Affordable Agricultural Aid: Using A Novel Machine Learning Model To Predict Crop Yield With The Use Of Satellite Imagery, Climate, And Soil Data

Shubham Yadav, Ayaan Bargeer, Labesh Baral

ABSTRACT

The demand for agricultural production has increased dramatically, putting a strain on developing nations that lack proper technology to increase their crop production. Newly developed technologies, such as crop yield prediction models, are impractical in developing countries that have insufficient spatial data and field surveys. However, new technologies that allow for crop yield prediction without the use of pre-collected data can allow for greater agricultural production. This project aims to create a device that will implement a novel deep learning algorithm to predict crop yields, calculated in bushels per acre, three months before the harvest. The device collects data from its sensors to be run through the crop yield predictive model. The use of deep learning to extract important features for estimating yield is expected to reduce the dependency on input data for future predictions. First, a novel deep learning model was created to predict soybean crop yields in Iowa using three multivariate datasets: Satellite Imagery, Soil, and Climate Data. Using a Random Forest Regression algorithm a climate model was created and a Regression Artificial Neural Network (ANN) was used to create models for both Satellite Imagery and Soil. All three models obtained a relatively high accuracy and were combined using ensemble learning. A device was created using an Arduino, multiple sensors, and a breadboard. The results of a field test for the device, where a prediction for the 2021 harvest of soybeans in Upstate New York was made, indicated that the device was indeed legitimate. This device can help developing countries increase crop production through crop yield predictions without the need for pre-collected data.

1. INTRODUCTION

Over 80% of the world's food supply comes from agricultural produce. Surprisingly, the majority of this surplus is managed not by institutions, but by a collection of more than 500 million family farms [1]. These farms are located throughout the world and serve as vital components of the global economy. In the United States, active monitoring of growing conditions and agricultural production has given farmers a technological advantage in competing with traditional, small-scale agricultural methods. Data recorded by weather stations and regional agricultural surveys have created a foundation for new agricultural technology, allowing farmers to dramatically increase their yearly output.

However, these new advancements are restricted to first-world countries. Developing nations currently lack proper spatial data and field studies to create or use any new crop yield predictive technology. The lack of agricultural advancements is one of the main reasons why the population of Africa currently sees over 257 million malnourished individuals [2]. Africa utilizes more agricultural land than any other continent yet it has consistently performed below the global average in agricultural production per capita since the 1960s (Figure 1). The 2016 survey conducted by the FOA showed that Africa designates over 1.2 billion hectares of land for agricultural needs. Thus, it comes as a concern when the continent reported an agricultural per capita of 100 tonnes in 2010, a time when the global mean was close to 140 tonnes. The lack of financial resources and education has prevented farmers in developing countries access and utilize advanced technology so that they may maximize their crop yield.

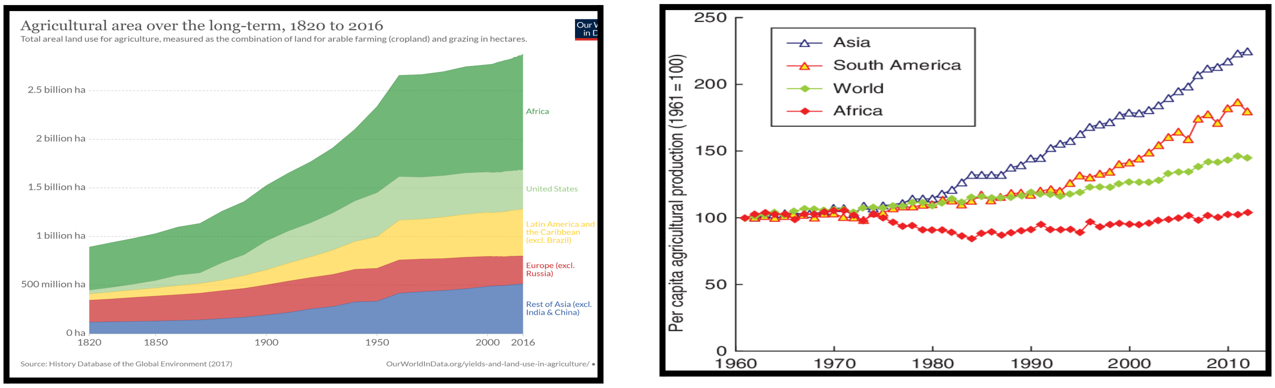


Figure 1: (left) Agricultural area from 1820-2016 in Africa, US, Latin America, Europe, and Asia(right) Comparison of per capita agricultural production in tonnes between Asia, South American, and Africa from 1960 to 2010[Include Reference].

