Crop Yield Prediction Using Machine Learning Algorithm

Abstract

Agriculture is the pillar of the Indian economy and more than 50% of India's population are dependent on agriculture for their survival. Variations in weather, climate, and other such environmental conditions have become a major risk for the healthy existence of agriculture. Machine learning (ML) plays a significant role as it has decision support tool for Crop Yield Prediction (CYP) including supporting decisions on what crops to grow and what to do during the growing season of the crops. The present research deals with a systematic review that extracts and synthesize the features used for CYP and furthermore, there are a variety of methods that were developed to analyse crop yield prediction using artificial intelligence techniques. The major limitations of the Neural Network are reduction in the relative error and decreased prediction efficiency of Crop Yield. Similarly, supervised learning techniques were incapable to capture the nonlinear bond between input and output variables faced a problem during the selection of fruits grading or sorting. Many studies were recommended for agriculture development and the goal was to create an accurate and efficient model for crop classification such as crop yield estimation based on the weather, crop disease, classification of crops based on the growing phase etc., This paper explores various ML techniques utilized in the field of crop yield estimation and provided a detailed analysis in terms of accuracy using the techniques.

Introduction:

Agriculture is the backbone of India's economy since its plays a vital role in the survival of every human and animal in India. The worldwide population was estimated at 1.8 billion in 2009 and is predicted to increase to 4.9 billion by 2030, leading to an extreme increase in demand for agricultural products. In the future, agricultural products will have higher demand among the human population, which will require efficient development of farmlands and growth in the yield of crops. Meanwhile, due to global warming, the crops were frequently spoiled by harmful climatic situations. A single crop failure due to lack of soil fertility, climatic variation, floods, lack of soil fertility, lack of groundwater and other such factors destroy the crops which in turn affects the farmers. In other nations, the society advises farmers to increase the production of specific crops according to the locality of the area and environmental factors. The population has been increasing at a significantly higher rate, so the estimation and monitoring of crop production is necessary. Accordingly, an appropriate method needs to be designed by considering the affecting features for the better selection of crops with respect to seasonal variation.

The core objective of crop yield estimation is to achieve higher agricultural crop production and many established models are exploited to increase the yield of crop production. Nowadays, ML is being used worldwide due to its efficiency in various sectors such as forecasting, fault detection, pattern recognition, etc. The ML algorithms also help to improve the crop yield production rate when there is a loss in unfavourable conditions. The ML algorithms are applied for the crop selection method to reduce the losses crop yield production irrespective of distracting environment.

