Ensemble Learning for Improving Generalization in Aeroponics Yield Prediction

Julio Torres-Tello*†, Suganthi Venkatachalam‡, Lyman Moreno‡ and Seok-Bum Ko*
*Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, Canada
†Departamento de Electrica, Electronica y Telecomunicaciones, Universidad de las Fuerzas Armadas ESPE, Sangolqui, Ecuador
‡Farm Boys Design Corp., Saskatoon, Canada
Email: {julio.torrestello, sugi.venkatachalam, seokbum.ko}@usask.ca, l.moreno@farmboysdesign.com

***Abstract*—Agriculture plays a crucial role in economy of several countries and yield prediction is essential for production management and operation planning. Machine Learning (ML) is a growing trend in determining yield as a complex function of multiple input variables. Aeroponics is one of the efficient sustainable farming methods and allows all season farming despite hostile outdoors growing environment. In this paper, yield prediction in aeroponics is studied using ML. We have compared and analyzed three popular supervised ML methods - Dense Neural Network (DNN), Random Forest based on decision trees (RF) and Support Vector Regression (SVR). Air quality and water quality measurements including temperature, humidity, CO_2, pH and Total Dissolved Solids (TDS) are used for yield prediction. Other static inputs such as number of days before and after transplant are also used. Six crops are studied (garlic chives, basil, red chard, rainbow chard, arugula, and mint). DNN performs particularly well with the prediction. The root mean square error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) are calculated to estimate the efficiency of the method. Mean square error and R^2 score of DNN are 0.10 and 0.67, RF follows DNN correctness with MSE and R^2 of 0.12 and 0.62, and SVR achieves 0.18 and 0.45 respectively, all of these values over the validation dataset. In addition to individual models, the two top performing models are combined as an ensemble model to improve overall performance, which shows an average R^2 score over the whole dataset divided by crop of 0.81.**

I. Introduction

There is an increasing trend in demand for healthy fresh foods. On a global scale, there is a need to meet food demands and predicting yield with better accuracy is important to plan import and export in case of deficit or excess [1]. Methods like green house, hydroponics, and aeroponics allow for year around harvest, protection from harsh cold or warm weather, portability, cultivation of diverse crops and disease free cultivation. Among these alternatives, aeroponics has been evolving as a promising and efficient modern day plant growing method in multiple countries [2]. Studies in [3] show that, compared to traditional farming, aeroponic farming shows an increase of yield ranging from 7% to 65% based on the type of crop. Along with faster crop cycle, they also have reduced water, pesticide and fertilizer usage [4].

Aeroponics systems are soil-less growing methods where plants get their nutrition from a mix of water and nutrient tonics. Aeroponics are more commonly closed or semi-closed systems, with automated controlled variables that influence the growth of the plants. Unlike traditional field agriculture, air quality variables such as temperature, humidity, CO_2 and light can be maintained within specific ranges by automated systems. Water quality variables including pH and TDS can be controlled as well, and sprayed to roots to deliver nutrients. They also allow for more dense growing environment since plants can be grown in a stacked tower structure.

Yield prediction is a complex task based on various parameters including crop type, environment and water quality. Human based yield prediction is time consuming and prone to error. The outcome of the harvest is hard to predict in advance, but yield knowledge can help growers model and plan price, supplies, and future techniques. Recently, ML is becoming an important tool for predictions in medicine, robotics, economic sciences, climatology and yield prediction is no exception.

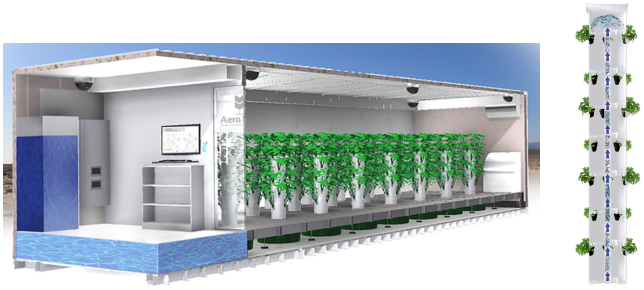
Furthermore, yield prediction could be the basis of a fully automated control system in which yield could be maximized by setting the variables under control (light, nutrients, etc.) in the best possible way. Such a system would increase revenues by using the minimum possible resources to produce the maximum yields. This is extremely valuable in aeroponics, where the farmer has much more control over the environmental conditions when compared to traditional farming.