1.1 Preserving Planet Earth through Agriculture Advancement

With a lack of advanced technology, developing countries often resort to extreme means to increase crop production which has severe environmental effects. Pesticides are excessively used when farmers are met with harsh environmental conditions [12]. Demand for agricultural land also increases, cutting into wild habitats to grow crops at a larger scale [12]. Harsh cultivation and planting methods often result in soil degradation, leading to even poorer crop yields [13]. By providing farmers with advanced technology, smarter and more environmentally friendly farming methods can be utilized in developing countries.

1.2 Lack of Available Data

A key factor for a lack of proper agricultural data lies in the difference in technological application in agriculture in comparison between the US and Africa. Unlike the United States, most of the villages in Ethiopia and Malawi, for example, rely on the use of subsistence farming. Farming is conducted on an individual basis and average farm sizes are much smaller than those of commercial growers in the states. Therefore, the government has less of an incentive to reinvest into the agricultural sector creating a lack of funding for research in agriculture. In addition, the large educational discrepancy when compared to the farmers of the western world reveals that most natives in Africa are illiterate and not accustomed to keeping agricultural records like land cultivation, crop yield, and farming techniques. Due to the lack of proper record-keeping and minimal oversight by weak governments, these regions lack the ingenuity and development that stimulates the agricultural economy of the United States. Furthermore, these differences and lack of proper data mean that western methods of using historical crop data to create accurate trends fail to appropriately predict crop yield in lesser developed countries.

1.3 Crop Yield Prediction

Crop yield prediction plays a large role in a country's economic systems and a farmer's ability to make appropriate financial and management decisions [9]. Agriculture contributes to 23% of Africa's GDP and is the primary labor force employing more than 60% of the population. Further analysis of the continent's economics shows that Africa is underperforming in relation to the global market. Most African countries place last in national GDP and have not shown any attempt at expanding. Failure to keep up with demand and adapt to growing economics continues to lag Africa behind the rest of the world. Africa's economics have left an indelible mark on society through poverty and malnutrition, in which more than half the population lives. With almost a quarter of its economy coming from agriculture, it is beneficial to give farmers access to the latest advancements in technology. Access to proper crop prediction technology in third-world countries will substantially increase knowledge on the success of their crop and enable them to increase productivity and allocate their resources efficiently. However, current efforts to use such technology have been largely unsuccessful.

In the presence of these challenges in forecasting, organizations like the FAO have looked towards local surveying in order to estimate crop yields[4]. This archaic form of data gathering requires scientists to analyze and record data individually. These methods are very time-consuming and expensive. In addition, the sheer number of farmers in the thousands of villages across each African country makes this an unrealistic approach to a national or a continental crop census. Process models like the CROPGRO-soybean model, GAEZ 63 model, and CERES-Maize model have been created as some of the most accurate prediction models. However, these models, like many others, demand an extensive amount of overly specific environmental conditions like plant photosynthesis in a given area [10]. Developing countries lack such data and are thereby unable to use these advanced models.

2. METHODOLOGY

The goal of this project is to create a device that can allow farmers anywhere in the world to accurately predict a crop's yield. The methodology of this project is structured into two parts. First, three deep learning models that predict crop yield will be created and combined, each using either satellite imaging, climate, or soil data. Testing will be conducted on the deep learning model, and optimization algorithms will be run to increase overall accuracy. Next, a device will be assembled using multiple sensors equipped for measuring climatic and soil conditions in the environment. The deep learning model will be implemented such that it will use the data obtained from the instrument along with satellite imaging in a given area to create a prediction for the expected yield of a particular crop. Field study tests will also be conducted to test the viability of the model.

2.1 Data Acquisition and Filtering

In order to create and train an accurate model, this project requires the gathering of 3 multivariate datasets: Satellite, Climatic, and Soil Data. Because this project serves as a proof of concept for the use of deep learning methods in agricultural production, the focus of the study was on predicting the soybean yield in the state of Iowa. Soybean was chosen as the base case for the model as its one of the most popular crops planted in African countries, with over 371 million hectares being used for production [8]. Iowa was chosen as the testing location as it is one of the leading soybean producers in the United States, and therefore had a large amount of data for the training process.

a) Satellite Data

Satellite Data was acquired from the United States Geological Survey (USGS) Agency's Earth Explorer platform. From this platform, Moderate Resolution Imaging Spectroradiometer

(MODIS) imagery was obtained from the Terra Satellite. Data were obtained from January 1st of 2012 to December 31st of 2018 with a multi-day temporal resolution. The MOD09A1 Version 6 data was acquired as it provided surface spectral reflectance of MODIS Bands 1 through 7 for atmospheric conditions (Table 1).