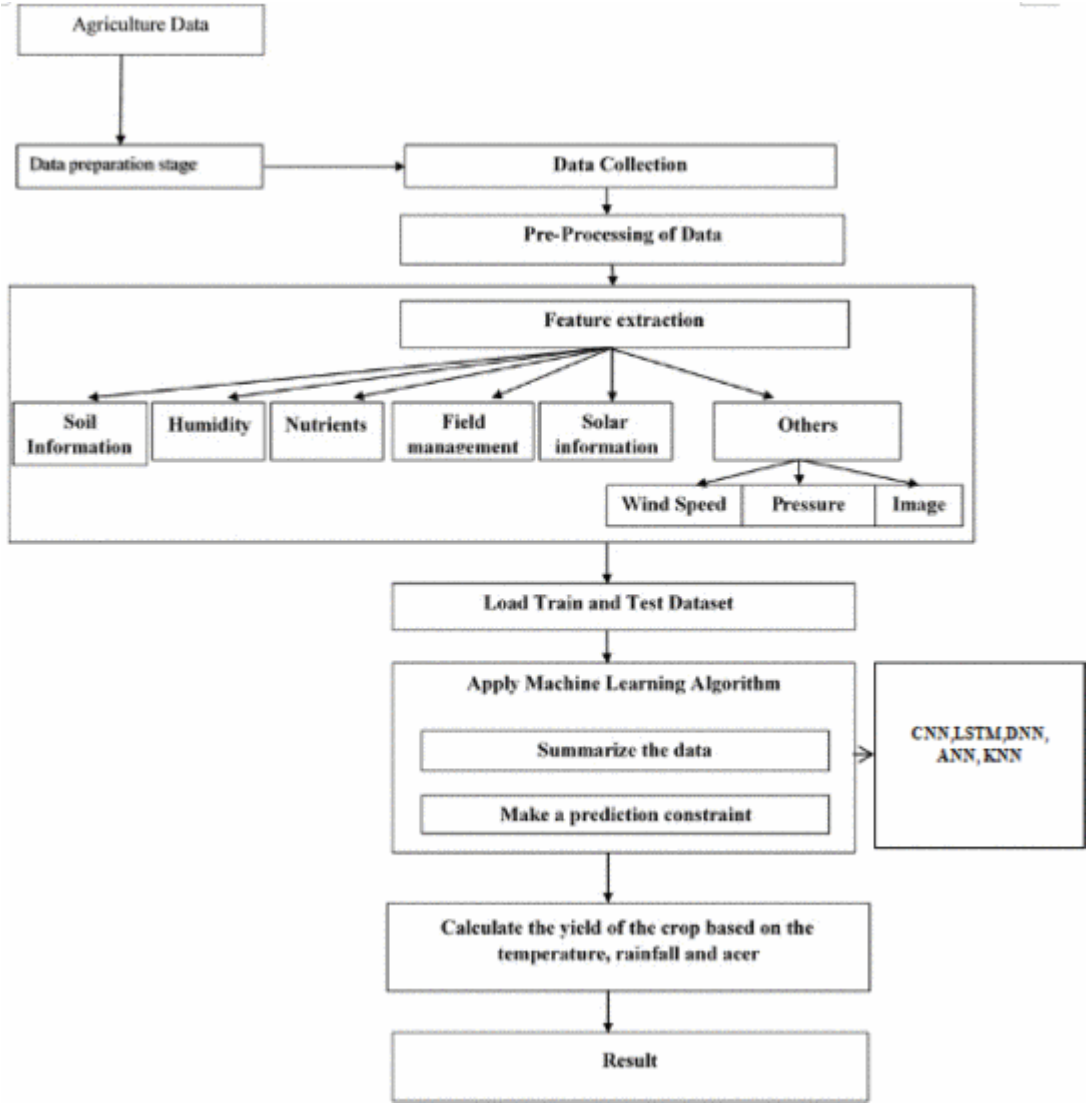
The existing model used SVM that classified the crop data based on the texture, shape, colour of patterns on the diseased surface as it includes an unambiguous perception of the defects. An

existing technique used CNN that reduced the relative error as well as decreased the prediction of crop yield. Similarly, the existing model used Back Propagation Neural Network (BPNNs) with the time series model and used smaller dataset size gained lower performance as a smaller number of samples was used for prediction. ML methods were applied in the field of stability of selection and greater precision. ML provides several effective algorithms which are used to find the input and output connection in yield and crop prediction. There are various machine techniques used in agriculture for yield prediction, smart irrigation system, Crop disease prediction, crop selection, weather forecasting, deciding the minimum support price, etc. These techniques will enhance the productivity of the fields along with a reduction in the input efforts of the farmers. Besides, the advances in machines and technologies were accurate as they used significant data and played an important role. This research work analyses the various agricultural methods that utilize ML, along with the merits and limitations.

This research paper is structured as follows: the stepwise process on crop yield analysis is explained in Section 2. The analysis of several ML methods used to examine Crop yield prediction is given in Section 3. The objectives and problem statement of crop yield prediction are shown in 4 and 5 and comparative analysis of several types of research are shown in Section 6. Section 7 describes the conclusion and future work.

Block Diagram:

The steps that are involved in crop yield prediction using machine learning methodology are stated as follows. Firstly, the agriculture Data is utilized for the crop yield prediction, Next, the data is undergone for pre-processing to remove the noisy data. The pre-processed data is undergone for feature extraction process that includes features such as soil information, nutrients, field management etc. which are used to perform the classification using ML algorithms. The results obtained by the existing models using ML algorithms are effectively described in the following section. Figure 1 shows the flow diagram of the crop yield prediction using ML algorithms.



Taxonomy for Analysing Crop Yield Using Various Machine Learning Algorithms

Tseng utilized intelligent agriculture Internet of Things (IoT) equipment to monitor the crop yield prediction. The crops were generally damaged by weather conditions and the existing models used big data in intelligent agriculture to predict the crop yield farm. The developed model utilized an IoT sensor device that monitored the overall agricultural farm and sensed the atmospheric pressure, humidity, moisture content, temperature and soil salinity. The objective of big data analysis in IoT was to analyze and understand crop growing methods practiced by the farmers along with examining environmental deviations. An advantage of the developed model was 3D cluster evaluated the relation between environmental factors and subsequently examined the guidelines obtained from the farmers. However, the developed model showed unusual distribution when it was exposed to potential risk in air humidity, soil moisture content, and temperature.

