Comprehensive Survey for Crop Yield Prediction Using Machine Learning

[1]Dr. Laxmi B Rananavare , [2]SpandanaM, [3]Dr. Sanjay Chitnis

[1,2]Dept. Of CSE, REVA University Bengaluru, India, [3]Dept. of CSE, R V University Bengaluru, India

[1]laxmib.rananavare@reva.edu.in, [2]spanduspandy5@gmail.com,
[3]sanjay.chitnis@reva.edu.in

Abstract.

A comparison examination of several agricultural production prediction models is proposed in this work. In the agricultural industry, a lot of study is done to estimate the production of different crops. For many years, the research has been conducted for agricultural areas all over the world. Different parameters such as rainfall, temperature, soil type and pH are taken into consideration. Most of crop dataset driven by the government websites consist fields such as crop name, state name, season, production and so on. Recent research has shown that remote sensing is an effective way for estimating yields, and that machine learning, particularly deep learning, can predict a decent forecast by combining multisource datasets including such temperature data, Satellite data, soil data, and etc. Addition of remote sensing imagery has a direct impact on accuracy of the models. For many models, both ground truth data and MODIS NDVI pictures are obtained from several sources. To compare and evaluate the performance of such models, R2 score which also known as the coefficient of determination and root-mean-square error (RMSE) were utilised. The impact of machine learning algorithms in evaluating yield is revealed in this study. This study discusses several yield predictions models in order to analyse previous work and come up with a fresh approach for properly predicting yield.

Keywords.
YieldPrediction;MachineLearning;deepneuralnetworks;regression;SVM;XGBregression;Linearregression;DecisionTree regression;

1. INTRODUCTION

Agriculture is the most important sector in India's economy. Predicting crop yields has become a critical challenge in agricultural field. Farmers everywhere is eager to know how much harvest they will receive. Prediction was made by the farmers based on their past experience cultivation expertise. However, the Yield is depended and influenced by a number of factors. characteristics like dirt (soil type, nutrients in the soil, and ph) meteorological conditions (rainfall, temperature, etc.) and soil level due to these unpredictably changing conditions throughout the year and Crop yield forecast is a difficult task in every season. Farmers face a challenge based solely on their previous experience knowledge.

Machine learning is the most suitable domain forforecasting real-time problems like weather forecasting, infection forecasting, and so on. It is divided into three categories: unsupervised learning, reinforcement learningandsupervised learning,each type has its own set of algorithms, each with its own set of goals. In the agricultural area, machine learning is used for yieldprediction of particular crops. It examines past years' agricultural data in order to extract useful information, employing algorithms from its various categories, a model is trained that predicts future yield. Random forest, support vector machine,KNN, decision tree classifiers, naive bayes classifier, and other techniques are commonly employed. Both regression and classification tasks are handled by machine learning. A regression problem is yield prediction based on agricultural data.Crop yield prediction is required not only to determine the crop's future production, but also which are important to do timely import and export decisions, as well as agricultural risk management decisions, in order to assure food security globally and avoid the waste of money and resources spent in cultivation. The purpose of this paper is to review existing crop yield prediction research in order to determine which factors have the most impact on yield outcomes and which algorithms provide higher yield estimation accuracy.

2. RELATED WORK

[1] India was selected as the study region due to its climatic variation. To perform the study major crops of India such as jowar, rice, tobacco, bajra and wheat were selected and the dataset for the same were collected originally from www.mospi.gov.in and https://data.gov.in.Features such as rainfall, crop names, area under irrigation, are, production and seasons were present in the dataset. Machine learning techniques such as Ridge regression, Linear regression, Lasso regression and Decision tree were applied and Decision tree gave more accurate results.

[2] In addition to Lasso regression this paper used other two techniques of regression such as E-Net and kernel regression. To enhance the performance of the algorithms Stacking Regression concept was induced. Among all the algorithms kernel ridge performed well and after applying stacked regression performance of the model was even more enhanced.

[3] This research explained the step-by-step process for building a Crop yield prediction model. Explained the importance of each and every process starting from data collection, feature selection till evaluation of different machine learning techniques such as DNN, KNN, ANN, CNN. Considering parameters such as rainfall, temperatureand acer was the main goal of the research.

[4] Embedded Machine learning Model was created for prediction of the crop yield. Soil management was the main focus and reducing the usage of fertilizers was the main goal. The system was able to predict that which crop is best suited to grow under the current soil conditions, so that the farmers will be benefited by the best yield. Real time Data from different sources were collected, combining IOT with ML contributed in collecting the real time data.

