IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AGRICULTURE EXPENDITURE VISUALIZATION AND CROP YIELD PREDICTION USING MACHINE LEARNING

S. Ramani[1], Dr.K. Merriliance[2]

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

Email[1]:ramanimithra53@gmail.com, Email[2]:merriliance@gmail.com

Abstract: Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research. In this work, we performed a Systematic Literature Review (SLR) to extract and synthesize the algorithms and features that have been used in crop yield prediction studies. Based on our search criteria, we retrieved relevant studies from six electronic databases, of which we have selected five studies for further analysis using inclusion and exclusion criteria. We investigated these selected studies carefully, analyzed the methods and features used, and provided suggestions for further research. According to our analysis, the most used features are temperature, rainfall, and soil type, and the most applied algorithm is machine learning in these models. After this observation based on the analysis of machine learning-based algorithm, to recognize machine learning, we conducted additional researches in databases on crop yields. To find studies that used machine learning, we also searched crop yield datasets. This further study reveals that the Decision Tree Method is the most frequently employed machine learning algorithm in these studies.

***Index Terms* - Component,formatting,style,styling,insert.**

I. INTRODUCTION

From ancient period, agriculture is considered as the main and the foremost culture practiced in India. Ancient people cultivate the crops in their own land and so they have been accommodated to their needs. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economical growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining also useful for predicting the crop yield production. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends. For example, summary information about crop production can help the farmers identify the crop losses and prevent it in future. Crop yield prediction is an important agricultural problem. Each and Every farmer is always tries to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depends on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. This research focuses onevolution of a prediction model which may be used to predict crop yield production. The proposed method use data mining technique to predict the crop yield production based on the association rules.

Machine learning approaches are used in many fields, ranging from supermarkets to evaluate the behavior of customers to the prediction of customers' phone use. Machine learning is also being used in agriculture for several years. Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety. This indicates that crop yield prediction is not a trivial task; instead, it consists of several complicated steps. Nowadays, crop yield prediction models can estimate the actual yield reasonably, but a better performance in yield prediction is still desirable .Machine learning, which is a branch of Artificial Intelligence focusing on learning, is a practical approach that can provide better yield prediction based on several features. Machine learning can determine patterns and correlations and discover knowledge from datasets. The models need to be trained using datasets, where the outcomes are represented based on past experience. The predictive model is built using several features, and as such, parameters of the models are determined using historical data during the training phase. For the testing phase, part of the historical data that has not been used for training is used for the performance evaluation purpose.

An ML model can be descriptive or predictive, depending on the research problem and research questions. While descriptive models are used to gain knowledge from the collected data and explain what has happened, predictive models are used to make predictions in the future. ML studies consist of different challenges when aiming to build a high-performance predictive model. It is crucial to select the right algorithms to solve the problem at hand, and in addition, the algorithms and the underlying platforms need to be capable of handling the volume of data.

II. LITERATURE SURVEY

In[1], In India there are different agriculture crops production and those crops depends on the various kind of factors such as biology, economy and also the geographical factors. And this several factors have the huge different impact on crops, which can be quantified using appropriate statistical methodologies. Applying such methodologies and techniques on historical yield of different crops, it is possible to obtain information or knowledge which can be helpful to farmers and government organizations for making better decisions and for make better policies which help to increased production. In this paper, our focus is on application of data mining techniques which is use to extract knowledge from the agricultural data to estimate better crop yield for major crops in major districts of India.

