A review on Estimation of Crop yield using different Machine learning approaches in precision agriculture

ANU C.S1 Dr. NIRMALA C.R2

[1]*Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, affiliated to Visvesvaraya Technological University, Belagavi-590018, India*

[2]*Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, affiliated to Visvesvaraya Technological University, Belagavi-590018, India*

anucsebiet@gmail.com, nirmala.cr@gmail.com

***Abstract*:**

Precision agriculture is an innovative approach of farming that makes use of data analytics, machine learning (ML) strategies to maximize utilization of resources, boost crop yield prediction, and reduce environmental impact. ML techniques are currently receiving more attention since remote sensing approaches uses large amount of data from different platforms. ML based systems have the ability to perform nonlinear tasks and process a huge number of inputs. In this review work, recent advances in ML based algorithms for precise crop yield prediction are discussed. The research comes to the conclusion that the rapid development in ML approaches will offer complete, cost-effective solutions for improved crop and environment status estimates and decision making. The goal of precision agriculture is to make farming practices more efficient, effective, and sustainable. Precision agriculture leverages ML and interpretable artificial intelligence to revolutionize conventional farming techniques, evolving them into data-centric, effective, and eco-friendly approaches.

Keywords: *Crop Yield Prediction, machine Learning, Deep Learning, Deep Leraning, , CNN-Recurrent Neural Networks (RNN), and CNN-Long Short-Term Memory (LSTM), Deep Neural Networks (DNN).*

1. Introduction

The most fundamental requirements of humanity are met by agriculture: food and fibre. Although the government is adopting financial measures to assist farmers, they are still facing difficulties since there is a lack of data analysis and crop forecast in our country, where a major portion of the people relies on agriculture. Increasing food demand, population expansion, and income levels will put more strain on natural resources. Over the past century, new farming techniques have been created, allowing agriculture compete with the rising demand for food and other agricultural products.

A critical use of agricultural data science and technology in Crop yield prediction (CYP) seeks to predict the crop's potential output before it is actually harvested. It is essential in modern agriculture for assisting farmers, decision-makers, and agribusinesses in making wise choices about the distribution of resources, crop management, and overall agricultural planning. CYP algorithms predict the production of different crops for a certain place and season using historical data, weather patterns, and advanced analytics.
Predicting crop yield is growing significantly in today's environment, where ensuring food security is crucial by the obstacles posed by climate change, growing global population, soil erosion, and dwindling water resources. The growth or yield of crops are influenced by factors such as temperature, radiation, water availability, and other environmental conditions in a complicated and nonlinear way.

Models for predicting agricultural crop yields can be continually updated and improved with new data and advancements in ML techniques. Additionally, the incorporation of data acquired through remote sensing methods, particularly satellite imagery, and IoT devices in agriculture further enhances the accuracy and granularity of these predictions. Ultimately, Precise forecasts of crop yields contribute to sustainable and efficient agricultural practices, reducing waste and ensuring food security for a growing global population

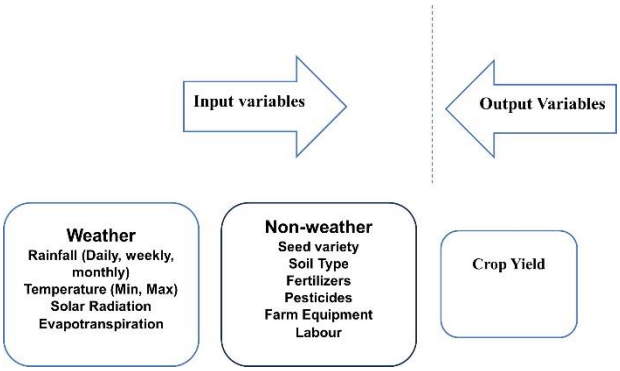

Figure 1: Variables and Framework for yield prediction

Previous works used various methods for forecasting agricultural crop production in precision agriculture. Multisensory Machine-Learning Approach (MMLA) are used for classifying multisensory data. Some of the ML algorithms are J48 Decision Tree, Hoeffding Tree, and Random Forest. Random Forest, SVM, Gradient Descent, long short-term memory, and Lasso regression technique, Linear Regression, Decision Tree Regression, GBR, RFR, Xgboost Regression, And Voting Regression, to integrate the outputs of multiple deep neural networks, including the 3DCNN (3D Convolutional Neural Network) and ConvLSTM (Convolutional Long Short-Term Memory), XGBoost ML (ML) algorithm, Convolutional Neural. Networks (CNN)-Deep Neural Networks (DNN), CNN-XGBoost, CNN-Recurrent Neural Networks (RNN), and CNN-Long Short-Term Memory (LSTM).

Depending on the area of the study and the accessibility of data, a wide variety of criteria are used in soil monitoring and CYP research. The outcomes of this work show variances in terms of scope, region, and the particular crops being researched. The dataset's accessibility and the study goals both influence the choice of attributes. It was interesting to note that models with maximum features did not consistently produce better yield forecasts. This includes the utilization of soil and meteorological data.

To identify the most effective model, researchers typically test models with varying feature set. various algorithms were used in numerous research, and the best model cannot be determined with certainty and it is clear that some ML models are preferred more. The frequently used models include the Deep Learning Multi-Layer Perceptron (DLMLP), RM, NN, LR, and. Most studies employ different ML models to identify which one yields the most accurate predictions.

Furthermore, among deep learning algorithms, Neural Networks hold the highest prevalence. According to surveys, CNN, LSTM, DNN, SVM, ADA Boost, and Bayesian/Naive Bayes Classifier algorithms are the most chosen DL methods in this context.

The success of CYP using ML can have significant implications for agriculture, allowing farmers and stakeholders to make better-informed decisions, optimize resource allocation, and ultimately increase agricultural productivity and food security. ML and AI are being used more frequently in agricultural research to make predictions. Many models, however, are frequently "black boxes," which means that we are unable to explain what the models discovered from the data or the causes of forecasts.

To get a general understanding of the work that has been done on the usage of ML in CYP, thorough examination of the literature is been carried out. An SLR identifies possible research holes in a specific area of study and offers advice to academics and practitioners who want to do new research in that area. An SLR study helps new investigators to grasp the cutting-edge novel perspectives.

2. Research questions

A research review work has been done in the area of Machine Learning (ML) and crop yield prediction. To achieve this, many studies from different viewpoints has been carried out. In this study, following five research questions are framed (Qs).

Q1: Which machine learning methods have been applied to crop yield prediction in the literature? (Preprocessing)

Q2: Which features have been applied to machine learning-based crop yield prediction in the literature? (Feature Extraction)

Q3: Which assessment criteria and methods have been applied for agricultural yield prediction in the literature? (Classification methods)

Q4: Which Evaluation Parameters have been used for yield prediction in precision agriculture? (Evaluation Parameters)

Q5: What challenges exist in the field of machine learning-based crop yield prediction?

3. Literature review

A paper titled "CYP via explainable AI and interpretable ML: Dangers of black box models for evaluating climate change impacts on crop yield" by Tongxi Hu was proposed in 2023 [1]. In this study, a Bayesian ensemble model (BM) was employed to dissect historical yield data in order to evaluate the combined effects of technical changes and climatic change on agricultural output. ElasticNet, Neural Network, MARS, SVM, Random Forests, and XGBoost were contrasted with BM. BM was great at both explaining and forecasting. BM was the only technique that, when tested on artificial data, revealed the real relationships: Other approaches accurately predicted, but for the wrong reasons, whereas BM has a stronger interpretability.

Estimating a crop output in the Indian Wheat Belt with explainable DL presented by Aleksandra Wolanin in 2023 [2]: In this study, they adopted a multivariate time series of vegetation and climatic data along with a DNN to estimate the wheat output. The characteristics and yield drivers that the model had learned were then visualised and examined using regression activation maps. In comparison to the other two models (random forest and ridge regression), the DL model performed better, and it also made it simpler to comprehend the variables and processes that affect yield variability. The primary teaching points were the length of the growing season, together with the temperature and illumination patterns during this time.

"Accelerating Crop Yield: Multisensor Data Fusion and ML for Agriculture Text Classification" is a proposal made by A. REYANA in [3] 2023. Multisensory Machine-Learning Approach (MMLA) is a technique used to categorise multisensor data. J48 Decision Tree, Hoeffding Tree, and Random RF are ML methods. With RMSE 13%, RAE 38.67%, and RRSE 44.21%, the RF method has the lowest error measure for identifying the agriculture text. Consequently, a multisensor data fusion technique based on crop suggestion offers increased prediction precision, leading to a noticeably higher crop output. To improve the forecast process, future research in the agriculture sector should take these characteristics into account

Kavita Jhajharia presented a work on “Crop Yield Prediction using Machine Learning and Deep Learning Techniques”[4]. In this study they used different technologies like Random Forest, SVM, Gradient Descent, long short-term memory, and Lasso regression technique. Random forest performed better than others with 0.963 R2, 0.035 RMSE, and 0.0251 MAE.

Bharati Panigrahi in 2023 [5] proposed” A ML-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning with Regression Models” .In this work they adopted different methods like Linear Regression, DTR, GBR, RFR, Xgboost Regression, and Voting Regression. The XGBoost Regression model was overfitting with a high R2 score and lower Cross Validation score. Furthermore, with an MAE score of 468.16, MSE score of 825.29, R2 score of 0.7952 and a Cross-Validation score of 0.6087. Future research involves using a variety of artificial intelligence (AI), deep learning (DL), and computer vision (CV) techniques on pictures of the field and crop to identify any diseases present in weeds, which can affect the crop's quality and can be distinguished from the healthy crops.

Iniyan in 2022, [6] demonstrated, “CYP using ML techniques”, methods used are Linear regression, DTR, GBR, Elastic Net, Lasso, Ridge, LSR and feature engineering-based LSTM models. Compared to other providers, this one had the best accuracy (86.3%), the lowest mean absolute error, and the lowest root mean square error. Accuracy can be increased along with data growth. The web application will be hosted on a Google Cloud platform in the future, and data will be stored in cloud buckets. Additionally, we intend to take into account factors like plant genotype, heritability, and other compositional traits.

Abbaszadeh in 2022 [7],proposed “Bayesian Multi-modelling of Deep Neural Nets for Probabilistic CYP Peyman” .In this study Bayesian Model Averaging (BMA) and a set of Copula functions to integrate the outputs of multiple DNN, including the 3DCNN (3D Convolutional Neural Network) and ConvLSTM (Convolutional Long Short-Term Memory) are used. The findings of this study demonstrate that, when uncertainties in the models considered, the suggested approach outperforms the 3DCNN and ConvLSTM networks in predicting soybean crop yield. The future work is to create a method for predicting probabilistic yields for various crops

Lontsi Saadio Cedric in-2022 [8] proposed “Crops yield prediction based on ML models: Case of West African countries”. Methods used are DT, multivariate logistic regression, and k-nearest neighbour models CMLR model gives a lowest performance, while the Ck-NN gives the highest performance. The future work is to add others features such as soil data, wind data, humidity, agricultural water data, wind data, pollution data, meteorological variations data, animal species data and agricultural economic data of those countries can probably improve the model quality.

Alexandros Oikonomidisa in 2022 [9] presented work on “Hybrid DNN for Crop Yield Prediction”: In this work they employed XGBoost ML (ML) algorithm, CNN, DNN, CNN-XGBoost, CNN-Recurrent Neural Networks (RNN), and (LSTM). The outcomes indicates that the hybrid CNN-DNN performs better than other models. Comparatively speaking, the other DL-based models take longer to execute than the XGBoost model. The usage of the soybean crop dataset across the Corn Belt in US is one of the work's limitations and results may vary slightly depending on the dataset.

Swati Vashisht- Taylors & Francis in 2022 [10] proposed a wok on “CYPUsing Improved Extreme Learning Machine”, Methods used are Kalman filter algorithm, Linear Discriminant Analysis (LDA), achieved better results in terms of accuracy, precision, and recall compared to SVM, NB, and Bayes Net. The proposed approach attained a 99.99% accuracy rate, a 99.24% precision rate, a 99.24% recall rate, and an F-measure of 0.9923.

Amna Ikram in 2022 [11] presented a work on “Crop Yield Maximization Using an IoT-Based Smart Decision”. In this study Methods used are Decision tree, SVM, KNN, Random Forest, and Gaussian Naïve Bayes. The SCS model is trained for 11 crops’ prediction, while its accuracy is 97% to 98%. In future, more parameters and crops can be added to this system. The more accurate and efficient ML algorithms like CNN and LSTM can also be studied. SCS model can be integrated with security to protect crop data.

Pallavi Kamath in 2021 [12] demonstrated “Crop yield forecasting using data mining”, A Yield prediction was done by taking parameters like Precipitation, Temperature, and other parameters such as season and location. The dataset having more features gives more accurate result. When compared to decision trees and MLR, which are alternative technologies, RM is the best classifier and prediction method. With the help of this model, accuracy of 98% has been attained.

Anjana, Aishwarya Kedlaya in 2021[13] presented a work on “An efficient algorithm for predicting crop using historical data and pattern matching technique”: In his study, a crop forecast system that makes use of historical data is presented. Datasets are processed using Xarray functions, and a pattern-matching technique is utilised to get the crop based on geography and season.

Zheng Chu in 2020 [14] demonstrated “An end-to-end model for rice yield prediction using deep learning fusion “which used BPNNs with an IndRNN, named BBI-model, The outcomes indicates that BBI-model achieved the least MAE and RMSE for the summer rice prediction (0.0044 and 0.0057, respectively) and corresponding values of 0.0074 and 0.0192. The time-series data set wasn't very large, the future study is to add more samples to boost forecast accuracy.

Akshar Tripathi in 2022 [15] has deliberated “A deep learning multi-layer perceptron and RS approach for soil health-based crop yield estimation. It used DL and RS to predict three crucial soil health indices including soil moisture, soil salinity, and soil organic carbon (SOC), using optical and microwave satellite data from Sentinel-1 and Sentinel-2 and field data. Wheat crop yield was evaluated using estimated soil health metrics, SAR backscatter, and optical RS satellite data characteristics.

Nihar Patel in 2021 [16] proposed “Crop Yield Estimation using ML”, where Crop Data from six states of India , only for five crops from 2009–2016 was used for training and validation.(used Traditional approach, also old data set) Linear models, Support Vector Regressor, K Neighbors Regressor, Tree-based models, Ensemble models and Shallow Neural Networks with R-squared score for evaluation. The study's findings offer insightful information for determining how vulnerable agriculture is to climate change.

E. Kanimozhi, D. Akila in 2020[17] illustrated an “Empirical Study on Neuro evolutional Algorithm Based on ML for Crop Yield Prediction, ML method for predicting crop yield”. Here, they demonstrated a technique for using an ANN-based neuroevolutionary model to estimate wheat crop productivity. From June to September, CYP is taken into account. The yield estimates are generated using meteorological and fertiliser usage data, but are only applicable to a particular wheat crop.

Ms. Kavith [18] in 2020 has proposed a paper on “Crop Yield Estimation in India Using ML to predict crop yield” using area, yield, production, and area under irrigation. Four ML techniques DT, LR have been applied to estimate the crop yield. MAE, MSE, RMS were employed as cross validation methods for validation.

Gunkirat Kaur [19], 2020 has proposed a work on "Soil Nutrients Prediction Using RS Data in Western India": our nutrients—N, K, P, and OC—were estimated for two districts of Maharashtra, India, using an evaluation model that used optical remote sensing data (Landsat-8 and Sentinel-2), terrain/climate data (precipitation, radiation, slope, etc.), and ground truth values. compared the estimates of NPK and OC using four linear and non-linear regression models: MLR, RFR, SVR, and GB. Comparative results indicate that, for all nutrients, GB and RFR outperformed other models with sMAPE in the range of 0.125-0.377, which is better or comparable with accuracy reported in the literature. As a result, the method can decrease the time and expense associated with soil sampling while producing a high resolution (\; ha) map of soil nutrients.

Mayank Champaneri [20] 2020 illustated the idea of "CYPusing ML", they developed model by using Random forest, the most popular and powerful supervised ML algorithm capable of performing both classification and regression tasks, that operate by constructing a multitude of decision trees during training time and generating output of the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Thomas van Klompenburga, Ayalew Kassahuna in 2020 [21] proposed a work on "CYP using ML: Based on the analysis it isobserved that the utilized factors in these models are temperature, rainfall and soil type. Moreover. ANN have emerged as the algorithm among researchers in this field. To further explore learning based studies an additional search was conducted in databases resulting in 30 relevant papers. From this analysis it was revealed that CNN were extensively used as a deep learning algorithm, in these studies LSTM and DNN were also identified .

Mohamad M. Awad in 2019 [22] has demonstrated an idea of "Innovative intelligent system based on remote sensing and mathematical models" for improving crop yield estimation. A new mathematical model of intelligent system is implemented that includes the use of energy balance equation several farmers are interviewed and information about their crops yield is collected.(Traditional Method)

Pushpa Mohan in 2018 [23] "Crop production rate estimation using parallel layer regression with deep belief network Parallel Layer Regression (PLR) along with Deep Belief Network (DBN) strategy" is proposed to perform crop productivity estimation. Here, DBN strategy is generated for top five growing crops in Karnataka namely, rice, ragi, and pulses. The proposed methodology forecasts each area in the applicable database. Finally, outcomes shows that the method has strong potential for accurate crop productivity prediction in terms of accuracy (ACC), sensitivity (SEN) and specificity (SPE) and also this method performance has verified in real time data and people interactions.

Jayantrao Mohite in 2018 [24] illustrated an idea of "Spatialization of rice crop yield using Sentinel-1 SAR and Oryza Crop Growth Simulation Model spatial estimation of rice yield "by assimilation of parameters derived from Synthetic Aperture RADAR (SAR) data from Sentinel-1 satellite into a process-based Oryza crop growth simulation model. The study has been carried out in four districts of coastal Andhra Pradesh, India viz., Guntur, Krishna, East Godavari and West Godavari during monsoon season locally called Kharif (mid-Jun. to midDec.) In the study area, rice is transplanted during mid-Jun to Aug. end and harvested

from Oct. to mid-Dec. months. The methodology for in-season regional rice area estimation using random forest classifier has been described in our previous work.

Yao Chunjing a, Zhang Yueyao, in 2017 [25] "Application of Convolutional Neural Network In Classification Of High Resolution Agricultural RS Images", proposed a classification method for high-resolution agricultural RS images based on CNN. For training, a vast amount of training samples were produced by panchromatic images of GF-1 high-resolution satellite of China. In the experiment, through training and testing on the CNN under the toolbox of DL by MATLAB, the crop classification finally got the correct rate of 99.66% after the gradual optimization of adjusting parameter during training. Through improving the accuracy of image classification and image recognition, the applications of CNN provide a reference value for the field of RS in PA

Miss.Snehal S.Dahikar in 2014 , Dr.Sandeep V.Rode[26] has illustrated "Agricultural CYPUsing Artificial Neural Network Approach" by considering various situations of climatologically phenomena affecting local weather conditions in various parts of the world. These weather conditions have a direct effect on crop yield. Various researches have been done exploring the connections between large-scale climatologically phenomena and crop yield. Artificial neural networks have been demonstrated to be powerful tools for modeling and prediction, to increase their effectiveness. Crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity.

Shunping Ji, in 2018 [27] proposed a "3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images, novel three-dimensional (3D) convolutional neural networks (CNN)" .This method automatically classifies crops from spatio-temporal RS images. First, 3D kernel is designed according to the structure of multi-spectral multi-temporal remote sensing data. Secondly, the 3D CNN framework with fine-tuned parameters is designed for training 3D crop samples and learning spatio-temporal discriminative representations, with the full crop growth cycles being preserved.

Saeed Khaki [28]in 2020 presented a work on "A CNN-RNN Framework for Crop Yield Prediction" presents a deep learning framework using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for crop yield prediction" based on environmental data and management practices. The proposed CNN-RNN model, along with other popular methods such as random forest (RF), deep fully connected neural networks (DFNN), and LASSO, was used to forecast corn and soybean yield across the entire Corn Belt (including 13 states) in the United States for years 2016, 2017, and 2018 using historical data.

F. H. Tseng, H. H. Cho and H. T. Wu, ha sillustrated "Applying big data for intelligent agriculture-based crop selection analysis". This research study leverages Intelligent Agriculture IoT technology to monitor factors present, on a farm. The data collected from this monitoring is then subjected to 3D cluster analysis, which helps us analyze and understand the characteristics of the farms environment. Our proposed approach includes features of achieving data normalization through a combination of moving average and average variance, utilizing 3D cluster analysis to examine the relationship between environmental factors and

investigate the knowledge held by farmers, determining if a selected crop belongs to an appropriate cluster; and establishing a critical value within each cluster based on future environmental conditions thereby providing guidance on whether a particular crop is suitable for the farm. To carry out our research we installed Intelligent Agriculture IoT equipment on the farm, for monitoring. Conducted practical analyses using our algorithm. The results obtained confirm that our proposed approach is indeed viable.

4. Publication Details

The Below graph displays the total number of publications released annually over the previous seven years.

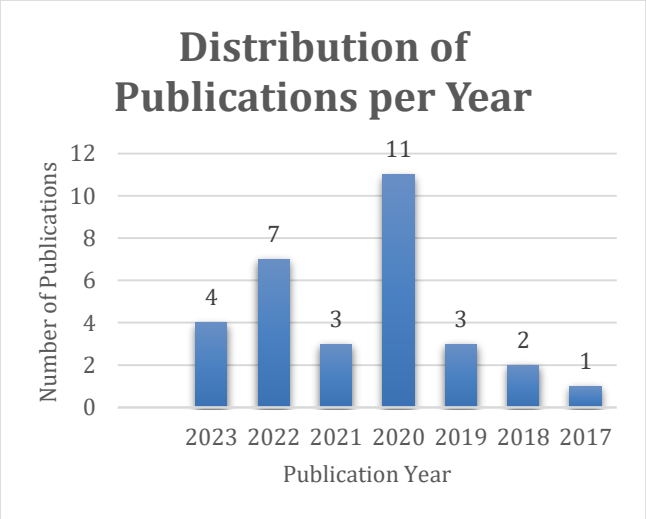

Figure 2: Publications released annually over the previous seven years

5. Q1: Preprocessing

Data preprocessing is a crucial step in the ML pipeline that involves cleaning, transforming, and integrating to remove inconsistencies and organizing raw data into a suitable format for training and building models. Proper data preprocessing can significantly impact the performance and reliability of ML algorithms. The preprocessing phase aims to improve accuracy of the results. some research, that uses a categorical variable takes one value from a limited set of values and is not quantifiable. As a result, to apply any algorithm, such variables must be encoded, because many ML algorithms cannot work directly on labelled data, so all input must be in the form of numeric values. There are several methods for encoding and handling such variables. Label Encoder, One Hot Encoder, and other methods fall into this category [4].

Each dataset underwent extensive pre-processing, which included performing crucial tasks like resolving missing values and encoding categorical data. Ultimately, the dataset will be divided into training and test sets [4]. By setting the parameters for numerical and categorical variables individually, the procedures for resolving missing values and encoding were done using a common pipeline.

The missing values in numerical data were replaced by the median while missing values in categorical data were replaced by the most frequently occurring value using an Imputer [5]. The encoder was used to encode the categorical data. It is critical in ML to distinguish between independent and dependent variables in a dataset. Crop, District, Year, Total Rainfall, and Maximum and Minimum Temperatures for all months of 2016, 2017, and 2018 are the independent variables in the dataset, whereas total yield is the dependent variable [5]. During the data pre-processing part, null, extreme outliers, and repeated values will be removed. They are removed using the python library. For analysing the model performance, the metric R-squared (Coefficient of Determination). For comparing the efficiency of these regression models, Mean Absolute Error, Root Mean Square Error were used. Decision rules are applied on dataset. In decision rules, standard ranges of parameters are defined for each crop. ML algorithms are used for dataset training.[11] Normalization is related to robust scaling, was also used but, use of the interquartile range instead of normalizing the data is because the data set contains numeric data. Normalization reduces the size of the data by a factor of 0 to 1[12].

6. Q2: Feature Extraction

To visualise the key characteristics and algorithms, groups are made for both features and algorithms. Clarity has been preserved despite this decision's loss of specific details. Soil type, precipitation, and temperature are the three most often used elements. There are features that were employed in certain studies in addition to those that were used in several investigations. GR, MODIS-EVI, predicted rainfall, humidity, photoperiod, pH value, irrigation, leaf area, NDVI, EVI, and crop information are some of these features. Studies have also used the nutrients magnesium, potassium, sulphur, zinc, nitrogen, boron, and calcium as characteristics. Not usually the same type of data is used as the most popular characteristics. For instance, average temperature is used to quantify temperature, but additional parameters like maximum temperature and minimum temperature are also measured. [21].

Once pre-processed, the vector components represent the associated words and their weights. Precision, recall, F-measure, and MCC are all used. In terms of precision, recall, F-measure, and MCC, the random forest algorithm outperforms the other two algorithms, the J48 decision tree, and the Hoeffding tree[3]. The StandarScaler method from the SciKit Learn library has been applied to the given dataset. This helps standardize features by removing mean and scaling to unit variance. Standardization of any dataset is common practice and a necessity at times for a lot of ML and deep learning methodologies [4]. CYPwas based on geological position, size, and crop features. Extra features don't necessarily boost accuracy.

The investigation assessed the best-performing traits. Combining ML with agriculture improved crop yield predictions. Climate change effects on agriculture feature selection could be improved. Fertilizer data should be included in crop yield predictions so farmers can make better decisions in case of low estimates [5]. As deep learning models require more data to train and deliver better results, features to create ones like Precipitation area ratio, precipitation humidity ratio, precipitation to mean temperature ratio are used. Along with adding these features, we reframed and normalised the dataset before training the model [6].

For the case study, we performed experiments on a public soybean dataset that consists of 395 features including weather and soil parameters and 25,345 samples. The results showed that the hybrid CNN-DNN model outperforms other models, having an RMSE equal to 0.266, an MSE of 0.071, and an MAE of 0.199. The predictions of the model fit with an R2 of 0.87. The second-best result was achieved by the XGBoost model, which required less time to execute compared to the other DL-based models[9].Major soil nutrient parameters such as NPK (nitrogen, phosphorus, and potassium) and other factors (T, H, pH, EC, CO2, soil type, and rainfall were used in the research work[11]. Area, Temperature, Precipitation, and Humidity. Some soil agronomical parameters, such as chalky, clay, loamy, sandy, and so on, as well as different seasons, are included. The data of these variables were given as input.

Initially dataset is collected which consisting of the parameters such as attributes like State Name, District name humidity, temperature, yield etc. Take into consideration any crops that will be planted in the region. This collected dataset is in csv format [12]. The dataset [12] consisting of the soil specific attributes (e.g. soil moisture) and weather specific attributes (e.g. rainfall, temperature, humidity etc.)[13]. Sentinel-1 and Sentinel-2 optical and microwave satellite data, along with field data, are used to estimate three key soil health parameters: soil moisture, soil salinity, and soil organic carbon (SOC). The estimated soil health parameters, SAR backscatter, and optical remote sensing satellite data parameters were utilized to estimate wheat crop yield [15]. Predictions can be made taking into account forecasts of climate, soil and its mineral content, moisture, crop historic performance, rainfall and others [16].

Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity [26]. The below diagram gives the information of the soil parameters. It includes crop information, soil information, humidity, nutrients, field management, solar information and others. Variable of these parameters are growth process, leaf area index, soil map (soil type, Ph, soil, CEC), crop map, rainfall, precipitation, Nutrient in soil, added nutrients, fertilization, gamma radiation, temperature, windspeed, pressure, Nitrogen, Phosphorus, Potassium, Wind speed, Pressure, Humidity [26] etc. Some of the features used in previous work are shown in table1 [21].

Table 1: Features used for estimating crop yields

Features	**#Number of times used**
Temperature	32
Soil type	32
Rainfall	29
Nitrogen	30
Phosphorus	30
Potassium	30
Crop information	25
Soil maps	12
Humidity	32
pH-value	32
Solar radiation	15
Precipitation	16
Images	14
Area of production	15
Fertilization	14
NDVI	10
Irrigation	11
Wind speed	25
Zinc	5
Magnesium	5
Shortwave radiation	4
Sulphur	4
Boron	4
Calcium	4
Organic carbon	2
EVI	2
Phosphorus	2
Gamma Radia metrics	3
MODIS-EVI	3
Forecasted rainfall	3
Photoperiod	3
Climate	3
Degree-days	3
Time	3
Pressure	3
Leaf area index	3
Manganese	3

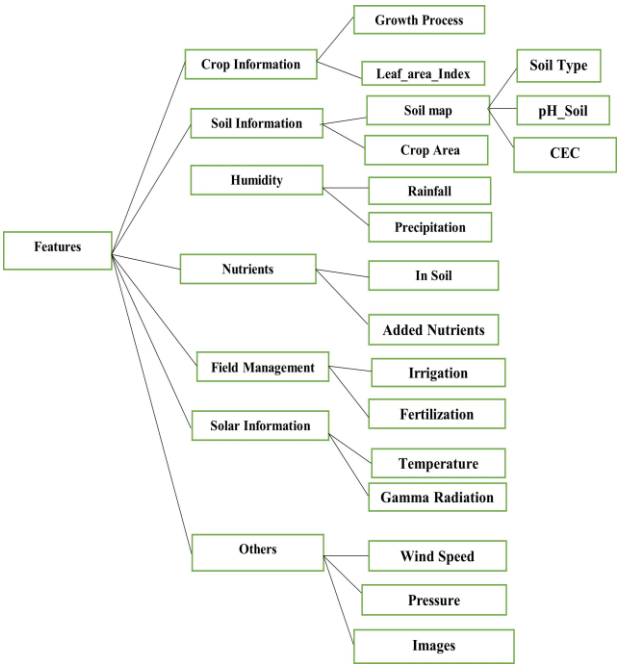

Figure 3: Feature diagram [21]

The above diagram gives the information of the soil parameters. It includes crop information, soil information, humidity, nutrients, field management, solar information and others. Variable of these parameters are growth process, leaf area index, soil map (soil type, pH, soil, CEC), crop map, rainfall, precipitation, Nutrient in soil, added nutrients, fertilization, gamma radiation, temperature, windspeed, pressure, Nitrogen, Phosphorus, Potassium, Wind speed, Pressure, Humidity etc.

7. Q3: Classification Methods

The ability of ML approaches is to autonomously resolve complex non-linear problems with datasets from numerous sources is one of the key advantages. CYP is a crucial application of agricultural data science and technology that aims to estimate the potential harvest output of crops before they are actually harvested. It plays a significant role in modern agriculture, helping farmers, policymakers, and agribusinesses make informed decisions about resource allocation, crop management, and overall agricultural planning. By using historical data, weather patterns, and advanced analytics, CYP models can forecast the yield of various crops for a specific region, season, or farming practice. Linear Regression, CNN, RF are the most used algorithms, in the previous works. Most of the time, linear regression is utilised as a benchmarking technique to see whether or not the suggested algorithm performs better than linear regression. Because of this, even though it is mentioned in numerous articles, it does not necessarily follow that it is the best algorithm. The phrase "most used" should be used with caution, as it may not always imply the best-performing ones. In fact, DL, which is a sub-branch of ML, has been used for the CYP problem recently and is believed to be very promising. In this study, we also identified several feature extraction techniques which are used the earlier research [21].

Bayesian ensemble model (BM) was used which aimed at decomposing historical yield data to jointly estimate technological trends and climate effects on crop yield. They compared BM with ElasticNet, Neural Network, MARS, SVM, Random Forests, and XGBoost.

BM excelled at both predicting and explaining. When tested on synthetic data, BM was the only method unveiling the true relationships: BM has stronger interpretability; other methods predicted well but for wrong reasons. [1] RAMs as a tool to visualize and interpret how the CNN models achieve their results. Activation mapping has been previously applied for image analysis in the classification (Zhou et al 2016) and regression problems (Wang and Yang 2017). Here, we show their application on time series data, as inspired by the application of class activation map to interpret the temporal data in Wang et al (2017)[2]. Furthermore, the text classification is performed based on three ML algorithms. Decision Tree, Hoeffding Tree, and Random Forest are the three types of trees. The text classification method organizes available information into appropriate categories in a systematic manner.
The classification, in this case, is based on a three-class approach. These are the seasons, the varieties, and the recommended zone[3]. Random forest is a very famous ML algorithm that felicitates in cases of both classification and regression issues [11]. This algorithm works on the notion of ensemble learning, which works on the principle of merging several classifiers to give the solution for any complex problem and improve the precision and performance of the applied model. some of the other classification methods used are Support Vector Machine (SVM), Gradient Descent, Long Short-Term Memory (LSTM), Lasso Regression [4]. After completing the necessary preprocessing and splitting the data into train and test sets, the model was trained on the training dataset. The most crucial phase in ML is training.
The prepared data is fed to the ML model during training so it can detect trends and make predictions. As a consequence, the model learns from the data and can complete the goal assigned. The model improves in predicting over time as it is trained. While predicting a continuous dependent variable from a set of independent factors by applying regression analysis. Six alternative regression models were used to train the data, including Linear Regression, Decision Tree Regression, Gradient Boosting Regression, Random Forest Regression, Xgboost Regression, And Voting Regression [5]. MLR, also called multiple linear regression, is a mathematical method for predicting the result of an answer parameter by integrating several logical factors. It depicts the relationship between a continuous dependent parameter and several independent parameters. Other classification models used are Decision tree regressor, Elastic net regression, Lasso regression, Ridge regression and Long short-term memory (LSTM) [6]. Deep learning-based models to evaluate how the underlying algorithms perform with respect to different performance criteria.

The algorithms evaluated in our study are the XGBoost ML (ML) algorithm, Convolutional Neural Networks (CNN)-Deep Neural Networks (DNN), CNN-XGBoost, CNN-Recurrent Neural Networks (RNN), and CNN-Long Short Term Memory (LSTM).An Ensemble Learning (EL) technique is applied on some distinct ML algorithms, i.e., Decision tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Random Forest. For rainfall prediction, a ML algorithm Multiple Linear Regression model is used [11].

Random Forest algorithm which is well known supervised learning algorithm that works on bagging technique [12]. The soil health-based DLMLP model gave satisfactory yield estimation accuracy in the absence of validation of soil health parameter values for the preceding years-2015–16 till 2018–19 wheat seasons[15]. The yields are estimated using Linear models, Support Vector Regressor, K Neighbors Regressor, Tree-based models, Ensemble models and Shallow Neural Networks with R-squared score for evaluation.[16]. Neuro evolution model based on ANN for predicting wheat crop yield [17]. 3D Convolutional Neural Network [27]. CNN-RNN model, along with other popular methods such as random forest (RF), deep fully connected neural networks (DFNN), and LASSO, was used to forecast corn and soybean yield across the entire Corn Belt [28]. The below figure gives some of the different classification algorithms used in the previous works.

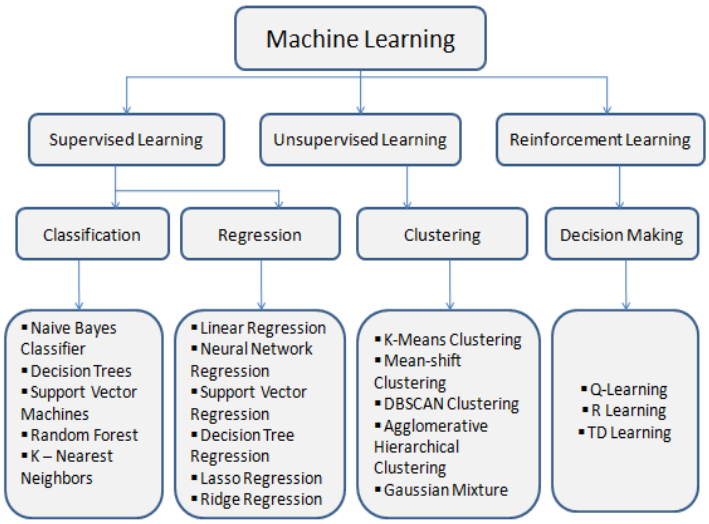

Figure 4: Machine Learning Algorithms Tree Structure

8. Q4: Evaluation parameters

Few evaluation parameters are mentioned in the previous publications are chosen. Almost all studies employed RMSE to measure the model's level of quality. MSE, R2, and MAE (Mean Absolute Error) are additional evaluation criteria. Some parameters, most of which resemble some of the previously stated parameters with a little change, were utilized in particular investigations. These are MCC, RSAE, RRSE, RCV, LCCC, MFE, SAE, and MAPE. The majority of the models produced results with good accuracy values for the assessment parameters, indicating accurate predictions. Researchers preferred the 10-fold cross-validation procedure as the evaluation strategy [21]. Few evaluation parameters/ Estimation Criteria are mentioned in the publications are shown in table 2 [21].

Table 2: All Estimation Criteria used

Key	Estimation Criteria
RMSE	Root- Mean Square Error
R2	R-squared
MAE	Mean Absolute Error
MSE	Mean Square Error
MAPE	Mean absolute percentage error
RSAE	Reduced simple average ensemble
LCCC	Lin's concordance correlation coefficient
MFE	Multi factored evaluation
SAE	Simple average ensemble
RCV	Reference change values
MCC	Matthew's correlation coefficient

9. Q5: Challenges

The challenges were reported based on the observations done in the previous paper. Crop yield prediction models should generalize well across different regions and crop types. Building models that work across diverse agricultural settings can be a significant challenge. Deploying machine learning models at scale in real-world agricultural settings can be challenging. Ensuring that the models work effectively and efficiently on a large scale is a significant practical challenge. Evaluating the performance of crop yield prediction models is challenging, as there is often a lack of comprehensive benchmark datasets. Developing standardized metrics for model evaluation is important. Human factors, such as market prices, government policies, and farmer decisions, can also significantly impact crop yields. Integrating socioeconomic data into models can be complex but essential for a comprehensive understanding of yield prediction. Crop yield is influenced by many factors, including seasonal variations in weather and climate. Models must account for these variations to provide accurate predictions.

10. Conclusion

This study shown that, depending on the area of the research and the accessibility of data, the chosen papers employ a number of features. Each paper examines yield prediction using ML; however, the features may vary. The research' scale, geological location, and crop also vary. The dataset's accessibility and the study's objectives influence the attributes used. Studies have also shown that models with additional features do not necessarily perform well for predicting yield. Models with more features should be evaluated in order to determine which one performs the best. Various research had employed a diverse set of algorithms. The findings indicate that no definitive inferences about the optimal model can be made.

REFERENCES

[1] Tongxi Hu, "CYPvia explainable AI and interpretable ML: Dangers of black box models for evaluating climate change impacts on crop yield, Agricultural and Forest Meteorology,Volume 336, 1 June 2023, 109458.

*[2] Aleksandra Wolanin has presented a work on "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt, 2020, Aleksandra Wolanin et al 2020 Environ. Res, Volume 15, Number 2, **DOI** 10.1088/1748-9326/ab68ac.*

*[3] Reyana, "Accelerating Crop Yield: Multisensor Data Fusion and ML for Agriculture Text Classification", 2023, IEEE Access, Volume: 11, **DOI:** 10.1109/ACCESS.2023.3249205.*

[4] Kavita Jhajharia ,"Crop Yield Prediction using Machine Learning and Deep Learning Techniques", 2023, Volume 218, 2023, Pages 406-417, Procedia Computer Science 218 (2023) 406–417.

[5] