Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 9 Issue 9, September - 2022

Machine learning Solution for Prediction of Soil Nutrients for Crop Yield: A Survey

Oladipe Ebenezer Oluwole
Computer Science Department
Federal University, Lokoja
Lokoja, Nigeria.
oladipebenezer@gmail.com

E.O. Osaghae
Computer Science Department
Federal University, Lokoja
Lokoja, Nigeria.
edgarosaghae@gmail.com

Fredrick D. Basaky
Computer Science Department
Federal University, Lokoja
Lokoja, Nigeria.
fdbasaky@gmail.com

***Abstract*-Crop yield prediction is a method for predicting crop production utilizing a variety of data such as temperature, rainfall, pH level, pesticides, fertilizer and other meteorological variables as well as features. Crop yield forecasting is based on a set of criteria. Crop yield forecasting is a science. one of precision agriculture's most difficult challenges. Furthermore, early accurate crop forecast is critical for agricultural production management; such predictions will also aid linked sectors in conceptualizing for their activities' logistics. In the past, several rules and applications for forecasting and showing agricultural yields were developed, with varied degrees of success. Because the majority of them are fundamentally empirical and so don't take into consideration weather and its features, and many crops were not examined. This has spurred a surge in interest in machine learning techniques. To increase estimation accuracy, environmental parameters such as temperature, rainfall, precipitation, relative, humidity and soil moisture must be used in Agriculture based on remote sensing yield estimating models. Hence recent techniques are introducing soil nutrients and location weather integration model for timely crop yield prediction to prevent blind planting and reduce farming time complexity.**

Keywords—crop yield, cultivation, forecasting, machine learning, plantation

I. INTRODUCTION

Crop yield refers to the per unit area of land, the quantity of crop harvested. It's a standard unit of measure for legumes, cereals, and grains, and it's commonly stated in bushels, pounds or tons per acre. Crop production is a key factor in determining the lengthy viability of agriculture. Crop yield is heavily influenced by environmental conditions. Intra-season yield variability is influenced by weather, which has an impact on crop growth and development. Soil quality interacts with weather in space to affect crop output, resulting in yield variability. (Prasad, 2020).

Crop yield forecasting is one of smart agriculture's most difficult challenges, and numerous models have been suggested and confirmed so far. Because crop production is affected by a range of factors such as weather, climate, seed, fertilizer, and soil type, this challenge necessitates the use of many datasets. (Xu *et al*., 2019). This suggests that predicting agricultural yields is not a simple operation; rather, it entails a series of complex stages. Crop yield prediction methods can now fairly approximate the actual yield, although greater yield prediction accuracy is still desired. (Klompenburg *et al*., 2020) That is why we want to integrate weather and location soil nutrients for crop yield prediction.

Data mining methodology is a versed computing domain, and is applicable where dataset is available. It is a multidisciplinary (statistics, machine learning and soft computing computational intelligent) domain of computing which uses its knowledge acquired from these disciplines to discover meaningful, interesting and useful patterns from specific domain data that are applicable in the domain.(Han *et al.,* 2011).

There's several drawbacks linked with using the conventional crop yielding approach because no consideration of sufficient environmental variables. This led to the renewed interests for the application of machine learning schemes in order to provide higher accuracy (such as random forest regression algorithm) for estimating crop yields (Sakamoto, 2020) That is why we want to integrate soil nutrients and location weather for early crop prediction even before planting using historical data.

ML is an AI that provides a strong framework for creating complex, automated, and objective

algorithms for analyzing high-dimensional and multimodal agriculture data. Machine learning approaches for making predictions, notably in precision agriculture in recent years. The bulk of these crop yield projections used supervised learning.

Data mining is described as the process of extracting previously unknown insights from a large amount of data. It is used in market research, manufacturing control, identity verification, client retention, and e-commerce, among other things. Based on wide customer Machine learning explores links and patterns in transaction records using queries. Depending upon the nature of data, data mining involves two sorts of functions being mined: descriptive functions that deal with generic features of data and predictive forms that identify patterns based on existing data. In agricultural, predictive types including as categorization, association, grouping, and regression are employed.

II. LITERATURE REVIEW

(Jayalakshmi and Devi 2021) used Supervised Machine Learning Algorithms like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree to estimate soil fertility using macro- and micronutrient status from a dataset. The decision tree generated greatest 99 percent accuracy with a low MSE rate.