No. Bands	Band Name	Band Width (nm)
B1	Red	620 - 670
B2	NIR	841 - 876
B3	Blue	459 - 479
B4	Green	545 - 565
B5	NIR	1230 - 1250
B6	SWIR	1628 - 1652
B7	SWIR	2105 - 2155

Table 1. MODIS MOD09A1 7 Reflectance Bands

Once this data was obtained, the imagery had to be filtered for data that contained cloud or snow coverage. Images with a cloud coverage higher than 20% were removed from the dataset. Next, Normalized Difference Vegetation Index (NDVI), Land Surface Water Index (LSWI), Enhanced Vegetation Index (EVI), Normalized Difference Snow Index (NDSI) values were calculated for every pixel representing a geographic coordinate.

Index	Formula	Formula with Band Substitutions
NDVI	$NDVI = \frac{NIR - Red}{NIR + Red}$	$NDVI = \frac{Band\ 2 - Band\ 1}{Band\ 2 + Band\ 1}$
LSWI	$LSWI = \frac{NIR - SWIR}{NIR + SWIR}$	$LSWI = \frac{Band\ 2 - Band\ 6}{Band\ 2 + Band\ 6}$
EVI	$EVI = G * \frac{NIR - Red}{NIR + C_1 * Red - C_2 * Blue + L}$	$EVI = G * \frac{Band\ 2 - Band\ 1}{Band\ 2 + C_1 * Band\ 1 - C_2 * Band\ 3 + L}$
NDSI	$NDSI = \frac{(VISIBLE - SWIR)}{(VISIBLE + SWIR)}$	$NDSI = \frac{(Band\ 2 - Band\ 6)}{(Band\ 2 + Band\ 6)}$

Table 2. Vegetation Indices and their respective formulas. Note: For EVI calculations G is the gain factor, L is the canopy background and, C1 and C2 are coefficients of the aerosol resistance term.

Images were inputted into QGIS, an open-source platform for geographic information as raster layers. The layers were then trimmed to Iowa state boundaries of latitude and longitude coordinates. The index values were calculated for each formula using the respective band substitution formulas (Table 2). New layers were then created that displayed each index value for the state of Iowa. (Figure 2) depicts an example of a conversion from MODIS imagery to vegetation value, this case being an NDVI value. Each layer was then converted into a .xyz file, which allowed for each index value to be paired with a respective latitude and longitude point. Finally, the .xyz data was converted into a CSV file with a sample row containing [Year, Latitude, Longitude, NDVI, LSWI, EVI, NDSI].

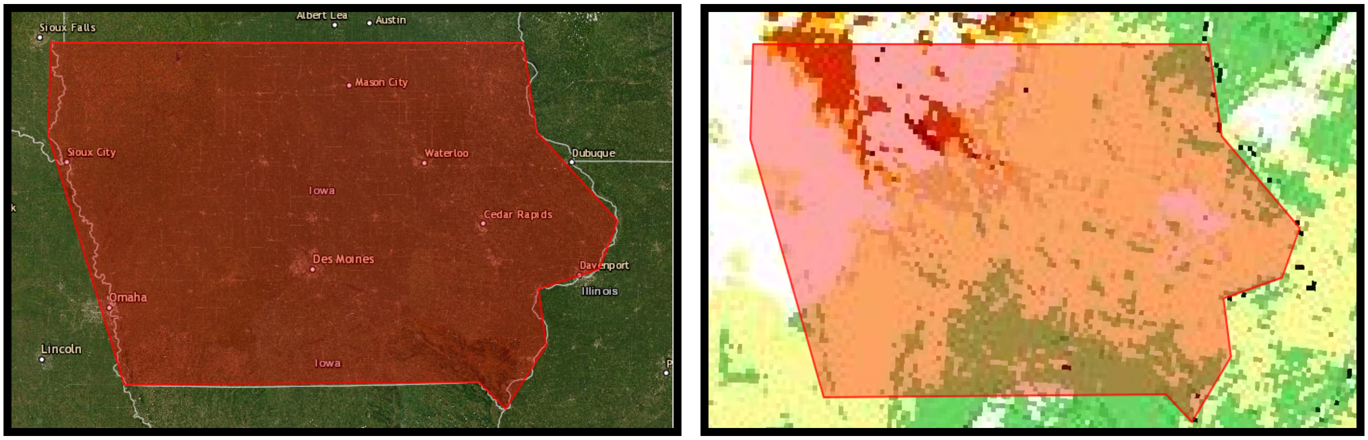


Figure 2. MODIS Imagery Converted into an NDVI Layer for December of 2013. The red overlay displays the bondings for Iowa State.

b) Climatic Data

Global climatic data was acquired from WorldClim, a database of high spatial resolution global weather and climate data. Annual bioclimatic data was obtained for 9 bioclimatic variables (Figure 3) in a .tif file. Using the Geospatial Data Abstraction Library (GDAL) within the Python console of QGIS, the data was exported into .xyz files that were organized by longitude, latitude, and bioclimatic value. The data was then cropped to only include values for the state of Iowa and finally, the .xyz files were converted to CSV files.

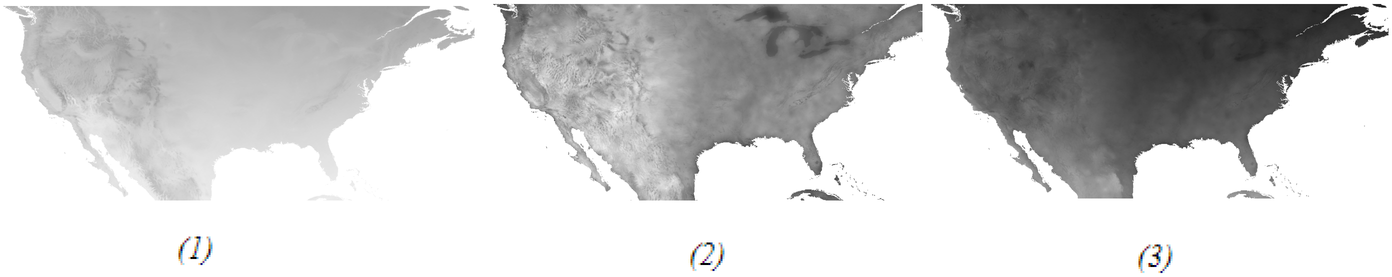


Figure 3. Maps Generated using QGIS depicting 3 of 9 Bioclimatic Variables. (1) - Annual Mean Temperature, (2) - Precipitation, (3) - Wind

c) *Soil Data*

Annual Soil Data was acquired from the Iowa Soil Properties and Interpretation Database (ISPAID) and the Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP). The success of a crop is dependent upon the suitability of its environment, therefore soil moisture data and pH levels were acquired from these datasets. The ISPAID database includes yearly survey data across each of the state's 99 counties and soil moisture data was taken on 2012-2018.

The MsTMIP project was used to acquire subsoil (30-100 cm deep) pH records in the state of Iowa from 2012-2018. The MsTMIP project is conducted with a spatial resolution of 0.25 degrees across the entire North America Region and includes a variety of soil characteristics for both topsoil and subsoil regions. PH data from the subsoil region was specifically chosen because it serves as a legitimate measure of soil quality. A study conducted by Dora Neina has found that subsoil pH gives important insight into soil health, nutrient availability, and the presence of pollution [3]. Humans are an important variable to consider as they inflict change upon the integrity of a landscape that may not be able to be recorded through satellite and climatic data. The use of soil pH fills this missing data and is appropriate [3]. Data from the MsTMIP project was found in .tif files and geographic masks were created in QGIS to target the Iowa region. This data was then converted into a .csv file and was combined with the soybean yield data. Figure 4 shows an example of the soil moisture and pH maps that were used in the collection.

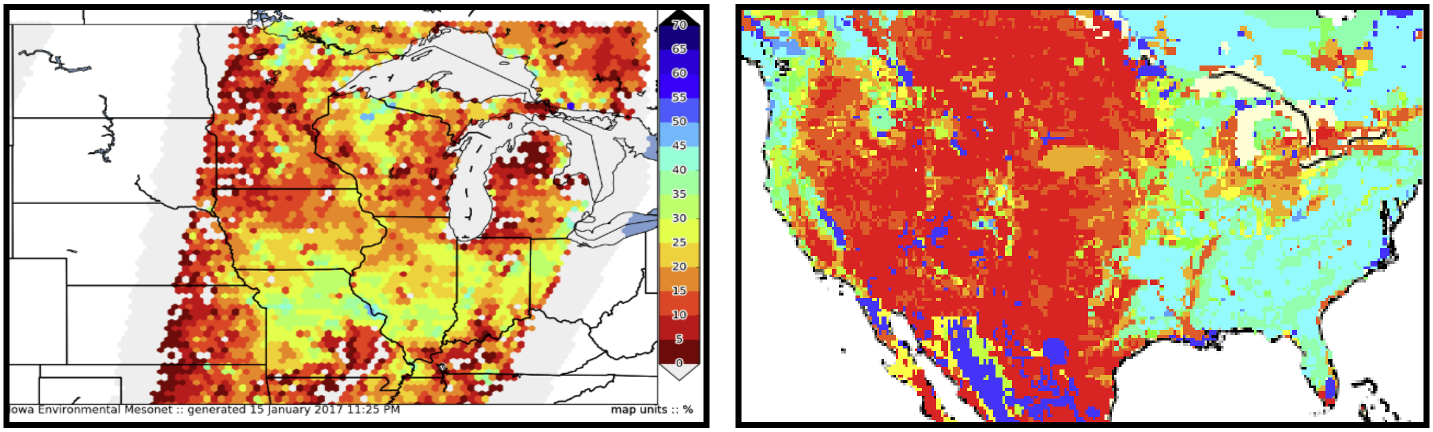

70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
owa Environmental Mesonet :: generated 15 January 2017 11:25 PM
map units :: %