[5] Nine Indian crops such as Jowar, Soyabean, Bajra, Sugarcane, Cotton, Corn, Wheat, Rice, Groundnut were used to build the prediction model. Separate dataset was formed by combing the dataset from different sources such as Kaggle and West Bengal government website. Deep Learning such as ANN was used to develop a crop estimation system. Including parameters of soil and rainfall data. Prediction was determined using a web-based system.

[6] Self obtained dataset was used to develop a crop yield prediction model. Different machine learning algorithms such as KNN, Decision tree and Naïve bayes was used to determine the relationship between various Physical factors and crops. The study says that physical factors such asSoil type, soil pH, temperature contribute a lot to the yield prediction. Prediction for crops such as tomato, chilli and potato were done in different types of soil such as red soil, black soil and alluvial soil and Decision tree and KNN performed well compared to naïve bayes.

[7] U.S was selected as study region, dataset used was having both seasonal yield and whether data of wheat. Bi-LSTM was built with CNN feature extraction sub-network which predicted the yield with more accuracy and less overfitting. Further theBi-LSTM model was compared with RF, KNN, Polynomial regression, SVR and Naïve Bi-LSTM. As the results all other models except for Bi-LSTM out performed. Performance Evaluation was based on R2 score and MSE.

[8] This research states that applying the prediction analysis before harvesting the crop would help the farmer with enhancement of production. Different machine learning algorithms such as Bayes Net, SVR, Kernel Ridge RBF, GPR, lasso regression and RNN were applied. Crops such as Rice, Alfalfa, Tomato and Maize were selected to build the model. Results showed that RNN worked well when compared by error rate and Bayes Net worked well when compared by accuracy rate.

[9] Maharashtra was the study are selected, major crops of Maharashtra such as Bajra, wheat, Jowar and rice were included for developing the model. Dataset for the research was derived from the Indian Government Website with consists of parameters like cultivation area, state, crop, season and district. ANN showed the best results after fitting the model using linear regression with Neural network.

[10] Four machine learning methods namely GPR, KNN regression, Decision Tree and Back Propagation Neural Network were used to compare the prediction Model. Barley production in Iran was the area of interest of this study. Combining field data with remote sensing data was carried out. Field data was collected from government website and remote sensing data was collected from GEE platform. Parameters such as EVI, Minimum Temperature and precipitation was taken into consideration. The study states that the production is also dependent on varying climatic changes. Results show that among all four algorithms GPR gave best results. The model was able to predict the yield one month before the harvest.

[11]Applied techniques such as, LSTM, Q-learning and RNN on the crop paddy for Vellore region in India. Collected Crop Related Weather data of Gujarat Paddy for time period (1997-2012) year were collected, rest of the such as Temperature, Precipitation were collected from government Websitehttps://www.kaggle.com/kpkhant007/gujarat-crop-related weather-data-199720. The results showed that for RNN different errors such as MAE, RSME and MSE were very high compared to other proposed models. The hybrid model LSTM- RNN with Q-learning acquired highest accuracy and R2-Score. The model was examined by executing Light GBM and SVR separately with the dataset, and then evaluating the LGBM-SVR hybrid model and analyzing the differences.

[12] Extracting the trends such as NDVI, average temperature, Air quality index and precipitation which were present in the metrological data was collected from government of India website and Remote sensing data was collected from MODIS for developing the framework. After data collection Image processing is followed by NDVI calculation and then applying the machine learning technique Extreme Gradient Boosting to predict the crop production. The model was used to predict the rice crop in Tamil Nadu.

[13] Multispectral data for wheat and barley from city of Pori was collected using sensors. CNN algorithm was used for prediction and the performance evaluation was done in the terms of Mean Absolute error.

[14] China was the study region, satellite remote sensing data for corn and soyabean were collected from NASA. The machine learning techniques such as ERT, RF ,SVM, and DL were used to build the model. When compared the performance of DL was better among all four.

[15] The large dataset of 7000 vineyard blocks from Australia was selected to train the model. Sentinel-1/2 were used to derive the images for the selected region. ML techniques namely Neural Networks, Linear Regression, Random Forest, SVM Regressor, Ridge Regression and Kernel Ridge Regression were used to estimate the soil moisture. The results showed that slightly similar results were given by neural networks and kernel regression. RF model was performing better among all.