In traditional agriculture, yield predicted from remote sensing data is extensively used. Normalized Difference Vegetation Index (NDVI) is used as an main indicator of yield projection [5], [6]. However, prediction of crop yield with better accuracy demands other inputs like water nutrients, fertilizer and pesticides information. In controlled aeroponics farming, where sensors and controllers are an integral part, this information is used for yield prediction and typically without the need for remote sensing.

In our paper, data is collected from a sustainable and organic aeroponic farming device called AeroPod of Farm Boys Design, a corporation based in Saskatchewan, Canada. A typical Aeropod system and a single stacked tower in AeroPod are given in Fig. 1.

Three ML techniques - Deep Neural Network (DNN), Random Forest (RF) and Support Vector Regression (SVR) are studied and compared. DNN is a neural network structure with multiple layers which has dramatic breakthroughs in

978-1-7281-3320-1/20/$31.00 ©2020 IEEE



(a) (b)

Fig. 1. Aeroponics container (a) and one of the towers in AeroPod (b).

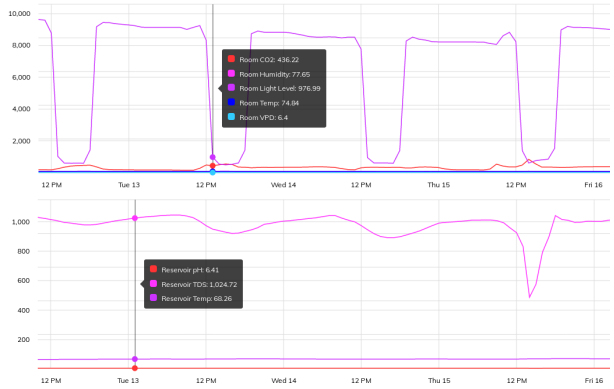


Fig. 2. Screenshot of the data acquisition platform.

multimedia recognition, object detection and classification [7], [8]; RF, proposed in [9] is used to improve prediction accuracy by taking in account large number of decision tree models. SVR [10] is an effective regression method with an advantage of one global optimum compared to neural network approach. Of the three methods, DNN is found to have better yield prediction, followed by RF. An ensemble regressor that puts together these two models is analyzed as well.

The paper is organized as follows: Section II discusses what data is used as inputs to the models and three ML models used in our work are explained in detail. Section III discusses mean square, mean absolute error and R2 score of all our models. Section IV concludes the paper.

II. Methodology

A. Growing Methods and Data Collection Materials

The crops taken for studying yield prediction are garlic chives, basil, red chard, rainbow chard, arugula and mint. Seeds are sown in rockwool grow cubes medium in a tray, until they germinate and attain early stages of growth. Then, they are transplanted to stacked towers in AeroPod. Rockwool absorbs nutrients and retains oxygen for fast growth. Each layer in a tower consists of few spots which can accommodate a rockwool cube. In the AeroPod, nutrient mixture comes in contact with rockwool cubes at constant intervals, thereby nourishing the roots. When the plants reach the expected harvest growth, they are harvested and yield is measured as weight per spot.

Air quality and water quality sensors are implemented in AeroPod container (an example of the data collection system is shown in Fig. 2), whose measurements are valuable inputs to ML algorithms. Environment and nutrient input variables include average (calculated for the time frame in which each plant has been in the AeroPod) of hourly values of room Carbon-di-oxide levels, room relative humidity, room light level, room temperature, room vapour pressure deficit, water pH, water TDS and reservoir temperature. Number of days in tray, number of days in tower, harvest number (how many times the plant has been harvested since first transplanted to AeroPod) and grow number (how many times a plant has been transplanted into that spot) are given as inputs to the model. The label of the ML model is yield measure of weight per spot (oz/spot). A block diagram of our model is given in Fig. 3. For this study, 200 samples have been collected between November 2018 and August 2019.

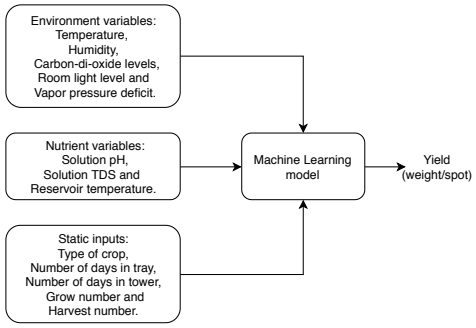


Fig. 3. Block diagram of the machine learning approach used in our analysis.

B. ML models and training parameters

The training process of all the algorithms has involved k-fold cross validation (k=10). This resampling technique without replacement is applied in order to obtain an error rate as independent as possible of the train-validation split of the dataset.

The only pre-processing technique applied to this dataset is the standardization of the features by subtracting the mean and scaling to unit variance, according to (1):

$$y = \frac{x - \mu}{\sigma}, \tag{1}$$

where x is the original feature, y the resulting one, μ is the mean of the samples, and σ is their standard deviation.

C. Support Vector Regression model description

SVR is used in yield prediction to provide an estimate of output as a non-linear function of inputs. The kernel function plays a crucial role in transforming inputs into higher dimensional space. Here, we use a non-linear kernel called Radial Basis Function (RBF). In our model, the influence of single training example gamma is taken as 0.1, factor C is 100 which decides the increase or decrease of margin, and a margin of tolerance epsilon = 0.1.

D. Random Forest model description

RF is an effective classification and regression method. It involves ensemble learning of multiple decision trees. We implemented a RF algorithm and tuned the hyperparameters in order to obtain the best possible results. The results of the tuning process indicated that the ideal number of trees for the ensemble is 100 and their maximum depth must be 20. One important difference between RF and the previous ML model is that RF does not require data normalization.

E. Deep Neural Network model description

The DNN model used in our work in given in Fig. 4. The model uses rectified linear unit (ReLU) as activation function for all layers, including the output. Rectified linear function can be given as:

$$R(z) = max(0, z), \tag{2}$$

which is basically a linear function that cancels out any negative value. Usually, regression problems use linear activation for the output neuron; however, in our case we had the issue that in some cases the DNN would predict negative yields (mathematically possible due to the linear response of the function, but impossible from the physical interpretation of the yield) so we decided to make those predictions zero (i.e. to use a ReLU activation function). This modification improved the results of the predictions of the DNN, and not only for the previously values predicted as negative, but for all of them.

Our DNN has three hidden layers with 48, 48 and 24 neurons respectively. After trying different combinations of number and size of layers and regularization techniques (we tested dropout and L2 with different coefficients), the architecture shown in Fig. 4 was our optimal solution, without the need of using any other method to fight over-fitting than the reduced size of the network itself, which in the end has 4,465 weights. Some of the important training parameters are the optimizer, loss function and metrics (adam, MSE and MAE, respectively), and that the model was trained for between 101 and 327 epochs per fold (early stopping was implemented), using a batch size of 4. The average of training and validation losses of this algorithm are shown in Fig. 5.

F. Ensemble of DNN and RF model description

A common technique used in the field of ML in order to increase the generalization performance of models is to combine them into a meta-model, in which the error probability of an ensemble is always better than the error of any of its components, provided that their individual error is better than a random guess [11]. As we will see in the following section, the best performing algorithms are RF and DNN, and they are used to create the ensemble learner finally used in this work.

III. Results and Discussion

Our ML models were implemented in Python 3, using Scikit Learn for SVR and RF, and Keras with Tensorflow backend for the DNN. All of them were trained on an machine with

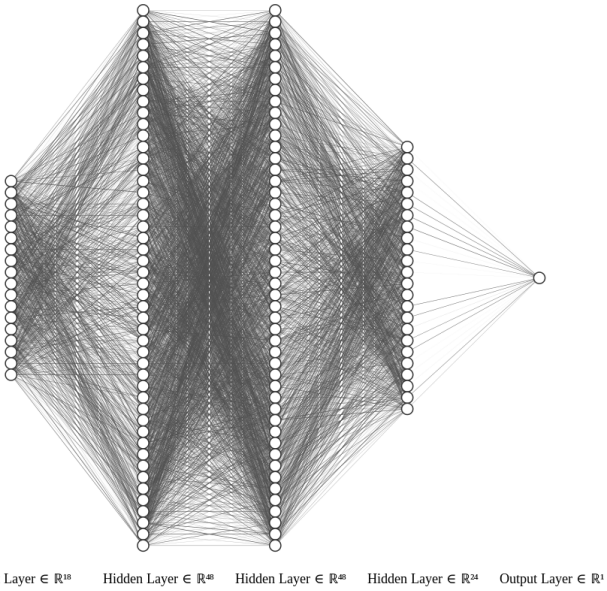


Fig. 4. Architecture of the DNN implemented for this work.

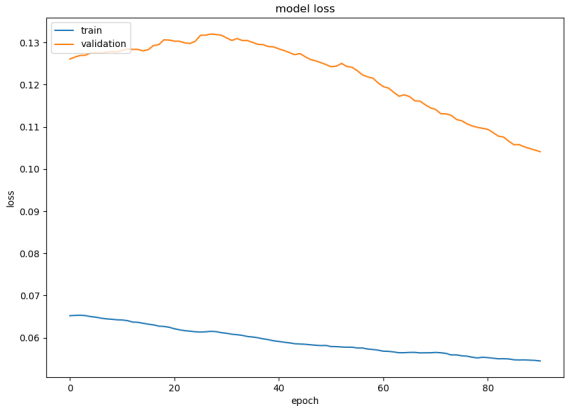


Fig. 5. Training and validation losses (10-fold average) of the DNN model.

12 Intel(R) Core(TM) i7-8750H CPU's at 2.20GHz, 16 GB of RAM, and a Nvidia GTX 1060 GPU used only for the DNN.

The main results of the regression process over the above described dataset are summarized in Table I, where we can see that the best performing model is DNN, when measuring both the error (either MAE or MSE) and the coefficient of determination. However, RF shows similar although not as good results as DNN. Fig. 6 is a plot of the predictions generated by the (a) DNN and (b) RF models for the training and validation sets, after the 10-fold training process.

Both, Table I and Fig. 6 show that these models have a good performance when dealing with the training set, but their error increase when dealing with the validation set (especially SVR shows a poor R^2 for validation), which means that they do not have a good generalization power. In our implementations we have tried to tackle overfitting in many ways, but we believe that in our case the only remaining options are the implementation of a meta-model by means of an ensemble of

TABLE I
MEAN SQUARED ERROR (MSE), MEAN AVERAGE ERROR (MAE) AND COEFFICIENT OF DETERMINATION (R^2), FOR THE TRAINING AND VALIDATION SETS.

	MSE		MAE		R^2	
	train	**val**	**train**	**val**	**train**	**val**
SVR	0.08	0.18	0.15	0.26	0.86	0.45
RF	0.05	0.12	0.13	0.21	0.90	0.62
DNN	0.05	0.10	0.11	0.18	0.91	0.67

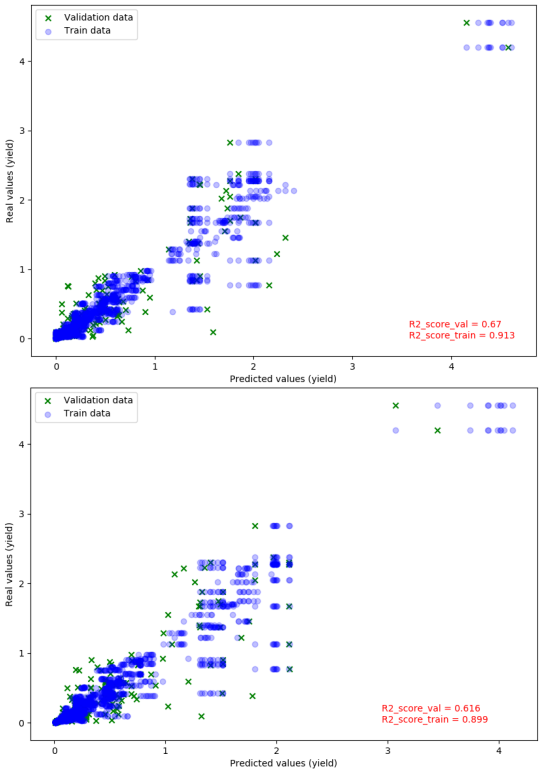


Fig. 6. Predictions over the training and validation sets with (a) the DNN model and (b) the RF model.

regressors, and as the main future work, to increase the size of our dataset by collecting new samples.

Additionally, it was important to evaluate how the models perform on the different crops that conform our dataset. Table II presents the coefficient of determination calculated with the two best performing models (RF and DNN) over each of the six different crops. It is interesting to note that the predictions are acceptable for most of them, except for the rainbow chard which will need further analysis of the phenotype of this plant.

Finally, we built an ensemble of these two models in order to improve the prediction and generalization capabilities of our ML models. This ensemble gave the best result with weights of 48% to DNN and 52% to RF, and obtained the results presented in Table II, which are better that the individual-model predictions, and it is quite obvious that this increases the coefficient of determination especially for the case of garlic chives. As a matter of example, Fig. 7 shows the results of the ensemble model for the best (Garlic Chives) and worst

TABLE II
COEFFICIENT OF DETERMINATION (R^2) CALCULATED OVER THE COMPLETE DATASET DIVIDED BY TYPE OF CROP, FOR THE TWO BEST PERFORMING ML ALGORITHMS AND THE ENSEMBLE OF THE TWO.

	RF	**DNN**	**Ensemble**
Garlic Chives	0.80	0.71	0.89
Basil	0.85	0.89	0.88
Red Chard	0.83	0.86	0.87
Rainbow Chard	0.70	0.61	0.66
Arugula	0.74	0.73	0.74
Mint	0.80	0.83	0.82
Average	0.79	0.77	0.81

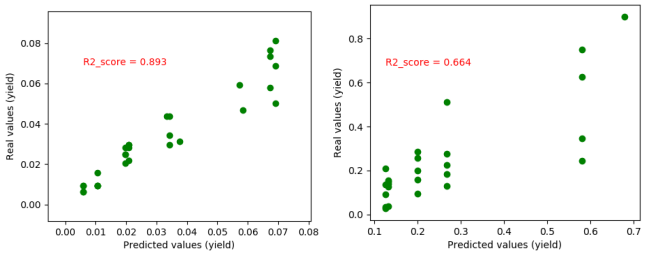


Fig. 7. Predictions with the ensemble of models, over the best -Garlic Chives- (a) and worst -Rainbow Chard- (b) performing crops.

(Rainbow Chard) predicted crops.

These results show that with an adequate tuning of weights in the ensemble model, we could tackle future generalization issues when we increase our dataset, as a future work.

IV. CONCLUSION

This work presents a yield prediction model for aeroponic crops in a controlled environment, based on the environmental variables that can be controlled and/or measured in the production system. For this purpose we have used 200 samples covering 6 different crops, and we have reached an average coefficient of determination value $R^2 = 0.81$ when testing an ensemble of the two best models (DNN and RF) over the whole dataset separated by type of crop. This and the other results presented in this paper show the potential for implementing a yield prediction model that could become the first step towards the full automation of a crop production system based on aeroponics, such as the one shown here.

This work draws the main path to be followed in order to have a robust yield prediction (and automated production) tool that would eventually require less human intervention with a bigger profit margin. The next step towards that goal is the collection of more data, organized in such a way that it does not present the issues that have had to be addressed for this work. We believe that it would allow us to generate more accurate and self-explanatory results, that could be easily implemented in a final control system.

ACKNOWLEDGMENT

The National Research Council of Canada Industrial Research Assistance Program (NRC IRAP) provided financial support for the project. J.T.T receives a scholarship from SENESCYT.

REFERENCES

[1] M. K. van Ittersum, K. G. Cassman, P. Grassini, J. Wolf, P. Tittonell, and Z. Hochman, "Yield gap analysis with local to global relevance—a review," vol. 143, pp. 4–17, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037842901200295X

[2] I. A. Lakhiar, J. Gao, T. N. Syed, F. A. Chandio, and N. A. Buttar, "Modern plant cultivation technologies in agriculture under controlled environment: a review on aeroponics," vol. 13, no. 1, pp. 338–352, 2018. [Online]. Available: https://doi.org/10.1080/17429145.2018.1472308

[3] S. Chandra, S. Khan, B. Avula, H. Lata, M. H. Yang, M. A. Elsohly, and I. A. Khan, "Assessment of total phenolic and flavonoid content, antioxidant properties, and yield of aeroponically and conventionally grown leafy vegetables and fruit crops: a comparative study," vol. 2014, p. 253875, 2014.

[4] National Aeronautics and Space Administration, "Spinoff," 2006. [Online]. Available: https://www.nasa.gov/pdf/164449main_spinoff_06.pdf

[5] A. X. Wang, C. Tran, N. Desai, D. B. Lobell, and S. Ermon, "Deep transfer learning for crop yield prediction with remote sensing data," in *COMPASS '18*, 2018.

[6] M. Guerif, M. Launay, and C. Duke, "Remote sensing as a tool enabling the spatial use of crop models for crop diagnosis and yield prediction," in *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, vol. 4, 2000, pp. 1477–1479 vol.4, ISSN: null.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://www.nature.com/articles/nature14539

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," vol. 61, pp. 85–117, 2015. [Online]. Available: http://arxiv.org/abs/1404.7828

[9] L. Breiman, "Random forests," vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., ser. Information Science and Statistics. Springer-Verlag, 2000. [Online]. Available: https://www.springer.com/gp/book/9780387987804

[11] S. Raschka, *Python Machine Learning, 1st Edition*. Packt Publishing, 2015.