Tiwari and Shukla developed a model for crop yield Prediction by using CNN and Geographical Index. The existing model faced a problem during a continuous breakdown in agricultural drifts for crop cultivation which were not suitable with environmental factors like temperature, weather and soil condition. The developed CNN model which used spatial

features as input were trained by BPNN for error prediction. An advantage of the developed model was that it was implemented on a real-time dataset that was taken from authentic geospatial resources. However, the developed model reduced the relative error but decreased the efficiency of crop yield prediction.

Fuentes et al. utilized the Robust Deep-Learning method to identify the pest infestation and tomato plant infections in crops. The existing model faced a problem for crop yield prediction due to the presence of pests and diseases in crops which substantially gave rise to economic loss. The developed model introduces a deep meta-architecture to predict the pests in plants. The developed model considers three key features of indicators: Single Shot Multibox Detector (SDD), Faster region-based CNN and Region-Based Fully CNN, which is known as deep meta-architecture. The execution of the deep meta-architecture and feature extractors furthermore suggested a method for a global and local period explanation. The data growth increases the precision and also reduced the number of false positives in training. The benefit of the developed model was it successfully identified different kinds of pests and diseases by dealing with complex situations from a nearby area. Due to the usage of complex pre-processing techniques, the robust deep learning method consumes more time and high computational price.

Sun et al. utilized the Deep CNN-LSTM method to predict the soybean yield estimation. The Yield prediction was an immense consequence for yield mapping, harvest management, crop insurance, crop market planning, and remote sensing. The developed CNN-LSTM approach improved its practicability and feasibility in order to forecast the Particulate Matter (PM2.5) concentration was also verified in the model. The DNN structure was developed that integrated LSTM and CNN based on the historical data such as cumulated wind speed, duration of rain, and concentration of PM 2.5The latest research in this area recommended that CNN could explore more spatial features and LSTM can reveal phonological features, which together play a significant role in crop yield prediction. However, the method employed histogram-based tensor alteration fused different remote sensing data which combined multisource data with a various resolution for feature extraction remained challenging,

Bondre and Mahagonkar utilized ML techniques to predict the crop yield and manure recommendation. The yield prediction was a major issue in agriculture which was overcome by developing a machine learning algorithm. The performance of the developed model was evaluated for estimating crop production in agriculture. An advantage of the developed model was that earlier data was utilized for crop prediction and by applying ML algorithms like random forest and SVM the data also recommended a suitable fertilizer for every particular crop. However, the smart irrigation system for farms to get a higher yield method was not implemented.

Devika and Ananthi utilized data mining techniques to predict the annual yield of major crops. Farmers were opposed to harvesting the yield because of insufficient availability of water sources and unpredictable weather variations but these issues were overcome by developing a data mining method. The developed model was gathering crop growing documents that used to be stored and analyzed for valuable crop yield prediction. In some of the data mining actions, the training data can be collected from the previous documents and the gathered documents were used in the phase of training which has to exploit. An advantage of the developed model was that the highest level of crop yield prediction was obtained only in sugarcane, cotton, and turmeric. However, the range was low for other crops such as wheat, rice, etc.

Pandith et al. utilized the calculation of ML technology for estimation of mustard crop yield from soil review. In agriculture, the soil is a significant factor for determining crop yield calculation and it was overcome by developing an ML technology. Several ML techniques were implemented to forecast mustard crop yield in advance from soil exploration, the techniques named multinomial logistic regression, K-nearest neighbor (KNN), ANN, random forest, Naive Bayes. An advantage of the developed model was that yield prediction was performed even in presence of fertilizer that also is implemented to support the soil analysis and farmers to take judgment accordingly in situations of low crop yield prediction. However, the developed model crop yield prediction with an enormous soil dataset was difficult in a big data environment that showed system complexity.

P.S. Maya Gopal and R. Bhargavi developed a novel approach for an effective CYP. The crop yield was predicted using ANN, statistical and Multi Linear Regression (MLR) algorithms. The model examined the intrinsic behaviour that integrated MLR-ANN model for CYP that analyse the accuracy based on the coefficient generated from MLR and ANNs input layer weights and bias. The Feed forward ANN with back propagation model was used for predicting the crop yield. Similarly, Khaki, S., & Wang, L [18] studied about the DNN for CYP for determining an accurate yield prediction model. The model performed fundamental understanding for setting up the relation among the yield and the interactive factors with respect to the powerful and comprehensive algorithm. The results showed and suggested that the regression trees outperformed better when compared with existing supervised models. However, the main limitation was to look for more advanced models were not showing accurate results.