Figure 4. Soil moisture data(left) and pH data (right) from July 1, 2018

2.2 Soybean Data Pairing

Soybean crop yield data was obtained from the United States Department of Agriculture's Agricultural Statistic Service. Soybean data ranged from 1970 - 2018 for the state of Iowa in each county. Soybean data was then paired with the satellite imagery, climate, and soil datasets with a three-month lag. Soybeans have a growing period of 2-3 months, therefore correlating crop yield data with satellite imagery, climate, and soil of three months prior to the harvest time would provide a dataset to train the model for a prediction at the beginning of the growing season. For satellite imagery each latitude and longitude point was converted into an Iowa county, and using a matching algorithm the data was paired with the soybean data. This process was repeated for the climate data. For soil data, the data was inputted directly into the matching algorithm and paired with the soybean data. The end result was three datasets that contained vegetation indices, bioclimatic data, and soil data respectively and a corresponding soybean yield for that time in Iowa.

2.3 Model Training and Testing

Using the matched datasets of satellite imaging, climate, and soil data with soybean production, three models were produced. The satellite imaging and soil data utilized a Neural Network Regressor algorithm, while the climatic data utilized the Random Forest Regressor.

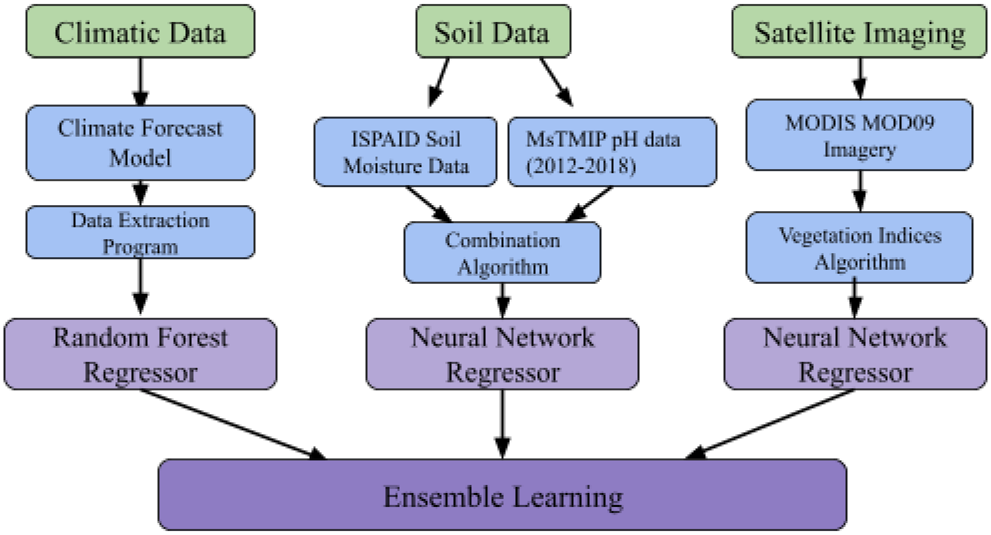


Figure 5. Diagram displaying methodology to create crop yield model.

Random Forest Training

Initially, the climate data was used as a training dataset - data that will be used to build the model. Then, in a process called bagging or bootstrap aggregation, the training dataset is sampled with replacement and distributed to different predictors. In Random Forests, these predictors are decision trees. A decision tree is a technique used in supervised machine learning in which a dataset is divided into smaller subsets based on decisions made at branches extending from each node. Each decision tree will be trained on its respective subset of the training data. When a prediction is made, a value traverses through all the decision trees. Since Random Forests is based on a "majority vote" system, whatever value the majority of decision trees outputs, is the value that is outputted by the model [14]. In this study, a regressor was used since the model will predict crop yield values based on the climatic data.

Figure 6: Equation utilized in Random Forest Regressor to produce an optimized model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Where *N* is the number of data points, *fi* is the value returned by the model and *yi* is the actual value for data point *i*.

To train the Random Forest model, the climatic data utilized an 80/20 split, where 80% of the data was used to train the model and 20% was used to test it. The model utilizes the mean squared error formula (Figure 6) to calculate the distance of each tree's prediction from the actual value. Using this equation the model was optimized to predict crop yield.