(Rushika *et al*., 2018) research technique seeks to assist farmers in cultivating the right crop for higher yields. Their initiative examines the nutrients in the soil as well as crop yield in relation to location. It compares the efficiency of various network learning methods and gives the user the most precise outcome. Their proposed system provides the end user with accurate recommendations for fertilizers that are appropriate for each crop.

(Geetka 2018).Use data mining techniques to estimate wheat crop yields. The paper discusses the clustering method of k-means grouping, principal components analysis feature extraction, linear discriminant analysis classification, and modeling in the MATLAB environment Their approach has a 0.5 MSE rate and a peak demodulation ratio of 33.18 decibels.Nitrogen,Phosphourous,Pottasium, pH and soil conductivity, as well as temperature were parameters use by (Archana and Saranya 2020). Crop rotation, harvest output predictions, and fertilization recommendations are all built into their system. In their research, they created a system that includes an agricultural dataset and uses a voting-based ensemble classifier algorithm to recommend acceptable crops. This method had a 92% accuracy rating.

Several ensemble models were designed and used by (Shahhosseini *et al.,* 2020).Out-of-bag forecasts are generated using a blocked sequential process. Their predictions were generated on a county-by-county basis, and then averaged for agricultural districts and states. With an RRMSE of 9.5%, their suggested optimized weighted; their most exact algorithms were the ensemble and the average ensemble. The least controversial forecasts are made by stacked LASSO (MBE of 53 kg/ha)

Saeed et al. (2020) present a deep learning framework for agricultural production prediction focuses on ecological data and management strategies that employ recurrent neural networks and convolutional neural networks. Their proposed CNN-RNN model, as well as several other prominent approaches including random forest (RF), deep fully connected neural networks (DFNN), and LASSO, were used to estimate soya beans and maize yield using historical data. Their most recent design had RMSE of 9% for convolutional neural networks and 8% for the rest of the group for recurrent neural networks.

The efficacy of a hybrid algorithm based on recurrent neural networks and convolutional neural networks was investigated by (Khaki et al., 2020).The models were used to anticipate soya bean with corn yields. The authors' model exhibited a coefficient of correlation validation of 85.82% to 88.24%, as well as training root mean mean square error of 11.48 to 13.26%.

(Kim *et al.,* 2019) combined to estimate the amount of soybean and corn; researchers used satellite data and meteorology. Their findings revealed that the artificial neural networks model performed worse than deep neural networks in terms of prediction accuracy, which had an average prediction error of 7.6% and 7.8%, correspondingly (for maize and soybean). The constructed model has coefficient of correlation (r) of 0.95 and 0.90 for maize and soybean respectively.

(Rahman *et al.,* 2018) employed the normalized difference red edge index (NDRE) index, among other things, to forecast mango fruit yield. They used a mixture of plant vegetative indices, such as tree crown area and normalized difference red edge index (NDRE) area, to develop an artificial neural networks model with 0.7 R^2 , 13.83 kg•tree1 was the RMSE for the total fruit body mass.

Potatoes cultivated in an organic system was assessed using soil factors such as soil resistivity, content of organic carbon and water in the soil, together with the soil's microbiological state (the numbers of mesophilic and thermophilic fungi and bacteria were found out).A model constructed by ANN algorithm were employed using a modular feed-forward circuit with two convolutional layers and seven neurons in its structures gives MSE of 0.01. (Abrougui *et al.,* 2019).

(You *et al*., 2017) Based their research on remotely sensed sequence photos obtained just prior harvest; they employed deep learning method such as recurrent neural networks and convolutional neural networks to estimate soya bean production. In terms of Mean Absolute Percentage Error (MAPE), their

model beat standard remote-sensing based approaches by 15%.

Data related to soil, weather and past year production were used. They did data analysis using —anaconda navigator with Jupiter. Different classifiers namely Support Vector Machine, Logistic Regression, Random Forest, K-Nearest Neighbor, Decision Tree and were utilized, the precision 0.788 is obtained. The performance evaluation for each classifier algorithm were done and it was found out that Logistict regression gives 100%,Decision tree gives 93.3%, Random Forest gives 93.3%, K-Nearest Neighbor gives 86.66% and Support Vector Machine gives 60%. (Suganya *et al.,* 2020)