[16] Dataset collected was the MODIS images of Aqua and terra which was from the satellite MODI31Q and MYDI31Q respectively from earth explorer website. SylhetHaor region from Bangladesh was selected as the study region for Boro rice prediction which was compared with the existing model. Brute force technique was used to find the NDVI threshold. By improvising NDVI threshold followed by the

accuracy model gave the best results.

[17] Winter wheat yield prediction from satellite images data was done for the cropland site in Poland and south Africa. Prediction of model was based on NDVI, from Sentinal-2 images 32 different types of indices were calculated. MODIS data allowed the model to predict with best accuracy.

[18] Different techniques such as spatial stream, temporal stream, fusion of spatial

and temporal stream was used for prediction. High spatial resolution (NAIP), High temporal resolution (250m spatial resolution), bi-weekly temporal resolution (MODIS) imagery was used for analysis. NDVI was also taken into reference which helped in predicting the yield more accurately.

[19] Corn yield prediction for the Midwestern US was done. Dataset consists of Satellite remote sensing data from NASA, ESA, CCI and few of the parameters such as NDVI, EVI, LAI, FPAR, GPP and ET was also obtained from images. Climate data from http.www.prism.organstate.edu/ was used which consisted parameters such as precipitation, maximum temperature, minimum temperature, TDmean, VPDmin, VPDmacx. Machine learning algorithmsnamely DL, RF, ERT, and SVM were implemented. Result analysis was done based on MAE, RSME, MAPE, and r which showed that DL method has the highest accuracy with correlation co-efficient of 0.776 and RSME of 0.844 ton/ha.

[20] Soybean yield prediction was carried out in Alabama state Lauderdale County of USA. Dataset used was Satellite images which was derived from NASA's MODIS. 3DCNN technique was used for building the model and the evaluation was done based on RMSE.

3. METHODOLOGY

"Fig. 1" Shows the flow diagram of crop yield prediction.

A. Data Collection

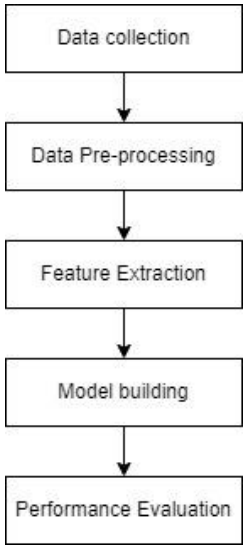

Fig 1. Flow diagram of yield prediction

All machine learning difficulties begin with the data collection process. Because the final model's correctness is a direct reflection of data, the two important things to be taken into consideration while gathering data are quantity and quality of data. Data can be collected from a different place, including Kaggle, Indian government websites, and so on. Files, databases or photos are some of the forms of data that can be collected.

B. Data Pre-processing

The crucial phase in machine learning is data preprocessing. Unstructured data,null values, excessively big values, duplicate values or missing values, may be found in the collected data, affecting the research's outcome. As a result, the data must be preprocessed to remove any contaminants and prepare it for model construction. This stage also involves data visualization, which is necessary for finding data imbalances and correlations among variables.

C. *Selection of Algorithm*

This is the final stage in the data pre-processing process. The problem of selecting

algorithms from a large number is critical. Each machine learning algorithm serves a distinct purpose. The algorithm we choose is determined by our run time, issue statement and data set.As per our requirements we have to select the most appropriate algorithm.

D. ModelBuilding

The dataset is divided into two sections: testing and training. The dataset followed the 80/20 rule. 80% of data is used to train the model,20% of data is used to test the model.Training was modelled using the chosen algorithm. The training stage is referred to as the "heart" of machine learning issues. The built model was then put to the test with the testing set to see how well it worked.

E. Performance Evaluation

Metrics from the machine learning domain were used to assess the model's performance. A range of measures are available for evaluating regression and classification difficulties. The Attributes such as MAE, MSE, RMSE, r2 score and Adjusted r2, are used to evaluate the regression problem's performance. classification problems are solved using confusion matrix, recall, Accuracy, sensitivity,F1 score,specificity, and precision.

F. Deploy the Model

After the model was built and completed the performance evaluations. The model with the lowest error and highest accuracy will be picked to deploy the model for future prediction of real-time problems.

4. CONCLUSION

Machine learning is primarily used to predict real-time issues. As a result, it will be valuable in predicting agricultural yields. Many studies have been conducted to forecast the yield of varieties of crops. The proposed project will look at some yield prediction models in order to evaluate prior work and produce comparison data.

