most frequent [30], and found that the median approach led to the most accurate predictions. The yield and environment datasets were complete and did not have missing data.

3.2 Weather Prediction

Weather prediction is an inevitable part of crop yield prediction, because weather plays an important role in yield prediction but it is unknown a priori. In this section, we describe our approach for weather prediction and apply it to predict the 2016 weather variables using historical data from 2001 to 2015. There were two reasons for the aggregation of 2,247 locations: (1) the majority of the locations were in the middle west region, so it was reasonable to make the simplifying assumption that the prediction models were uniform across locations, (2) combining historical data for all locations allows sufficient data to train the 72 neural networks more accurately.

For each weather variable $w$, the neural network model explains the weather variable $X_{l,y}^w$ at location $l$ in year $y$ as a response of four previous years at the same location: $\{X_{l,y-1}^w, X_{l,y-2}^w, X_{l,y-3}^w, X_{l,y-4}^w\}$. We have tried other parameters for the periodic lag and found four years to yield the best results. As such, there were 24,717 samples of training data for each weather variable. The resulting parameters of the networks were then used to predict $X_{l,y=2016}^w$ using historical data of $X_{l,y=2012}^w$ to $X_{l,y=2015}^w$ for all $l$ and $w$. The structure of a shallow neural network is given in Figure 1.
The reason for using neural networks for weather prediction is that neural networks can capture the nonlinearities, which exist in the nature of weather data, and they learn these nonlinearities from data without requiring the nonlinear model to be specified before estimation [31]. Similar neural network approaches have also been used for other weather prediction studies [31, 32, 33, 34, 35, 36].

3.3 Yield Prediction Using Deep Neural Networks

We trained two deep neural networks, one for yield and the other for check yield, and then used the difference of their outputs as the prediction for yield difference. These models are illustrated in Figure 2. This model structure was found to be more effective than using one single neural network for yield difference, because the genotype and environment effects are more directly related to the yield and check yield than their difference.
The following hyperparameters were used in the training process. Each neural network has 21 hidden layers and 50 neurons in each layer. After trying deeper network structures, these dimensions were found to provide the best balance between prediction accuracy and limited overfitting. We initialized all weights with the Xavier initialization method [21]. We used SGD with a mini-batch size of 64. The Adam optimizer was used with a learning rate of 0.03%, which was decayed exponentially with a rate of 95% [37]. Batch normalization was used before activation for all hidden layers except the first hidden layer. Models were trained for 300,000 iterations. Residual shortcuts were used for every two stacked hidden layers [20]. To alleviate the vanishing gradient problem, we used the rectified linear unit (ReLU) activation function for all neurons in the networks except for the output layer, which did not have any activation function [18, 38, 39]. In order to avoid overfitting, we used the $L_2$ regularization [40] for all layers, and we also added dropouts [26] with a keep probability of 0.92 for the last four hidden layers. Figure 3 depicts the detailed structure of the deep neural network, which was the same for yield and check yield prediction.

Figure 3: Deep neural network structure for yield or check yield prediction. G, W, and S stand for genotype, weather, and soil data, respectively. Every two layers has one residual shortcut connection.

4 Results

The two deep neural networks were implemented in Python using the Tensorflow open-source software library [41]. The training process took approximately 1.5 hours on an i7-4790 CPU for each neural network. We also implemented three other popular prediction models for comparison: The least absolute shrinkage and selection operator (Lasso) [42], shallow neural network (having a single hidden layer with 300 neurons), and regression tree [43]. To ensure fair comparisons, two sets of these three models were built to predict yield and check yield separately, and the differences of their outputs were used as the prediction for the yield difference. All of these models were implemented in Python in the most efficient manner that we were capable of and tested under the same software and hardware environments to ensure fair comparisons.

The following hyperparameters were used for the regression tree. The maximum depth of the tree was set to 10 to avoid overfitting. We set the minimum number of samples required to split an internal node of tree to be 2. All features were used to train the regression tree.