Regression Artificial Neural Network

Initially, the Random Forest Regression algorithm was used to train a model for both soil and satellite imaging. However, after receiving low accuracy values a Regression Artificial Neural Network [ANN] algorithm was used (Figure 7). Each multivariate regression model was trained 1000 times on a random 90% and 10% split.

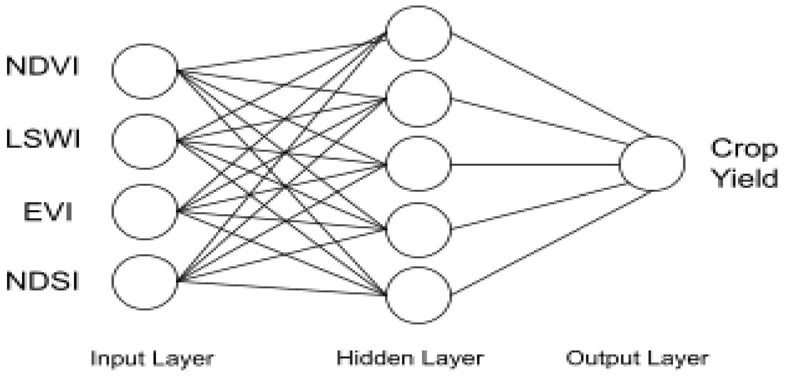


Figure 7. Image of the ANN Regression model's neurons and layers

Initial testing indicated a median error of 15% and 36% for satellite imaging and soil data respectively. To further increase the accuracy of the models, sigmoid was used as an activation function and the epoch size was increased to 6000.

Each model was then combined using stacking, a form of ensemble learning that uses a meta-regression. Each base level model was trained based on an overarching training set, and the meta-regression model was trained using the outputs. A loss function and mean square error were then computed for the regression ANN's and the Random Forest Regressor model respectively.

2.3 Crop Yield Prediction Device

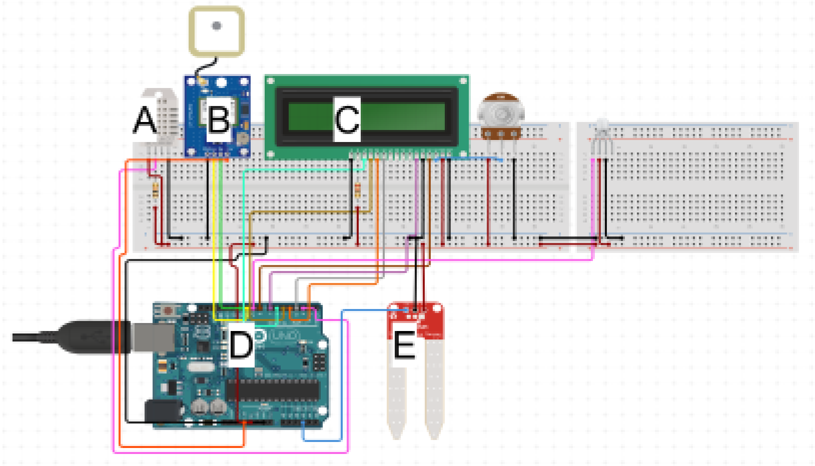


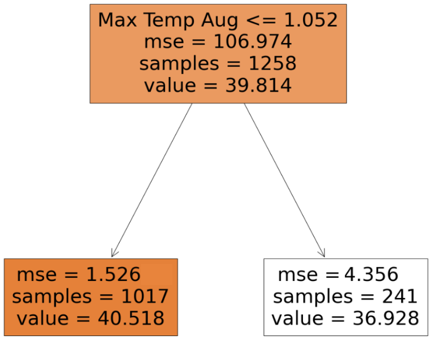
Figure 8. Schematic displaying wiring for Device. A) Temperature sensor. B) GPS and altitude sensor C) LCD Display. D) An Arduino board that connects to all the other parts as well as a computer via a USB connector. E) Soil moisture sensor

The device is composed of an array of sensors that were connected to a microcontroller. The microcontroller depicted is an Arduino Uno R3, chosen for its simplicity and cost-effectiveness. First, three sensors (Figure A) were attached to the microcontroller to accurately produce values for three bioclimatic variables; temperature, humidity, and soil moisture. The sensors additionally produce values for location in latitude and longitude, as well as altitude. Next, an LCD display and RGB LED diode were connected to the device. In the future, a ph sensor will be added, further increasing the practicality of the device.