T. Vijayakumar studied Posed Inverse Problem Rectification Using Novel Deep CNN. The existing methodologies showed an excellent outcome, but imposed challenges in terms of computational cost, parameter selection for adjoint operators and forward operators. The developed model used CNN directly was inverted found a solution for solving the convolution inverse problem. The developed model utilized physical model for analyzing direct inversion, but the combination of multi-resolution decomposition and the combination of residual learning led to artifact generation. Therefore, the model was declined as the noise level was high.

T. Senthil Kumar developed a data mining-based marketing decision support system using hybrid ML techniques that solves the problem respective finance and marketing applications. The decision making is done based on the decision support system which enhanced the organization performance that analyses the ground reality. In the existing models, globalization, privatization, and liberalization dragged the organization more competitively. The competition is balanced and withstand for achieving marketing strategies planned, executed properly. However, an optimization model was required for the model as it posed difficulty during the process and showed lowered assessment performance.

By analyzing the studies, various feature groups related with soil information such as soil maps, soil type, and area of production were discussed. The soil maps will give an information related to type of nutrients present in soil and also location of soil found. The features related to crop information is about the crops such as mustard crops, wheat, rice, tomato plants etc, were analysed in terms of crop density, growth process in terms of weight, leaf area index. Similarly, weather features such as humidity, rainfall, precipitation and forecaster rainfall. Based on these environmental factors, the nutrients components play an important role. The nutrients include,

Nitrogen, potassium, magnesium, zinc, boron etc., The solar information includes features related with the temperature and radiation (gamma), shortwave radiation, solar radiation, degree days are utilized for calculation of features. The features used are less including wind speed, images, and pressure are calculated.

Pseudo Code for CYP Using ML

Learning phase:

Create a training instance data set

Classification phase:

For every unknown instance x^n

Identify x^1, x^2,.. x^n which are the most best instances obtained using ML algorithms from data set are the data points

Set class label until it is equal to the most repeated class

Return class;

End for

Problems Faced in Existing Researches

The problems faced in existing research for crop yield prediction using machine learning are stated below:

1. Creation, repair and maintenance of ML algorithms required huge costs as they are very complex.

2. ML technique used for Crop yield prediction (mustard, wheat) combined input and output data but failed to obtain better results statistically

3. Due to the nature of linear connection in the parameters, the regression model was failed to provide the exact prediction in a complex situation such as extreme value data and nonlinear data.

4. The existing K-NN models were used for classification for yield prediction but lowered the performance due to nonlinear and highly adaptable issues present in KNN. They were operated in a locality model that incremented the dimensionality of the input vector made confusion for classification.

5. An appropriate decision was not taken during classification because a fewer quantity of data was available for estimation of crop yield.

Objectives to be Followed in Future

Objectives to be followed in the future are given below:

1. Depending on the dissimilar crop feature divisions, the modulating factor values of ML algorithms differ to attain perfect approximation.

2. When the quantity of input elements is reduced, ANN is utilized. The optimal feature was being empirically selected for appropriate crop yield estimation.

3. The advantage of ML method regression is to avoid difficulties of using a linear function in large output sample space and optimization of complex problems transformed into simple linear function optimization.

4. ML algorithm can be executed with an enormous soil dataset for crop yield estimation.

5. The ML techniques, through observation of the agricultural fields, provided the necessary support to the farmers in increasing crop production to a great extent.

Result:

```
Mean Squared Error (SVR): 0.33793405801090526
R^2 (SVR): 0.9464327437456398
Accuracy (SVR): 94.64327437456397
```

```
Mean Squared Error (BPNN): 1.404253888786807
R^2 (BPNN): 0.7774061947186237
Accuracy % : 77.74061947186237
```

```
Mean Squared Error (Random Forest): 0.2735781249999999
R^2 (Random Forest): 0.9566340557275542
Accuracy % : 95.66340557275542
```

