Wheat Yield Forecasting using Regression Algorithms and Neural Network

Cheng Dai
School of Information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
daichengzyw@gmail.com

Yinqin Huang
School of Information and Communication Engineering,
University of Electronic Science and Technology of China
Chengdu, China
hyinq@hotmail.com

Minghao Ni
School of Information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
845101809@qq.com

Xingang Liu*
School of Information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
hanksliu@uestc.edu.cn

***Abstract*—The demand for food is growing rapidly which calls for a more efficient method to grow crops. Precisely predicting the yield of corps can provide the farm operators advice to optimize their method of planting. However, existing methods have limitations in achieving high prediction accuracy and obviating overfitting. In this work, we consider the wheat yield forecasting problem of the U.S and compose a Bi-LSTM model with convolutional neural networks feature extraction sub-network to predict the wheat yield. The proposed model and other common models are tested base on the U.S wheat yield dataset which contains weather data and seasonal yield. The results of the experiment show that the model outperforms the other regression algorithms and can overcome the overfitting problem which is commonly found in other algorithms. The R^2 score of the model reaches 0.87 and the MSE is 19.6252.**

***Index Terms*—Prediction, Agricultural Yield, Random Forest, LSTM RNN**

I. Introduction

The food issue has always been a hot topic and directly affects many aspects of people's lives. With the explosive growth of the world population, countries all over the world are exploring methods to grow crops with less natural resource consumption while increasing production under the guarantee of environmental protection. In such situation, the precision and intelligence of agriculture are becoming increasingly important. Therefore, researchers have focused on the collection, combination and processing of external information including data of climate, satellites and geography in recent years. Such researches do improved the modern agricultural technology and also helped increase agricultural yield and crop quality.

Over the past few decades, regression algorithms and sensor technologies have developed rapidly with the advancement of computing hardwares and theories, that development directly leads to the emergence of more accurate algorithms of prediction. At the beginning, statistical-based models with probability theory or regression methods were widely used to implement yield forecasting [1], [2]. Since the explosive development of space technology and electronic industry, remote sensing and sensor network are employed to farmland monitoring. The data from remote sensing and sensor network have played an important role in the improvement of the prediction accuracy [3], [4]. Such models can not find the inner relations between features and yield, that drawback leads to a low prediction accuracy comparing to deep learning models.

Besides the aforementioned methods, a new technique which is recently gaining momentum thanks for the application of GPU in calculation is deep learning. Many deep learning models have been applied to this area, which allows more complicated data, such as genetic information, to be processed comparing to the statistical weather data [5], [6]. Another advantage of deep learning is the feature extraction, that data from higher levels of the hierarchy are formed by the composition of lower level features [7]. However, the existing deep learning model are suffering from overfitting problem which leads to the deterioration of performance on data that model haven't seen.

In this paper, we proposed a Bi-LSTM model with CNN feature extraction sub-network which can obviate the overfitting and maintain high accuracy which outperforms the other tested models base on the real-world wheat yield problem. To get usable data, we preprocessed the data by removing some entries that includes abnormal data and filling the missing items following the principle of time and location proximity.

The rest of this paper is organized as follows. Section.II introduces the common used methods in crop yield prediction. Section.III introduces the mathematic backgrounds of the LSTM and the bidirectional RNN, Section.IV introduces the architecture of the Bi-LSTM model and in Section.V experiment results are revealed and we discuss its feasibility and advantages. Finally, the conclusion and some directions of future work are given about our method in Section.VI.

II. Related works

The most common models which has been used for decades to forecast the crop yield are statistical-based models. Such methods feed agrometeorological data to a regression model, and finally are in the form of a function whose independent variables are agrometeorological parameters such as precipitation and temperature. The simplicity of regression models results in the applicability in crop yield forecasting. On the contrary, the limitations of data sources and observation boundaries make it hardly to extrapolate results to other locations. USDA (2011) [8] have been using such statistical-based regression model to forecast yield of corn in 10 major corn-producing countries for years. Tack (2015) [9] used a quadratic regression model base on the weather data and wheat yield of Kansas during 1985 to 2013 to research on how the temperature effects the wheat yield during the growing season in the context of global warming, the result shows that the largest factor of yield loss are freezing temperatures in the Fall and extreme heat events in the Spring, they also found that the overall rising temperature imposes negative influence on wheat yield even after accounting for the benefits of reduced exposure to freezing temperatures.

Recently, given the increasing amount of climate parameters collected by sensors with evolving performance, data which is more accurate to explain the crop condition like biomass and growth stage have been added into the models. Giri (2017) [10] developed a district-wise regression model to predict the wheat and rice yield in seven districts of eastern Madhya Pradesh India using combined climatic data and agrometeorological data. Their experiment result shows that the deviation between reported and forecasted yield was less than 15 percent. Singh (2017) [11] went a step further, they calculated two derivative variables for each weather variable, one as simple accumulation of weather variable and the other one as weighted accumulation of weekly data on weather variable. Then they train the model separately in different districts, and the R^2 score varies from 0.32 to 0.88 in different districts. Schauberger(2017) [12] concentrate on detecting weather influence on yield anomalies and project yields with unknown weather, they build an ordinary least squares regression scheme based on the Cobb-Douglas production function with different model specifications and their model can explain 63%-81% of observed yield anomalies of the US in the experiment.

Except for naive statistical-based regression algorithms, advanced machine learning techniques also have been used in this area. Sisodia et al. (2018) [13] build a multiple linear regression model with discriminant function analysis to predict the rice yield base on biometric characters. The experiment result has shown that the percent standard errors of prediction with real records is below 5%. Mokarram and Bjianzadeh (2016) [14] compare and analyze the multiple linear regression and fully connected neural network with one hidden layer of predicting yield of barely, their experiments have proved that neural network models outperform the naive regression models with the same data inputs.

With GPU getting involved to boost the computing neural network training procedure, an increasing number of researchers have focused on deep learning algorithms which allow to input more complicated data to get more accurate forecasting results. Basso(2013) [15] uses deep neural networks to predict the maize yield with a dataset containing 2267 hybrids planted in 2247 locations between 2008 and 2016. Their DNN model has a superior prediction accuracy, with 12% root-mean-square-error (RMSE) and 50% standard deviation of the average yield on validation datasets. You(2017) [3] proposed a dimensionality reduction technique based on CNN or LSTM, which can automatically learn features with scarce labeled data. This algorithm incorporates a Gaussian Process component to explicitly model the spatial-temporal structure of the data and improves accuracy on the prediction of soybean yield. The experimental results verify that this model outperforms state-of-the-art algorithms. Elavarasan and Vincent (2020) [16] propose a deep recurrent Q-network model with the construction of a recurrent neural network over the Q-Learning reinforcement learning algorithm, and the prediction accuracy on the original data distribution is 93.7%.

III. Mathematic backgrounds

Since The Bi-LSTM structure is employed to predict the yield, some mathematic backgrounds about LSTM and Bidirectional wrapper will be introduced in this part.

A. LSTM Node

Recurrent Neural Network is a kind of neural network that is good at processing time-series data. It can well process tasks with continuous-time and contextual relationships which is exactly in line with the characteristics of the crop growth. The connections existing between each layer make the network better in learning time series relations, but on the other hand, this leads to a gradient explosion in backpropagation [17].

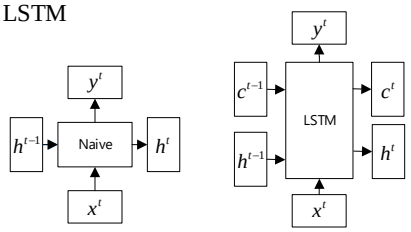


Fig. 1. LSTM and RNN structure

Fig. 1 shows the structure of LSTM and RNN. The structure of t-th RNN node on the first image, where x^t is the input of t-th node, h^{t-1} is the input of t-th node received from $t-1$th node, y^t is the output and h^t is the output which will transfer to next node. The structure of t-th LSTM node is shown on the second image, that structure have two more

states c^t(cell state) and h^t(hidden state), state c in LSTM is the counterpart of state h in RNN, and normally state c changes slowly in front propagation and add a number to c^{t-1} while h^t changes drastically.

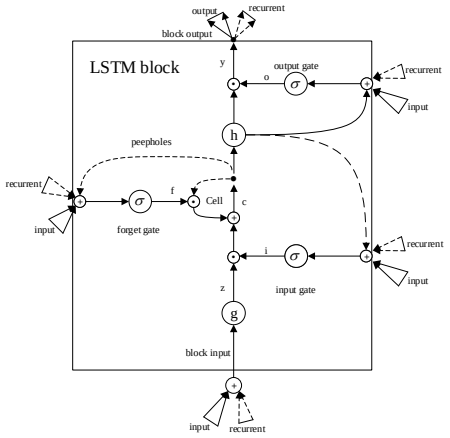


Fig. 2. LSTM node

$$\begin{cases} f_t = \sigma\left(w_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t = \sigma\left(w_i \cdot [h_{t-1}, x_t] + b_i\right) \\ o_t = \sigma\left(w_o \cdot [h_{t-1}, x_t] + b_i\right) \end{cases} \tag{1}$$

$$\sigma\left(x\right) = \frac{1}{1 + e^{-x}} \tag{2}$$

For a LSTM node shown in Fig. 2, four states f_t(forget gate), i_t(input gate), c_t(current state) and o_t(output gate) are generated by concating x^t and h^{t-1} via (1). Forget gate f_t controls the remaining of information transferred from c_{t-1}, and $[h_{t-1}, x_t]$ means to concat vector h_{t-1} and vector x_t. w_f and b_f are the weight and bias of forget gate f_t, σ is the Sigmoid activation function whose expression is (2).

To calculate c_t, the input state $\widetilde{c_t}$ are needed. Use activation function $\tanh$ and (3), (4) and (1) to calculate c_t, the $\odot$ in (4) means corresponding element multiplication.

$$\tilde{c}_t = \tanh\left(W_c \cdot [h_{t-q}, x_t] + b_c\right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t \tag{4}$$

All these features of LSTM meet the crop growing process, there are weather fluctuations or natural disasters in a short time, while there are stable natural conditions over a long period.

B. Bidirectional RNN

Mike Schuster [18] proposed the Bi-directional RNN(BRNN) in 1997 as shown in Fig. 3. In his paper, the conventional RNN neurons are splited into two parts, one for negative time direction and the other for positive time direction. The output of the forward states will not be send to the backward states.

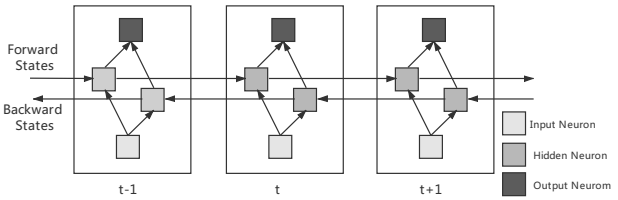


Fig. 3. The structure of Bi-RNN, the forward states are not connected to the backward state

IV. Model Architecture

The architecture of the model is shown in Fig. 4, the whole model consists of a 2D CNN sub-network and a Bi-LSTM sub-network which separately extracts high level features and predicts the yield. The input of the network is the preprocessed data which has the same planting duration of 186 days, the output of the CNN network is fed to the Bi-LSTM network which consists of 3 stacked Bi-LSTM layers followed by a fully-connect layer to give the prediction.

The CNN network is employed for feature extraction, there are two same Conv2D structure stacked in the CNN sub-network which contains a Conv2D layer, Relu activation layer, BatchNormalization layer and Maxpooling layer. The kernel size of the Conv2D layers is 1×2, and the first one has 32 filters while the second one has 64 counterparts. The Conv2D layers are in the timedistributed wrapper which makes the Conv2D layer do the same calculation on every time pieces. Then, Conv2D layers are first followed by a activation layer with the activation function ReLU (Rectified Linear Units) (5) and then followed by a batch normalization layer and a max-pooling layer whose kernel size is 1×2. And at the end of the CNN sub-net, the output is reshaped to 184×64 by a reshape layer to fit the input shape request of Bi-LSTM sub-network. Batch normalization regularize the CNN to mitigate the overfitting and the pooling layers improve network's generalization, robustness to small distortions and reduces dimensionality.

$$ReLU = \max(0, x) \tag{5}$$

There are three stacked Bi-LSTM layers followed by two dense layer to predict the yield in Bi-LSTM sub-network. By the Bidirectional wrapper, the three LSTM layers can utilize the information from both positive time direction and negative time direction. ALl of three layers have 64 filters and both of the kernel regularizer and the recurrent regularizer are set to be the L2 with the property of 0.0001. The weight initializer we choose the Glorot normal initializer, also called Xavier normal initializer, proposed by Glorot in 2010 [19]. It initialize the weights by drawing samples from a uniform distribution within $[-B, B]$, where B can be calculated by (6), in which the F_i represents the number of input units in the weight tensor and F_o represents the number of output units. The three stacked Bi-LSTM structure are followed by two dense layer, the first one has 32 units and the second one

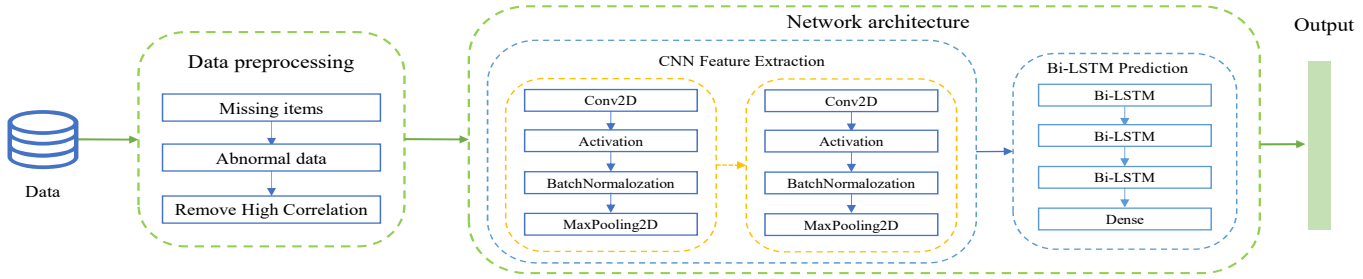


Fig. 4. The architecture of the model are splited into two sub-networks, one is the CNN sub-network for feature extraction and the other is Bi-LSTM sub-network for yield prediction.

has 1 units, both employ the l2 kernel regularization and are initialized by the Glorot normal initializer.

$$B = \sqrt{\frac{6}{F_i + F_o}} \tag{6}$$

In the procedure of training, the accuracy on the training dataset might improve, but the performance begins to reduce at a certain point on the data not yet seen by the model. To make sure the real-world performance of the model, earlystopping callback function was employed, such function was set to monitor the cross validation MSE. The train will end when that value has stopped improving after 10 epochs consecutively. The total epoch with batch size of 16 was set to be 1000 under the regulation of earlystopping callback, and the gradient descent on top of the adaptive momentum (ADAM) optimizer was choose.

V. Experiment

A. Data Preprocessing

The dataset includes data of winter wheat yield in the U.S with local climate conditions in 2013 and 2014, and contains 150 counties from 5 states and has 26 features, 360042 entries in total.

1) Missing items: There are 654 missing items in the dataset, accounting for 1.81% of the total data volume. Deleting an entry with missing items may result in the decline of continuity of data in a certain location due to the characteristics of crop growth that disaster of someday will cause a great reduction of output. Filling the missing items with the nearest and most recent entry due to the continuity of climate data in both space and time. For a isolated miss item, it is filled by its spatial and temporal adjacent items, and for continuous time missing items, they will be filled from both ends to the middle one.

2) Abnormal data: By going through the data, an apparent anomaly was found in the yield which is shown in Fig. 5(a). There are entries having nonzero yield with a relatively short period of time from being seeded to harvest which was recorded by column *DayInSeason*. To ensure a uniformed input shape and drop such abnormal data, we dropped data from locations who have recorded for less than 186 days.

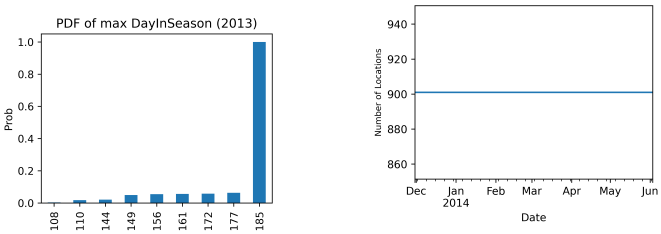


Fig. 5. (a) shows the location

After doing so, there are 901 locations left in the dataset which all have the same planting duration of 186 days.

3) Map visualization: For data with location, map visualization can always help us find some intuitive relations between location and target. According to Fig. 6-(b), there are two intuitive conclusions:

- The yield in the northern growing areas was significantly higher than that in the south.
- In the south growing areas, the more east the locations are the high the yields are.

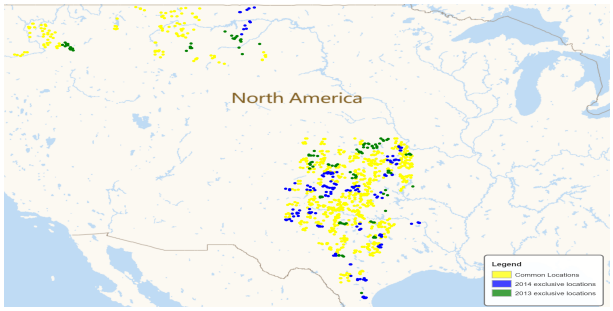
4) Cut high correlation features: There is significant multicollinearity between some features in the dataset, that multicollinearity can lead to errors and distortions in results [20]. By calculating the correlation coefficient between the features, some features with abnormal high correlation which reaches 99%. Table. I shows the pairs of features whose correlation high than 90%. Before the data is feed to the model, these columns with high correlation accompany with some yield irrelevant columns including *State*, *Location* e.t. are cut out from the original dataset.

B. Algorithm Evaluation

Before testing, use the *Min-Max* method (7) to standardize the data, then the data was fed to the model.

$$x_i := \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{7}$$

To evaluate the performance of the algorithms, a certain part (20%) of data was randomly set apart as validation set. Such setting can make the algorithms have the same baseline



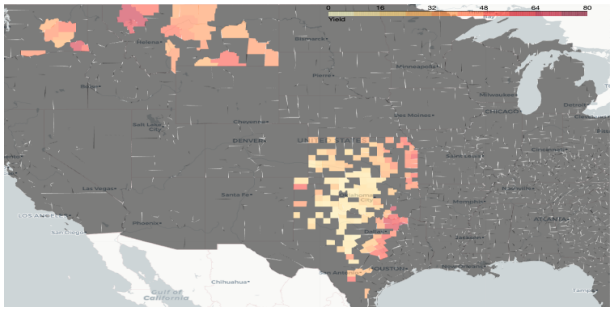


Fig. 6. (a)shows the location of all data being collected, yellow points represent locations monitored for two years which accounts for most of the locations (b)shows the yield varying by locations

TABLE I
FEATURES CORRELATION HIGH THAN 90%

Feature 1	Feature 2	cor1	cor2	cor3
apparentTempMin	TempMin	0.99	-0.04	-0.06
apparentTempMax	TempMax	0.99	-0.14	-0.15
precipIntensity	precipIntensityMax	0.91	0.04	0.03
dewPoint	tempMin	0.91	0.01	-0.06
apparentTempMin	dewPoint	0.90	-0.04	0.01

[a]cor1: correlation between feature1 and feature2
[b]cor2: correlation between feature1 and yield
[c]cor3: correlation between feature2 and yield

TABLE II
PERFORMANCE OF ALGORITHMS

Algorithm	MSE	R^2
Random Forest [22]	47.85	0.71
KNN [23]	60.65	0.43
Polynomial regression	108.48	0.24
Linear regression	179.43	0.13
SVR [24]	188.46	0.11
SVR RBF kernel [25]	193.51	0.07
SVR linear kernel	197.06	0.06
Naive LSTM	45.38	0.73
CNN-Bi-LSTM(Our method)	12.37	0.84

to compare to each other. We use both mean-square-error (MSE) and the coefficient of determination denoted as R^2 to estimate the performance of the algorithms. The R^2 score can be calculated by (8)-(10).

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{8}$$

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \tag{9}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{10}$$

C. Result

Besides the Bi-LSTM model, we also tested commonly used ML algorithms and naive LSTM models, each algorithms are fed the same data except SVR whose computing complexity is too high to fed all the data [21]. But during the test procedure, the randomly forest shows a severe overfitting, extra efforts are done to solve such problem in Section.V-D. The results of the algorithms is shown in Table. II.

D. Overfitting in Random Forest

The Random Forest have shown a severe overfitting in the test whose MSE reaches 0.0034. in this part, we try to find out what leads to this problem and solve it. Going through the contribution of the features in the prediction, we found that the location features (latitude and longitude) contribute more than 99% in total, which means that the algorithm can give a exact "prediction" just base on where the farm is.

We have tried three progressive methods to solve the problem. First, simply remove the longitude and latitude columns from the dataset and run the test again. Learning curve of the result shown in Fig. 7(a), the problem was not solved but tell us that it is not the location but some other factors cause the overfitting.

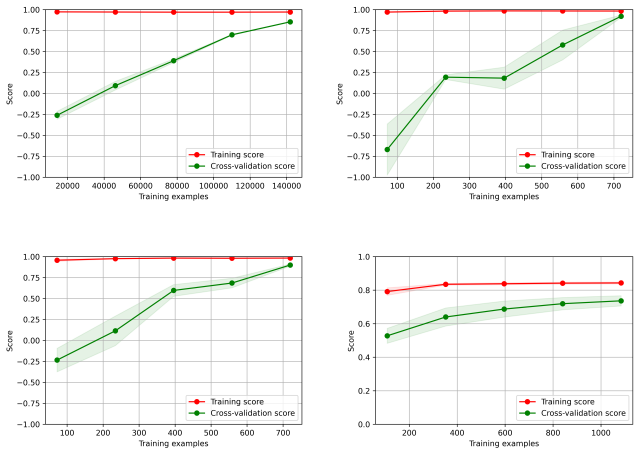


Fig. 7. Learning Curve of Random Forest. (a) generated by the origin data, obviously, there's severe overfitting. (b) generated by setting data of different days to a new feature, performance gets worse. (c) generated by compress the data, the CV-score gets better. (d) generated by feature selection, the overfitting problem being solved

Go through the dataset, finding that the yield is bound

to everyday log, which means that there are 186 different weather conditions with the same yield for every location. So we reshape the data to shrink the disparity in volume between yield and data entries. For data in 2013, the original shape is $(166685, 16)$ (186 days $\times$ 901 locations, 16 features). Transform the shape of data into (901×2960) (901 locations, 185 days $\times$ 16 features) and run the test again. Learning curve of the result is shown in Fig. 7(b). These two methods can't solve the overfitting problem.

Try to aggregate the data to shrink the number of features. Calculate the mean values of ever 60 days and shrink the number of features to 48 and run the test again. Learning curve of the results is shown in Fig. 7(c). There is some improvement after half of the dataset getting involved.

To extract some indirect features from the origin features, we calculate the minimum and maximum value of features in 30 days, minimum maximum NDVI in 30 days, mean temperature difference, and variance e.t using a rolling window method from pandas. Then, run the test base on these features get the learning curve Fig. 7(d). The gap between training score and cross-validation score are narrowed, and improved the performance when

VI. Conclusion and Future work

We composed a Bi-LSTM model with a CNN feature extraction sub-network to predict the wheat yield in the U.S and compared it with some commonly used regression algorithms and naive LSTM. The results show that the model outperforms the other algorithms and is less likely to get overfitting which is commonly found in other regression algorithms. Compare to the naive Bi-LSTM, the CNN feature extraction sub-network can help the Bi-LSTM fit the data better and reduce the disparity of performance between that on training set and test set.

However, there are still some factors we do not consider due to the limitation of data. Some features crucial to crop growth, such as how farmers plant the crop and the condition of the sow in a certain location e.t, are not involved in the models due to the lack of data sources. The weather, undoubtedly, greatly influence the yield of corps, but some subjective behaviors of human can also influence the yield. The corps is not planted in an environment free of human intervention, water the crops when precipitation is a basic natural behavior that people influence the planting. If can get access such sort of data, the accuracy of prediction will be higher.

References

[1] J. H. Matis, T. Saito, W. E. Grant, W. C. Iwig, and J. T. Ritchie, "A Markov chain approach to crop yield forecasting," 1985.

[2] D. J. Stephens, "Crop yield forecasting over large areas in Australia," Ph.D. dissertation, Murdoch University, 1995.

[3] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data," in *AAAI*, 2017, p. 7.

[4] B. Baruth, B. Baruth, A. Royer, A. Klisch, and G. Genovese, "The use of remote sensing within the MARS crop yield monitoring system of the European Commission Monitoring and forecasting agricultural resources," Tech. Rep.

[5] A. Mateo-Sanchis, M. Piles, J. Muñoz-Marí, J. E. Adsuara, A. Pérez-Suay, and G. Camps-Valls, "Synergistic integration of optical and microwave satellite data for crop yield estimation," *Remote Sensing of Environment*, vol. 234, p. 111460, Dec. 2019.

[6] K. A. Steen, P. Christiansen, H. Karstoft, and R. N. Jørgensen, "Using Deep Learning to Challenge Safety Standard for Highly Autonomous Machines in Agriculture," *Journal of Imaging*, vol. 2, no. 1, p. 6, Mar. 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[8] D. L. Good and S. H. Irwin, "USDA corn and soybean acreage estimates and yield forecasts: Dispelling myths and misunderstandings," Tech. Rep., 2011.

[9] J. Tack, A. Barkley, and L. L. Nalley, "Effect of warming temperatures on US wheat yields," *PNAS*, vol. 112, no. 22, pp. 6931–6936, Jun. 2015.

[10] A. Giri, M. Bhan, and K. Agrawal, "Districtwise wheat and rice yield predictions using meteorological variables in eastern Madhya Pradesh," *Journal of Agrometeorology*, vol. 19, no. 4, pp. 366–368, 2017.

[11] M. Singh and S. Sharma, "Forecasting the maize yield in Himachal Pradesh using climatic variables," *Journal of Agrometeorology*, vol. 19, no. 2, pp. 167–169, 2017.

[12] B. Schauberger, C. Gornott, and F. Wechsung, "Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting," *Global Change Biology*, vol. 23, no. 11, pp. 4750–4764, 2017.

[13] B. Sisodia and S. Kumar, "Pre-harvest forecast model for rice yield based on biometrical characters: An application of discriminant function analysis," *International Journal of Chemical Studies*, p. 4, 2018.

[14] M. Mokarram and E. Bijanzadeh, "Prediction of biological and grain yield of barley using multiple regression and artificial neural network models:," *Australian Journal of Crop Science*, vol. 10, pp. 895–903, Jun. 2016.

[15] B. Basso, D. Cammarano, and E. Carfagna, "Review of Crop Yield Forecasting Methods and Early Warning Systems," 2013.

[16] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," *IEEE Access*, vol. 8, pp. 86 886–86 901, 2020.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," 1997.

[18] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov./1997.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, Mar. 2010, pp. 249–256.

[20] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: The problem revisited," 1967.

[21] S. M. Clarke, J. H. Griebsch, and T. W. Simpson, "Analysis of support vector regression for approximation of complex engineering analyses," 2005.

[22] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomForest," 2002.

[23] L. E. Peterson, "K-nearest neighbor," 2009.

[24] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," 1997.

[25] N. Cristianini, J. Shawe-Taylor *et al.*, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge university press, 2000.