3. RESULTS

3.1 Model Testing and Results

Each model was tested separately in order to measure accuracy based on each dataset. To test the climate model, feature importance was utilized to show that Max Temp Aug. had the greatest importance within each decision tree. Mean squared errors were computed for each node and the results indicate that the climate model had a mean square error as low as 1.526, with a progressive decrease in each split (Figure 9). These minimal errors indicate that the model predicted crop yields with a very high accuracy using climatic data.



$$R^2 = 1 - \frac{RSS}{TSS}$$

Figure 9. Mean Squared Error values and r^2 score equation for Climate (left), Satellite (right), and Soil Model (right)

The image on the right of (Figure 9) displays the r^2 score equation used to calculate the correlation between the predicted and actual values for the Soil Model. During testing, the soil model had an r^2 value of .776, indicative of a strong correlation between what was predicted and the actual. A similar success was achieved in the Satellite Model which had an r^2 value of

.865. The r^2 score indicates that the model predicts with fairly high accuracy and can be used as a viable tool for predicting crop yield.

3.2 Device Application, Testing, and Results

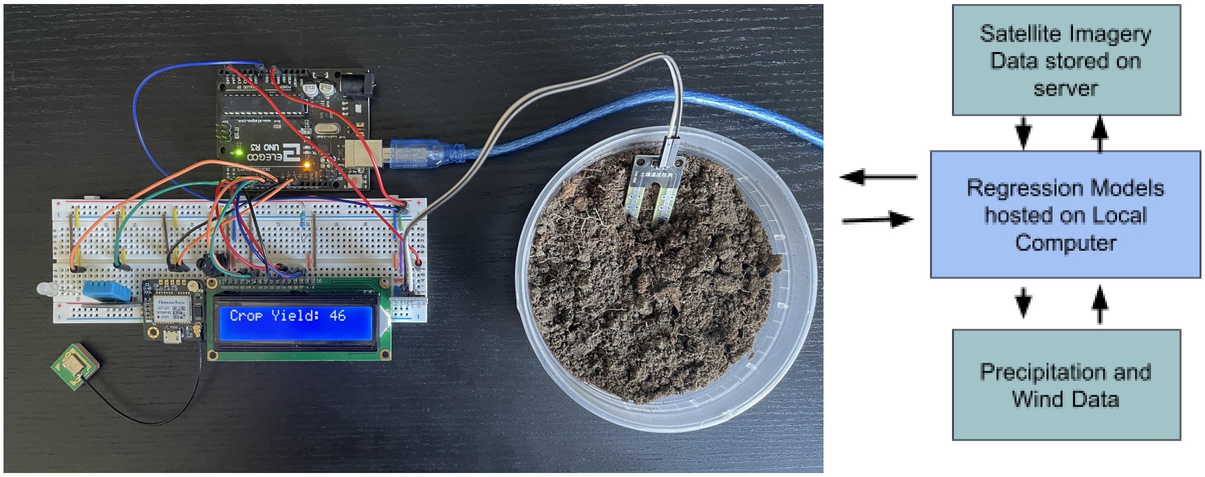
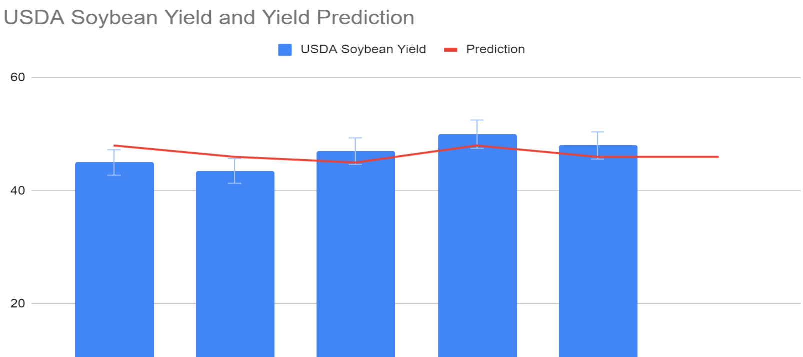

Crop Yield: 46
Satellite Imagery Data stored on server
Regression Models hosted on Local Computer
Precipitation and Wind Data

Figure 10. Prototype Device and Diagram displaying accurate crop yield output

The Arduino utilizes the 0.5kb bootloader to execute the preloaded data retrieval program. This program retrieves the corresponding data from the sensors connected to the microcontroller. This data is then sent back via the USB cable to the computer. The computer then retrieves the wind, precipitation, and satellite data from a local server. This data, originally obtained from the WorldClim and Earth Explorer database, is processed alongside the device inputted data within the regression model on the local computer. The value derived from the model indicates the crop yield, measured in bushels per acre, of the crop three months in the given future. This result is then sent back to the device to be displayed on the LCD (Figure 10).