For Random Forest Algorithm, we have obtained Accuracy of 96 %

For Support Vector Regression, we have obtained Accuracy of 95%

For Back Propagation Neural Networks, we have obtained Accuracy of 78 %

Comparative Analysis

Authors Methodology Advantage Limitation Performance Metrics Kumar et al SVM The SVM method has implemented a cascade of two SVM classifiers for achieving the accuracy, specificity and precision metrics The developed model was not given the proper extensive analysis of the defective outlines such as colour, shapes and texture. Hence, it is failed to identify the infected surface on the defective patterns Accuracy=97.77% Sensitivity=96.55% Precision =99.24% Tiwari and Shukla [7] CNN, Modified Convolutional Neural Network (MCNN) The CNN model was developed which utilized spatial features as input and trained by backpropagation that reduced error of prediction as well. The developed model reduced the relative error as well as decreased the prediction efficiency of crop yield. MCNN RMSE value = 1396.4 Relative Error=9.8465 Shastry and Sanjay [13] (H-ANN) Hybridized ANN, the developed (H-ANN) was used to forecast agricultural data such as air temperature and crop yield estimation. IN H-ANN, the LN algorithm was used to train the ANN The developed model was incapable of capturing the nonlinear bond between input and output variables. RMSE=4.72 Gopal and Bhargavi [17] ANN and Multiple Linear Regression (MLR). The developed model is a combination of backpropagation algorithm with ANN to evaluate the exact crop yield. The developed model showed difficulties in training the neural network model MLR RMSE=9.8%MAE=6.9%R=89% ANN RMSE=5.1%MAE=6.4%R=99% Khaki and Wang [18] Deep Neural Network (DNN) The DNN model was performed for the feature selection. Next, the DNN model has reduced the measurement of input space without affecting the accuracy. The developed model had a black box which was shared through several ML methods Training RMSE=10.55 Validation RMSE=12.79

Conclusion

The present research work discussed about the variety of features that are mainly dependent on the data availability and each of the research will be investigated CYP using ML algorithms that differed from the features. The features were chosen based upon the geological position, scale, and crop features and these choices were mainly dependent upon the data-set availability, but the more features usage was not always giving better results. Therefore, finding the fewer best performing features were tested that also have been utilized for the studies. Most of the exiting models utilized Neural networks, random forests, KNN regression techniques for CYP and a variety of ML techniques were also used for best prediction. From the studies most of the common algorithms used were CNN, LSTM, DNN algorithms but still improvement was still required further in CYP. The present research shows several existing models that consider elements such as temperature, weather condition, performing models for the effective crop yield prediction. Ultimately, the experimental study showed the combination of ML with the agricultural domain field for improving the advancement in crop prediction. However, still more improvement in feature selection was required in terms of temperature variation aspects effects on agriculture. In the further studies, the key possibility that should be concentrated such as firstly the delay to border topographical areas required additional-explicit treatment. Next, a nonparametric portion of the model using machine learning algorithm and thirdly, using features from deterministic crop models to get perfect statistical CO_2 fertilization. By following above-mentioned objectives, the crop yield estimation would be improved by further

researchers. Additionally, in the crop yield estimation, fertilizer should also be considered for executing soil forecasts that agriculturalist to make a better judgment based on the situation of low crop yield estimation. Based on the outcomes obtained for the study further we need to build and develop a model based on DL for CYP.

References:

1.
R. Ghadge, J. Kulkarni, P. More, S. Nene and R. L. Priya, "Prediction of crop yield using machine learning", *Int. Res. J. Eng. Technolgy*, vol. 5, 2018.
2.
F. H. Tseng, H. H. Cho and H. T. Wu, "Applying big data for intelligent agriculture-based crop selection analysis", *IEEE Access*, vol. 7, pp. 116965-116974, 2019.
3.
A. Suresh, N. Manjunathan, P. Rajesh and E. Thangadurai, "Crop Yield Prediction Using Linear Support Vector Machine", *European Journal of Molecular Clinical Medicine*, vol. 7, no. 6, pp. 2189-2195, 2020.
4.
M. Alagurajan and C. Vijayakumaran, "ML Methods for Crop Yield Prediction and Estimation: An Exploration", *International Journal of Engineering and Advanced Technology*, vol. 9, no. 3, 2020.
5.
P. Kumari, S. Rathore, A. Kalamkar and T. Kambale, "Predicition of Crop Yeild Using SVM Approch with the Facility of E-MART System", *Easychair*, 2020.

6.
S. D. Kumar, S. Esakkirajan, S. Bama and B. Keerthiveena, "A microcontroller based machine vision approach for tomato grading and sorting using SVM classifier", *Microprocessors and Microsystems*, vol. 76, pp. 103090, 2020.
Show in Context CrossRef Google Scholar