We tried different values for the coefficient of $L_1$ term [40] in the Lasso model, and found that 0.5 led to the most accurate predictions.
Table 1 compares the performances of the four models on both training and validation datasets with respect to the RMSE and correlation coefficient. These results suggest that the deep neural networks outperformed the other three models to varying extents. The weak performance of Lasso was mainly due to its linear model structure, which ignored epistatic or $G \times E$ interactions and the apparent nonlinear effects of environmental variables. SNN outperformed Lasso on all the performance measures, since it was able to capture nonlinear effects. As a non-parametric model, RT demonstrated comparable performance with SNN with respect to yield and check yield but was much worse with respect to the yield difference. DNN outperformed all of the three benchmark models with respect to almost all measures; the only exception was that SNN had a better performance for the training dataset but worse for the validation dataset, which was a sign of overfitting. The DNN model was particularly effective in predicting yield and check yield, with RMSE for the validation dataset being approximately 11% of their respective average values. The accuracy for the check yield was a little higher than that for the yield because the former is the average yield across all hybrids and all years for the same location, which is easier to predict than the yield for individual hybrid at a specific location in a specific year. The model struggled to achieve the same prediction accuracy for yield difference as for the other two measures, although it was still significantly better than the other three benchmark models.

To evaluate the effects of weather prediction on the performance of the DNN model, we obtained prediction results using the predicted weather data rather than the ground truth weather data. As shown in Table 2, the prediction accuracy of DNN deteriorated compared to the corresponding results in Table 1 which suggested how sensitive yield prediction
is to weather prediction and the extent to which a perfect weather prediction model would improve the yield prediction results.

| Model  | Response | Training RMSE | Training Correlation | Validation RMSE | Validation Correlation |
|--------|----------|---------------|-----------------------|------------------|------------------------|
| DNN    | Yield    | 11.64         | 85.66                 | 13.94            | 78.65                  |
|        | Check yield | 9.49         | 78.35                 | 12.51            | 75.04                  |
|        | Yield difference | 12.80      | 37.64                  | 15.54            | 19.86                  |

Table 2: Prediction performance with predicted weather variables.

5 Conclusion

We presented a machine learning approach for crop yield prediction, which demonstrated superior performance in the 2018 Syngenta Crop Challenge using large datasets of corn hybrids. The approach used deep neural networks to make yield predictions (including yield, check yield, and yield difference) based on genotype and environment data. The carefully designed deep neural networks were able to learn nonlinear and complex relationships between genes, environmental conditions, as well as their interactions from historical data and make reasonably accurate predictions of yields for new hybrids planted in new locations with known weather conditions. Performance of the model was found to be relatively sensitive to the quality of weather prediction, which suggested the importance of weather prediction techniques.

A major limitation of the proposed model is its black box property, which is shared by many machine learning methods. Although the model captures $G \times E$ interactions, its complex model structure makes it hard to produce testable hypotheses that could potentially provide biological insights. Our future research is to overcome this limitation by looking for more advanced models that are not only more accurate but also more explainable.

References

[1] T Horie, M Yajima, and H Nakagawa. Yield forecasting. *Agricultural Systems*, 40(1-3):211–236, 1992.
[2] Syngenta. Syngenta Crop Challenge In Analytics. [https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php/](https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php/) 2018.
[3] I H DeLacy, K E Basford, M Cooper, J K Bull, and C G McLaren. Analysis of multi-environment trials, an historical perspective. *Plant Adaptation and Crop Improvement*, 39124, 1996.
[4] Nicolas Heslot, Deniz Akdemir, Mark E Sorrells, and Jean-Luc Jannink. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127(2):463–480, 2014.
[5] S C Chapman, M Cooper, G L Hammer, and D G Butler. Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. *Australian Journal of Agricultural Research*, 51(2):209–222, 2000.
[6] M Cooper and I H DeLacy. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88(5):561–572, 1994.
[7] Jose Crossa and Paul L Cornelius. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. *Crop Science*, 37(2):406–415, 1997.
[8] José Crossa, Paul L Cornelius, Ken Sayre, R Ortiz-Monasterio, J Iván, et al. A shifted multiplicative model fusion method for grouping environments without cultivar rank change. *Crop Science*, 35(1):54–62, 1995.