In[2], Real time, accurate and reliable estimation of maize yield is valuable to policy makers in decision making. The current study was planned for yield estimation of spring maize using remote sensing and crop modeling. In crop modeling, the CERES-Maize model was calibrated and evaluated with the field experiment data and after calibration and evaluation, this model was used to forecast maize yield. A Field survey of 64 farm was also conducted in Faisalabad to collect data on initial field conditions and crop management data. These data were used to forecast maize yield using crop model at farmers' field. While in remote sensing, peak season Landsat 8 images were classified for landcover classification using machine learning algorithm. After classification, time series normalized difference vegetation index (NDVI) and land surface temperature (LST) of the surveyed 64 farms were calculated. Principle component analysis were run to correlate the indicators with maize yield. The selected LSTs and NDVIs were used to develop yield forecasting equations using least absolute shrinkage and selection operator (LASSO) regression. Calibrated and evaluated results of CERES-Maize showed the mean absolute% error (MAPE) of 0.35-6.71% for all recorded variables. In remote sensing all machine learning algorithms showed the accuracy greater the 90%, however support vector machine (SVM-radial basis) showed the higher accuracy of 97%, that was used for classification of maize area. The accuracy of area estimated through SVM-radial basis was 91%, when validated with crop reporting service. Yield forecasting results of crop model were precise with RMSE of 255 kg ha^{-1}, while remote sensing showed the RMSE of 397 kg ha^{-1}. Overall strength of relationship between estimated and actual grain yields were good with R^2 of 0.94 in both techniques. For regional yield forecasting remote sensing could be used due greater advantages of less input dataset and if focus is to assess specific stress, and interaction of plant genetics to soil and environmental conditions than crop model is very useful tool.

In[3], More than 80% of agricultural land in Ireland is grassland, which is a major feed source for the pasture based dairy farming and livestock industry. Many studies have been undertaken globally to estimate grassland biomass by using satellite remote sensing data, but rarely in systems like Ireland's intensively managed, but small-scale pastures, where grass is grazed as well as harvested for winter fodder. Multiple linear regression (MLR), artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) models were developed to estimate the grassland biomass (kg dry matter/ha/day) of two intensively managed grassland farms in Ireland. For the first test site (Moorepark) 12 years (2001-2012) and for second test site (Grange) 6 years (2001- 2005, 2007) of in situ measurements (weekly measured biomass) were used for model development. Five vegetation indices plus two raw spectral bands (RED=red band, NIR=Near Infrared band) derived from an 8-day MODIS product (MOD09Q1) were used as an input for all three models. Model evaluation shows that the ANFIS (RM2moorepark = 0.85, RMSEMoorepark = 11.07; RGrange2 = 0.76, RMSEGrange = 15.35) has produced improved estimation of biomass as compared to the ANN and MLR. The proposed methodology will help to better explore the future inflow of remote sensing data from spaceborne sensors for the retrieval of different biophysical parameters, and with the launch of new members of satellite families (ALOS-2, Radarsat2, Sentinel, TerraSAR-X, TanDEM-X/L) the development of tools to process large volumes of image data will become increasingly important.

In[4], To solve a problem on a computer, we need an algorithm. An algorithm is a sequence of instructions that should be carried out to transform the input to output. For example, one can devise an algorithm for sorting. The input is a set of numbers and the output is their ordered list. For the same task, there may be various algorithms and we may be interested in finding the most efficient one, requiring the least number of instructions or memory or both. For some tasks, however, we do not have an algorithm—for example, to tell spam emails from legitimate emails. We know what the input is: an email document that in the simplest case is a file of characters. We know what the output should be: a yes/no output indicating whether the message is spam or not. We do not know how to transform the input to the output. What can be considered spam changes in time and from individual to individual.

In[5], Agricultural researchers over the world insist on the need for an efficient mechanism to predict and improve the crop growth. The need for an integrated crop growth control with accurate predictive yield management methodology is highly felt among farming community. The complexity of predicting the crop yield is highly due to multi dimensional variable metrics and unavailability of predictive modeling approach, which leads to loss in crop yield. This research paper suggests a crop yield prediction model (CRY) which works on an adaptive cluster approach over dynamically updated historical crop data set to predict the crop yield and improve the decision making in precision agriculture. CRY uses bee hive modeling approach to analyze and classify the crop based on crop growth pattern, yield. CRY classified dataset had been tested using Clementine over existing crop domain knowledge. The results and performance shows comparison of CRY over with other cluster approaches.

III. METHODOLOGY

To get an overview of what has been done on the application of ML in crop yield prediction, we performed a systematic literature review .A Systematic Literature Review shows the potential gaps in research on a particular area of problem and guides both practitioners and researchers who wish to do a new research study on that problem area. By following a methodology in SLR, all relevant studies are accessed from electronic databases, synthesized, and presented to respond to research questions defined in the study. An SLR study leads to new perspectives and helps new researchers in the field to understand the state-of-the-art.