The paper reviews the effect of climate change from place to place; Historical data of the climate and the crop of a specific location were gotten for system implementation. There data were gotten from websites of different government. The climatic parameters they use are temperature, precipitation, vapour pressure, cloud cover and wet day frequency. As a result, data on all of these weather indicators was collected on a regular basis. It was decided to employ the Random Forests for International and Provincial Crop Yield Forecasting model. They created a website, and the accuracy of forecasts is greater than 75% in each of the plants and regions chosen. (Mayank *et al.,* 2020)

The study discusses that (Eswari *et al.,* 2018) said that many elements such as temperature, humidity, rainfall, and precipitation might impact crop output. Crop evapotranspiration is a new attribute she's introduced. Crop evapo-transpiration is a result of both the atmosphere and the growth of the plant. Predictive methods such as crop production prediction employing drought indicators across various time periods have shown that wheat and barley yields have at least 88% and 82%flexibility, respectively.

(Jayalakshmi and Devi 2019) used C5.0,K-Nearest Neighbor and Random Forest were used to develop a model to forecast if soil is "Ideal" or "Not Ideal" for growing crops depending on soil feature with high accuracy and efficiency. C5.0 accurately forecasted 96 percent of the time rate.

Machine learning and deep learning approaches were used in the suggested framework to forecast the optimal crop yield. The suggested model conducts an experiment on a crop dataset. They discovered that when using the recurrent neural networks (RNN), long-short term memory (LSTM) and support vector machine (SVM) methods, the precision was determined to be 97%. (Sonal & Sandhya 2021)

(Zhang *et al.* 2019) used XGBoost, random forest (RF), long-short term memory (LSTM), and LASSO algorithms to estimate county of China yield of Maize using environmental, fluorescence, thermal satellite and optical data. In the research, it was found out that solar-induced chlorophyll fluorescence (SIF) produced superior results than enhanced vegetation index (EVI) because to its low signal-to- noise and coarse spatial resolution ratio. Their performance metrics give the follow result- Value for R^2 of SIF: - RF: 0.75, LASSO: 0.39, LSTM: 0.68, XGBoost: 0.77 R^2 value for EVI: - RF: 0.76, LASSO: 0.38, LSTM: 0.69 XGBoost: 0.75.

III. PROPOSED METHODOLOGY

The methodology for our model follows the following steps which are the common techniques use in data mining project.

i. Agriculture Understanding- Bumper harvest depends on some agricultural conditions which must be strictly follow. We have understanding that not every farm are suitable for all crops, Individual crop can only yield maximally in a plantation that contain appropriate require nutrient and also weather condition suitable for its proper yield. Hence, we collected data from Agriculture Development Project (ADP) office Lokoja and we use the information for the training of the crop yield prediction Model.

ii. Data Understanding- In this step, a data understanding of the data collected will be carried out through the exploratory data analysis to report what the dataset entails by tabulating all the necessary parameters and also visualize the behaviors within the dataset. Here we will make use of plots and diagrams to see the relationship between the various features in the dataset.

iii Data Preparation- We will check for missing value and fill it, then we move to dropping features that might reduce the accuracy of our model keeping in mind the domain knowledge of the kind of system we are working on. Using Pearson Correlation, we are going to do a plot for the correlation between the features in our dataset in order to view the distinctfulness and usability to the model in making prediction. Furthermore, we normalize the available information by dropping the redundant data or the data with no importance to the model building. We also then divided the dataset into two-training set received 80% of the vote, whereas the testing set received 20%. We used training set for building the model and the testing data set will be used for the model validation.

iv Modeling- Keras classifier (Using TensorFlow backend), Random Forest Classifier, Decision Tree Classifier, and Gradient Boost Classifier algorithms from the “modelselection” library of the “SKlearn” package installed into the Python for model training.

v. Evaluation- We will use Classification accuracy to determine the accuracy, precision, recall, F1-score and support for the four (4) classifiers algorithms we want to use for building the model. We will visualize result by plotting a boxplot of all the

models according to their accuracy score using the seaborn and matplotlib libraries

IV. HIGH-LEVEL MODEL OF THE PROPOSED SYSTEM

The system's architecture is explained in detail in the high-level design. The architectural figure explains a full system, including the critical parts that should be created and also their relationships. They are straightforward models designed to aid comprehension, analysis, communication, and decision-making. The high-level design of the crop yield prediction system and its numerous components is depicted in the diagram below.

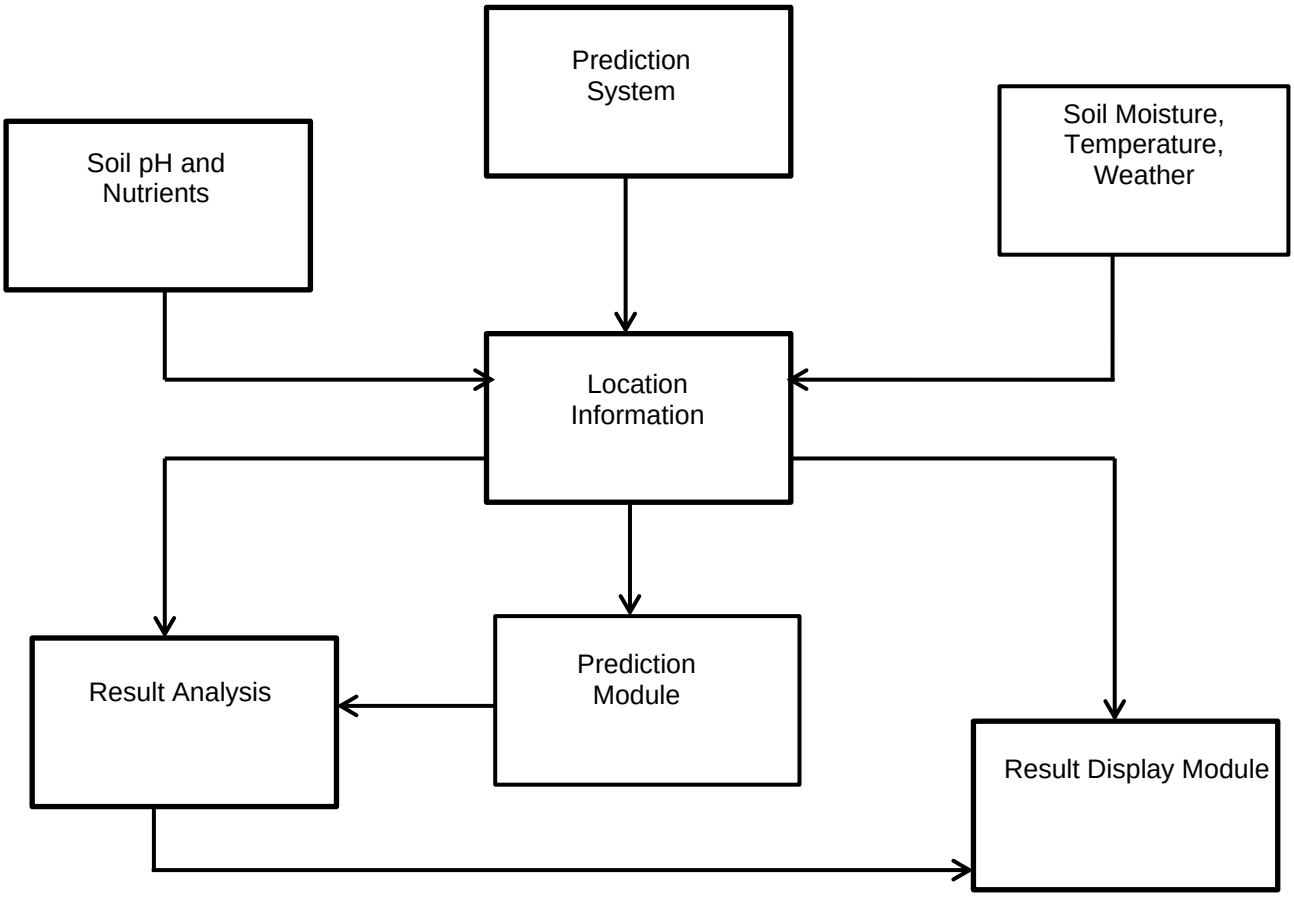

Figure 1: High Level Design of Integration of Soil Nutrient and Weather for Crop Yield Prediction System

- Location Information Module: This module handles the location or proposed plantation information which consists of Soil information include soil Nutrients(N.P.K), soil pH and location weather which includes temperature, relative humidity and moisture or rainfall.
- Crop yield Prediction Module: This is the system's main module; it takes the location information from the location information module and analyzes it using the created model, producing a prediction result that would be passed to the result analysis module.
- Result Analysis Module: In this module, the outcome of the prediction from the prediction module would be interpreted into an understandable format. The location details also would be analyzed and remarks are given to identify the level of each feature. All these interpreted details would be sent to the result display module.
- Result Display Module: The outcome from the prediction module alongside the detail from the analysis module are all received at the result display module and then organized by the display module to give output on the screen. The findings are presented in the form of bar graphs with percentage accuracy.