[9] Juan Burgueño, Jose Crossa, Paul L Cornelius, and Rong-Cai Yang. Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. *Crop Science*, 48(4):1291–1305, 2008.

[10] Juan Burgueño, José Crossa, José Miguel Cotes, Felix San Vicente, and Biswanath Das. Prediction assessment of linear mixed models for multienvironment trials. *Crop Science*, 51(3):944–954, 2011.

[11] José Crossa, Rong-Cai Yang, and Paul L Cornelius. Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(3):362–380, 2004.

[12] Osval A Montesinos-López, Abelardo Montesinos-López, José Crossa, José C Montesinos-López, David Mota-Sanchez, Fermín Estrada-González, Jussi Gillberg, Ravi Singh, Suchismita Mondal, and Philomin Juliana. Prediction of multiple trait and multiple environment genomic data using recommender systems. *G3: Genes, Genomes, Genetics*, 8(1):131–147, 2018.

[13] Jing Liu, C E Goering, and L Tian. A neural network for setting target corn yields. *Transactions of the ASAE*, 44(3):705, 2001.

[14] Scott T Drummond, Kenneth A Sudduth, Anupam Joshi, Stuart J Birrell, and Newell R Kitchen. Statistical and neural methods for site specific yield prediction. *Transactions of the ASAE*, 46(1):5, 2003.

[15] Oskar Marko, Sanja Brdar, Marko Panic, Predrag Lugonja, and Vladimir Crnojevic. Soybean varieties portfolio optimisation based on yield prediction. *Computers and Electronics in Agriculture*, 127:467–474, 2016.

[16] José R Romero, Pablo F Roncallo, Pavan C Akkiraju, Ignacio Ponzoni, Viviana C Echenique, and Jessica A Carballido. Using classification algorithms for predicting durum wheat yield in the province of buenos aires. *Computers and Electronics in Agriculture*, 96:173–179, 2013.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, volume 1. MIT Press Cambridge, 2016.

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[22] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[23] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–50. Springer, 1998.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Cvpr*, 2015.

[26] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[27] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, pages 4559–4566, 2017.

[28] Helena Russello. Convolutional neural networks for crop yield prediction using satellite images. 2018.

[29] Oskar Marko, Sanja Brdar, Marko Panić, Isidora Šašić, Danica Despotović, Milivoje Knežević, and Vladimir Crnojević. Portfolio optimization for seed selection in diverse weather scenarios. *PLOS One*, 12(9):e0184198, 2017.

[30] Paul D Allison. *Missing Data*, volume 136. Sage Publications, 2001.
[31] Kumar Abhishek, MP Singh, Saswata Ghosh, and Abhishek Anand. Weather forecasting model using artificial neural network. *Procedia Technology*, 4:311–318, 2012.

[32] Imran Maqsood, Muhammad Riaz Khan, and Ajith Abraham. An ensemble of neural networks for weather forecasting. *Neural Computing & Applications*, 13(2):112–122, 2004.

[33] Mohammed Bou-Rabee, Shaharin A Sulaiman, Magdy Saad Saleh, and Suhaila Marafi. Using artificial neural networks to estimate solar radiation in kuwait. *Renewable and Sustainable Energy Reviews*, 72:434–438, 2017.

[34] Amanpreet Kaur, JK Sharma, and Sunil Agrawal. Artificial neural networks in forecasting maximum and minimum relative humidity. *International Journal of Computer Science and Network Security*, 11(5):197–199, 2011.

[35] Rosmina Bustami, Nabil Bessaih, Charles Bong, and Suhaila Suhaili. Artificial neural network for precipitation and water level predictions of bedup river. *IAENG International Journal of Computer Science*, 34(2), 2007.

[36] S Santhosh Baboo and I Kadar Shereef. An efficient weather forecasting system using artificial neural network. *International journal of environmental science and development*, 1(4):321, 2010.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[38] Kevin Jarrett, Koray Kavukcuoglu, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153. IEEE, 2009.

[39] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[40] Andrew Y Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine learning*, page 78. ACM, 2004.

[41] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. TensorFlow: A system for large scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[42] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[43] Leo Breiman. *Classification and Regression Trees*. Routledge, 2017.