In order to test the functionality of the device, a field test was conducted. (Figure 11) shows the results of a test that was conducted on soybean yield from 2015-2019 for Upstate New York. Results from the error bar in blue show that predictions made by the model lie within 5% of the ground truth. To conclude this test, a prediction was made for the upcoming harvest in the state of New York using soil in Albany, New York. The device outputs a yield prediction of 46 bu/ac (Figure 11). This output will be later compared with the actual output once surveys are released for 2021.

4. DISCUSSION AND CONCLUSION

The purpose of this project was to create a device that could accurately predict the specific crop yield in a given area. In order to do this, we created a device to work in conjunction with a novel machine learning ensemble that takes in climactic, soil, and satellite imagery to create a soybean yield prediction that will enhance production to any and all areas that are applied. The model was given testing data and displayed low errors, indicating a high accuracy. Testing results from soybean data from upstate NY support the practicality of this prediction device.

This device and the model have the potential to create a positive impact in preserving planet earth. The device will allow farmers to better plant their crops to maximize the harvest and use fewer environmentally dangerous resources during the planting season. In addition, maximizing crop yield results in a decrease in the necessity of agricultural land and environmentally degrading techniques. More efficient use of land indirectly preserves wildlife and reduces the negative environmental effects of excessive agricultural techniques. In the near future, we plan on experimenting with the success of our device on different crops like corn, wheat, and cereals.

5. REFERENCES

[1] Arsenault, C. (2014, October 16). *Family farms produce 80 percent of the world's food, speculators seek land*. U.S. https://www.reuters.com/article/us-foundation-food-farming-idUSKCN0I516220141016

[2]*Hunger Relief in Africa*. (2020, July 14). Action Against Hunger. https://www.actionagainsthunger.org/africa-hunger-relief-facts-charity-aid#:%7E:text=239.1%20million%20people%20in%20sub,513.8%20million%20people%20face%20hunger.&text=22.8%25%20of%20the%20population%20of,all%20regions%20in%20the%20world

[3] Neina, Dora. "The Role of Soil pH in Plant Nutrition and Soil Remediation." Applied and Environmental Soil Science, 2019. Gale Academic OneFile.

[4] Keita, N. Improving cost-effectiveness and relevance of agricultural censuses in africa: Linking population and agricultural censuses. 04 2019.

[5]J. (2021, April 9). *Power of 10: Top 10 Produce Crops in the U.S.* AgAmerica. https://agamerica.com/blog/power-of-10-top-10-produce-crops-in-the-u-s/

[6] *Soybeans - an overview | ScienceDirect Topics*. (2020). Food Supply Systems in Africa. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/soybeans

[7] Khaki, S. (2019). *Crop Yield Prediction Using Deep Neural Networks*. Frontiers. https://www.frontiersin.org/articles/10.3389/fpls.2019.00621/full

[8]*Agriculture food and nutrition for Africa - A resource book for teachers of agriculture*. (2020). Food Supply Systems in Africa. http://www.fao.org/3/w0078e/w0078e05.htm#:%7E:text=Under%20the%20current%20conditions%20in,sorghum%20(354%20million%20hectares)

[9] Horie, T., Yajima, M., and Nakagawa, H. (1992). Yield forecasting. Agric. Syst. 40, 211–236. doi: 10.1016/0308-521X(92)90022-G

[10] Cai, Yiqing & Moore, Kristen & Pellegrini, Adam & Elhaddad, Ayman & Lessel, Jerrod & Townsend, Christianna & Solak, Hayley & Semret, Nemo. (2017). Crop yield predictions - high resolution statistical model for intra-season forecasts applied to corn in the US.

[11] *The Economic Decline in Africa*. (2004). NBER. https://www.nber.org/digest/jan04/economic-decline-africa

[12] Ecobichon, D. J. (2001, March 7). *Pesticide use in developing countries*. PubMed. https://pubmed.ncbi.nlm.nih.gov/11246121/#:%7E:text=Chemical%20pesticides%20have%20been%20a,%2C%20clothing%2C%20etc.

[13] Yousaf, M. (2017, April 28). *Effects of fertilization on crop production and nutrient-supplying capacity under rice-oilseed rape rotation system*. Scientific Reports. https://www.nature.com/articles/s41598-017-01412-0

[14] Geron, Aurelien (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and Tensor Flow*. Sebastopol: O'Reilly.

[Fig 1] Ritchie, H. (2013, November 13). *Land Use*. Our World in Data. https://ourworldindata.org/land-use

[Fig 4] daryl herzmann akrherz@iastate.edu. (2021). *IEM :: Time Machine*. IEM Time Machine. https://mesonet.agron.iastate.edu/timemachine/#53.202104260000