IV. CONCLUSION

Different systems employ various data mining methods to modify data in order to draw insights and assist farmers in making decisions. However, the main issue is that they somehow focus solely on a single crop or confine their considerations to soil minerals or local weather. Based on local historical data acquired from Agriculture Development Programme Lokoja, Kogi State Nigeria, this project is

utilized to estimate the optimum crop to be grown in a suggested farm area that will give a bumper harvest. Farmers can obtain the estimated production using a software platform. This makes it easier for farmers to choose which crop to sow.

Acknowledgement

I want to appreciate the untiring effort of my supervisors Dr. E.O Osaghae and Dr.F.D Basaky for their support and also my appreciation goes to Federal University Lokoja, Kogi State Nigeria for making their resources available for my use.

References

[1] Archana K. & Saranya K.:.Crop yield prediction, forecasting and fertilizer recommendation using voting based ensemble classifier. International Journal of Computer Science and Engineering (SSRG-IJCSE) 7(5). ISSN: 2348 – 8387 May 2020.

[2] Eswari K.,Swarna L.,Wilczek V., & Sudheer K.:Combining ability analysis for seed yield and its component characters in greengram. International Journal of Chemical Studies. 6(2) 237-242 February 2018.

[3] Geetika R.: Crop yield prediction using data mining: an efficient data modeling approach International Journal of Engineering & Technology. 7(2.27) 128-131 August 2018.

[4] Jayalakshmi R. & Devi M.:Relevance of machine learning algorithms on soil fertility prediction using R. International Journal of Computational Intelligence and Informatics, 8(4), 193-199. 2019

[5]Jayalakshmi R. & Devi M.,:Predictive model construction for prediction of soil fertility using decision tree machine learning algorithm.Kongunadu Research Journal 8(1): 30-35 June 2021

[6] Kim N., Ha K.,Park N.,Cho J.,Hong S. & Lee Y.:A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. International Journal of Geo-Information 8(5) 240. May 2019.

[7] Klompenburg, T., Kassahun, A., & Catal, C.: Crop yield prediction using machine learning : A systematic literature review. Computers and Electronics in Agriculture. 177(10) 105709. https://doi.org/10.1016/j.compag.2020.105709. October 2020.

[8] Mayank C., Darpan C., Chaitanya C., & Mansing R.:Crop yield prediction using machine learning. International Journal of Science and Research (IJSR) 9(4) 645-648 April 2020.

[9] Paliwal, A., & Jain, M.:The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. Frontiers in Sustainable Food Systems, 4, 1–10. https://doi.org/10.3389/fsufs.2020.00025 March 2020.

[10] Rahman M.,Robson A. & Bristow M.: Exploring the Potential of High Resolution WorldView-3 Imagery for Estimating Yield of Mango. Remote Sensing Journal. 10(12), 1866. November 2018.

[11] Rushika G., Juilee K., Pooja M., Sachee N. & Priya R.: Prediction of crop yield using machine learning. International Research Journal of Engineering and Technology (IRJET).5(2) 2237-2239. February 2018.

[12] Saeed K., Lizhi W. & Sotirios V.: A CNN-RNN framework for crop yield prediction. Frontiers in Plant Science 10: 1750 January 2020.

[13] Sakamoto, T.: Sensing incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. ISPRS Journal of Photogrammetry and Remote Sensing, 208–228. https://doi.org/10.1016/j.isprsjprs.2019.12.012 February 2020.

[14] Shahhosseini M., Hu G. & Archontoulis S.: Forecasting corn yield with machine learning ensembles. Frontier Plant Science. 11:1120. July 2020.

[15] Sonal, A. & Sandhya, T.: A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. Journal of Physics.: Conference. Ser. 1714 012012. 2021

[16] Suganya M., Dayana R., & Revathi.R.: Crop yield prediction using supervised learning techniques.International Journal of Computer Engineering & Technology (IJCET) 11(2) 9-20. July 2020.

[17] You, J., Li, X., Low, M., Lobell, D., & Ermon, S.: Deep gaussian process for crop yield prediction based on remote sensing data. Conference on Artificial Intelligence (San Francisco, CA), 4559–4566 .2017

[18] Zhang L., Zhao Z., Yuchuan L., Juan C., & Fulu T. : Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in china using machine learning approaches. Remote Sensing, 12(1):21, 2020. https://doi.org/10.3390/rs12010021. December 2019.