5. FUTURE WORK

The majority of the study work forecasts rice, sugarcane and maize yields based on comparative yield prediction results. This research will be expanded in the future to

estimate crop yields such as jowar, bajada, and others. Only a few input characteristics are taken into account in current crop yield predictions, such as wheat and rice. To improve accuracy, we will evaluate all of the relevant input characteristics for yield prediction in the future.

6. REFERENCES

[1] Kavita,M.,&Mathur,P.(2020,October).CropyieldestimationinIndiausing machine learning. In *2020 IEEE 5th International Conference onComputingCommunicationandAutomation(ICCCA)* (pp.220-224).IEEE.

[2] Nishant, PotnuruSai, PinapaSaiVenkat, Bollu Lakshmi Avinash, and B. Jabber. "Crop yield prediction based on Indian agriculture using machine learning." In 2020 International Conference for Emerging Technology (INCET), pp. 1-4. IEEE, 2020.

[3] Malik, P., Sengupta, S., & Jadon, J. S. (2021, January). Comparativeanalysis of soil properties to predict fertility and crop yield using machinelearning algorithms.In *202111thInternationalConferenceonCloudComputing, Data Science & Engineering (Confluence)* (pp. 1004-1007).IEEE.

[4] Nalwanga, Rosemary, Jimmy Nsenga, Gerard Rushingabigwi, and IgnaceGatare. "Design of an Embedded Machine Learning Based System for an Environmental-friendly Crop Prediction Using a Sustainable Soil Fertility Management." In *2021 IEEE 19th Student Conference on Research and Development (SCOReD)*, pp. 251-256. IEEE, 2021.

[5] Mondal, A., & Banerjee, S. (2021, October). Effective Crop Prediction Using Deep Learning. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1-6). IEEE.

[6] Malik, P., Sengupta, S., & Jadon, J. S. (2021, January). Comparativeanalysis of soil properties to predict fertility and crop yield using machinelearning algorithms.In *202111thInternationalConferenceonCloudComputing, Data Science & Engineering (Confluence)* (pp. 1004-1007).IEEE.

[7] Dai, C., Huang, Y., Ni, M., & Liu, X. (2020, November). Wheat Yield Forecasting using Regression Algorithms and Neural Network. In *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)* (pp. 129-134). IEEE.

[8] Chandraprabha, M., &Dhanaraj, R. K. (2020, November). Machine learning based Pedantic Analysis of Predictive Algorithms in Crop Yield Management. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1340-1345). IEEE.

[9] Kale, S. S., & Patil, P. S. (2019, December). A machine learning approach to predict crop yield and success rate. In *2019 IEEE Pune Section International Conference (PuneCon)* (pp. 1-5). IEEE.

[10] Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, *101*(3), 891-896.

[11] Patel, J., Vala, B., &Saiyad, M. (2021, April). LSTM-RNN Combined Approach for Crop Yield Prediction On Climatic Constraints. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1477-

1483). IEEE.

[12] Shah, A., Agarwal, R., &Baranidharan, B. (2021, March). Crop Yield Prediction Using Remote Sensing and Meteorological Data. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 952-960). IEEE.

[13] Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Computers and electronics in agriculture*, *163*, 104859.

[14] Kim, N., & Lee, Y. W. (2016). Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, *34*(4), 383-390.

[15] Efremova, N., Seddik, M. E. A., &Erten, E. (2021). Soil moisture estimation using Sentinel-1/-2 imagery coupled with cyclegan for time-series gap filing. *IEEE Transactions on Geoscience and Remote* Sensing, 60, 1-11.

[16] Kalpoma, K. A., & Rahman, A. (2021, July). Web-Based Monitoring of Boro Rice Production Using Improvised NDVI Threshold of MODIS MOD13Q1 and MYD13Q1 Images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* (pp. 6877-6880). IEEE.

[17] LK, Sowmya Sundari, Mukul Rana, Syed Thouheed Ahmed, and K. Anitha. "Real-Time IoT Based Temperature and NPK Monitoring System Sugarcane-Crop Yield for Increasing." In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1-5. IEEE, 2021.

[18] Gurdak, R., Dabrowska-Zielińska, K., Bochenek, Z., Kluczek, M., Bartold, M., Newete, S. W., &Chirima, G. J. (2021, July). Crop Growth Monitoring and Yield Prediction System Applying Copernicus Data for Poland & South Africa. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* (pp. 6564-6567). IEEE.

[19] Gadiraju, K. K., Ramachandra, B., Chen, Z., &Vatsavai, R. R. (2020, August). Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3234-3242).

[20] Kim, N., & Lee, Y. W. (2016). Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, *34*(4), 383-390.