Dataset Collection:

Data is collected from a variety of sources and prepared for data sets. And this data is used for descriptive analysis. Data is available from several online abstract sources such as Kaggle.com and data.gov.in. We will use an annual summary of crops for at least 10 years. The data sets used in this paper are soil dataset, rainfall dataset, and crop yield data.

Preprocessing step:

This step is very important in machine learning. Preprocessing consists of inserting the missing values, the appropriate data range, and extracting the functionality. The kind of the dataset is critical to the analysis process. In this work we have used isnull() method for checking null values and lable Encoder() for converting the categorical data into numerical data.

Feature Selection:

Feature extraction should simplify the amount of data involved to represent a large data set. The soil and crop characteristics extracted from the pre-treatment phase constitute the final set of training. These characteristics include the physical and chemical properties of the soil. Here, we have used Random Forest Classifier() method for feature selection. This method selects the features based on the entropy value i.e., the attribute which is having more entropy value is selected as an important feature for yield prediction.

Split the Dataset into Train and Test Set:

This step includes training and testing of input data. The loaded data is divided into two sets, such as training data and test data, with a division ratio of 80% or 20%, such as 0.8 or 0.2. In a learning set, a classifier is used to form the available input data. In this step, create the classifier's support data and preconceptions to approximate and classify the function. During the test phase, the data is tested.The final data is formed during preprocessing and is processed by the machine learning module.

IV. Table: 1 Crop yield dataset

State_Name	District_Name	CropYear	Season	Crop	Area	Production
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165

Step 1: Initialize the sample weights

In the first step of AdaBoost, each sample is associated with a weight that indicates how important it is with regards to the classification. Initially, all the samples have identical weights (1 divided by the total number of samples).

Step 2: Build a decision tree with each feature, classify the data and evaluate the result

Next, for each feature, we build a decision tree with a depth of 1. Then, we use every decision tree to classify the data. Afterward, we compare the predictions made by each tree with the actual labels in the training set. The feature and corresponding tree that did the best job of classifying the training samples becomes the next tree in the forest.For example, assume that we built a tree that classifies people as attractive if they're smart and unattractive if they're not.The decision tree incorrectly classified 1 person as being attractive based on the fact that they were smart. We repeat the process for all trees and select the one with the smallest number of incorrect predictions.

Step 3: Calculate the significance of the tree in the final classification

Once we have decided on a decision tree. We use the proceeding formula to calculate the amount of says it has in the final classification.

Significance = ½ log (1-totalerror/ Total error)

Where the total error is the sum of the weights of the incorrectly classified samples.

Total error= sum of weights for No option classified samples.

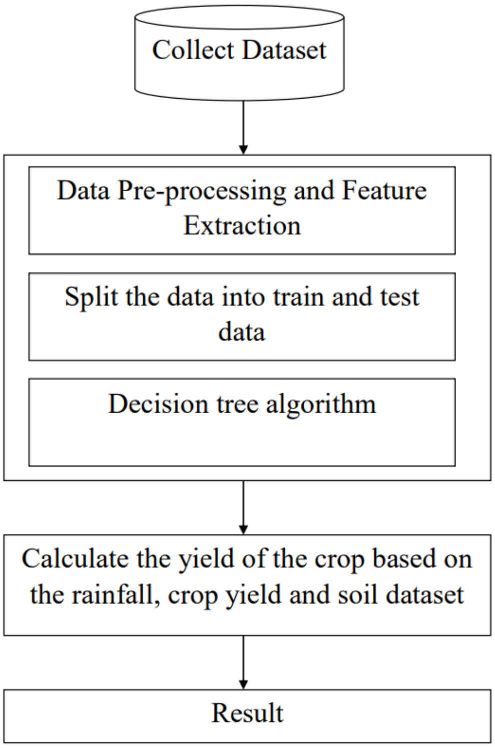
Going back to our example, the total error would be equal to the following.

Total error=1/3

By plugging the error into our formula, we get:

Significance= ½ log (1-1/3/1/3)=0.025

The below figure shows the complete flow of this work.


Collect Dataset
Data Pre-processing and Feature Extraction
Split the data into train and test data
Decision tree algorithm
Calculate the yield of the crop based on the rainfall, crop yield and soil dataset
Result

IV. EXPERIMENT AND ANALYSIS

The implementation of the project was divided into two .i.e. crop yield prediction and rainfall prediction. Crop Yield Prediction This module returns the predicted production of crops based on the user's input. If the user wants to know the production of a particular crop, the system takes the crop as the input as well. Else, it returns a list of crops along with their production as output. These are the following steps of the algorithm implemented:

Step 1 : Choose the functionality i.e., crop prediction or yield prediction.

Step 2 : If the user chooses crop prediction:- ϖ Take soil type and area as inputs. ϖ These values are given as input to the random forest implementation in the backend and the corresponding predictions are returned. ϖ The algorithm returns a list of crops along with their production predicted.

Step 3 : If the user chooses yield prediction:- ϖ Take crop, soil type and area as inputs. ϖ These values are given as input to the random forest implementation in the backend and the corresponding crop yield prediction is returned. ϖ The algorithm returns the predicted production of the given crop.

ACCURACY: It is the ratio of correctly classified points (prediction) to the total number of predictions. Its value ranges between 0 and 1.

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$$

PRECISION: In Information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents and a set of relevant documents.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

RECALL: Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

$$Recall = \frac{TP}{TP + FN}$$

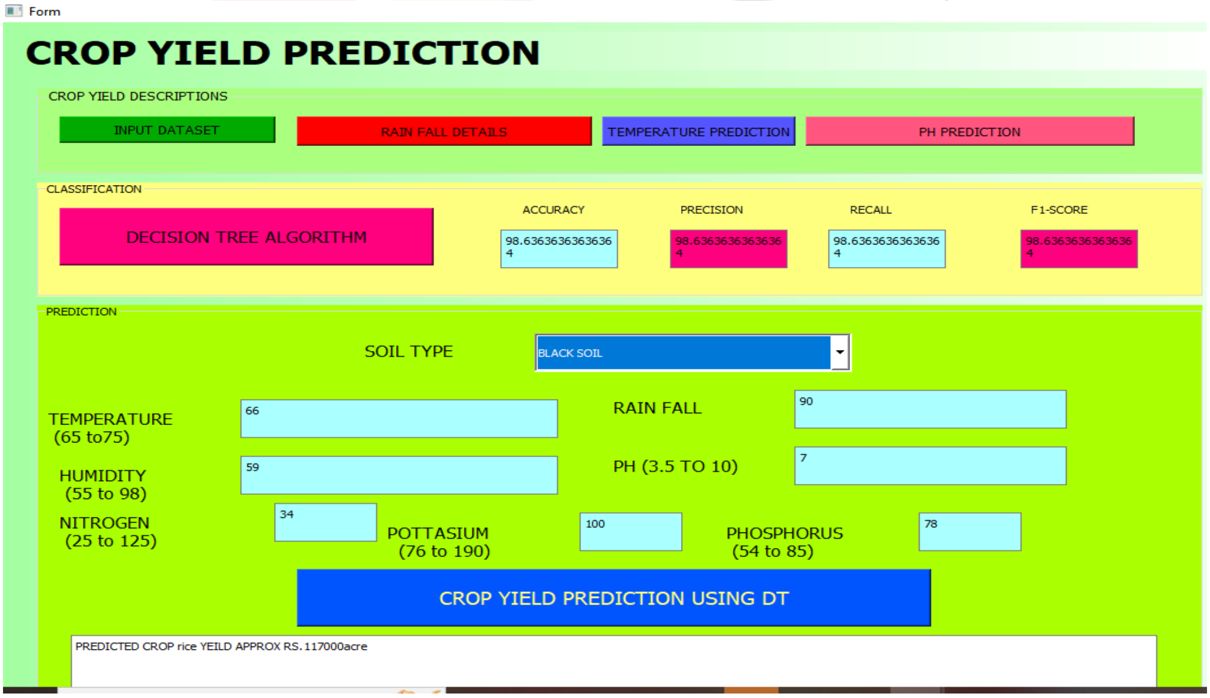
IV.

F1-SCORE FORUMLA: In the F1 score, we compute the average of precision and recall. They are both rates, which makes it a logical choice to use the harmonic mean.

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

V.

The below output shows the main options of this work. The input dataset reads the input data from the crop yield dataset.



ACCURACY	PRECISION	RECALL	F1-SCORE
98.6363636363636 4	98.6363636363636 4	98.6363636363636 4	98.6363636363636 4

SOIL TYPE: BLACK SOIL

TEMPERATURE (65 to75): 66

RAIN FALL: 90

HUMIDITY (55 to 98): 59

PH (3.5 TO 10): 7

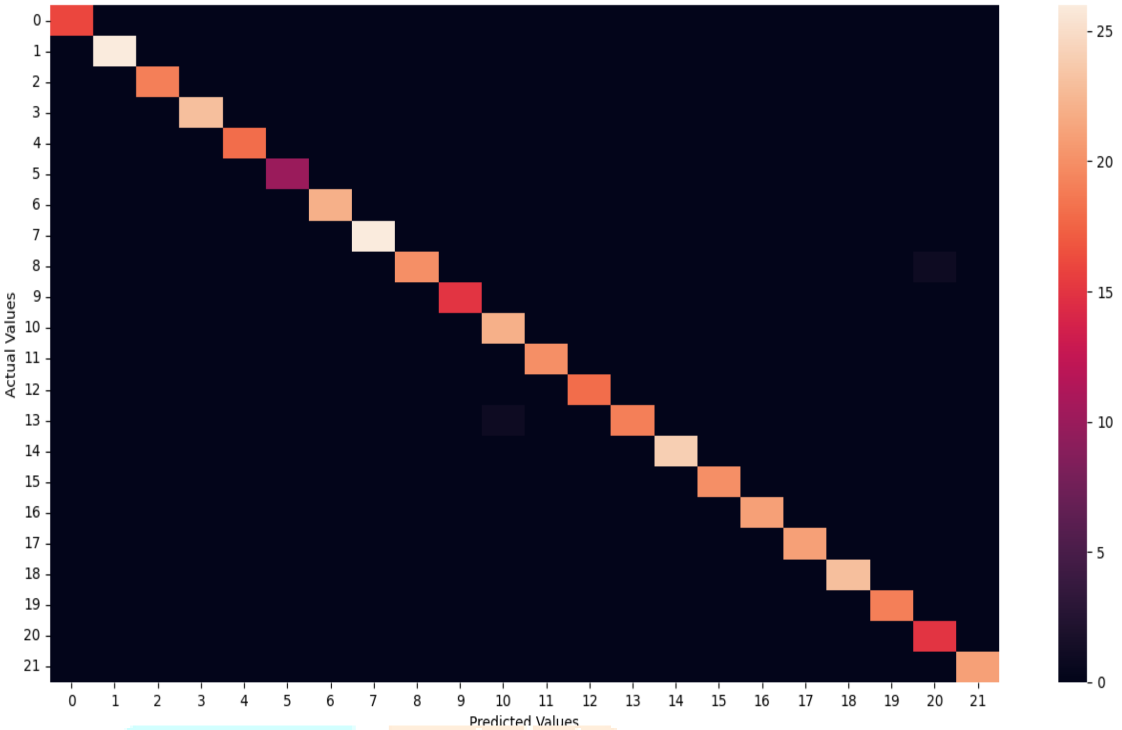
NITROGEN (25 to 125): 34

POTTASIUM (76 to 190): 100

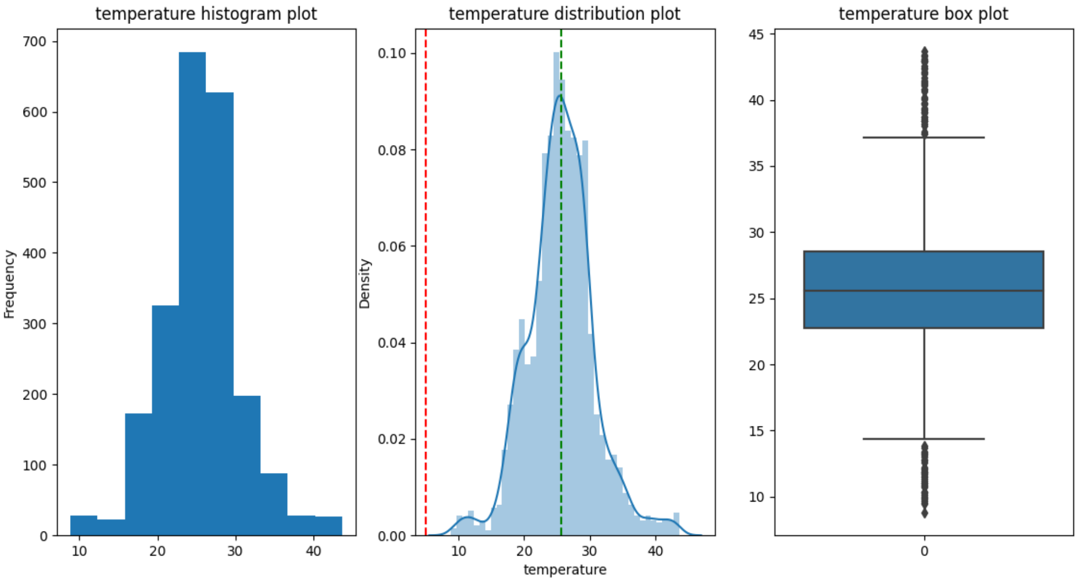
PHOSPHORUS (54 to 85): 78

PREDICTED CROP rice YEILD APPROX RS.117000acre

The above figures shows the rainfall details, temperature details and ph prediction as chart. The decision tree algorithm was tested and predicted the accuracy as 98.63% and precision, recall and fi score. Finally, this system is used to predict the crop and crop yield by using decision tree algorithm.



The above figure illustrates the correlation feature of the dataset.



The above figure illustrates the temperature histogram chart and marked maximum histogram value.

VI. CONCLISION

This study showed that the selected publications use a variety of features, depending on the scope of the research and the availability of data. Every paper investigates yield prediction with machine learning but differs from the features. The studies also differ in scale, geological position, and crop. The choice of features is dependent on the availability of the dataset and the aim of the research. Studies also stated that models with more features did not always provide the best performance for the yield prediction. To find the best performing model, models with more and fewer features should be tested. Many algorithms have been used in different studies. The results show that no specific conclusion can be drawn as to what the best model is, but they clearly show that some machine learning models are used more than the others.

The most used models are the random forest, neural networks, linear regression, and gradient boosting tree. Most of the studies used a variety of machine learning models to test which model had the best prediction. Since Neural Networks is the most applied algorithm, we also aimed to investigate to what extent deep learning algorithms were used for crop yield prediction. After the identification of 30 papers that applied deep learning, we extracted and synthesized the applied algorithms. We observed that other algorithms are the most preferred deep learning algorithms. However, there are also other kinds of algorithms applied to this problem. We consider that this article will pave the way for further research on the development of crop yield prediction problem.

In our future work, we aim to build on the outcomes of this study and focus on the development of a ML-based crop yield prediction model.

REFERENCES

1. Wang, Yi-Hsin, et al. “Applying Data Mining Techniques to WIFLY in Customer Relationship Management.” *Information Technology Journal*, vol. 9, no. 3, Mar. 2010, pp. 488–93. *DOI.org (Crossref)*, https://doi.org/10.3923/itj.2010.488.493.
2. Ahmad, Ishfaq, et al. “Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan.” *Journal of the Indian Society of Remote Sensing*, vol. 46, no. 10, Oct. 2018, pp. 1701–11. *DOI.org (Crossref)*, https://doi.org/10.1007/s12524-018-0825-8.
3. Ali, Iftikhar, et al. “Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 7, July 2017, pp. 3254–64. *DOI.org (Crossref)*, https://doi.org/10.1109/JSTARS.2016.2561618.
4. Alpaydin, E., 2010. Introduction to Machine Learning, 2nd ed. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=TtrxCwAAQBAJ&oi=fnd&pg=PR7&dq=introduction+to+machine+learning&ots=T5ejQG_7pZ&sig=0xC_H0agN7mPhYW7oQsWiMVwRnQ#v=onepage&q=introduction to machine learning&f=false.
5. Ananthara, M. G., et al. “CRY &#x2014; An Improved Crop Yield Prediction Model Using Bee Hive Clustering Approach for Agricultural Data Sets.” *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, IEEE, 2013, pp. 473–78. *DOI.org (Crossref)*, https://doi.org/10.1109/ICPRIME.2013.6496717.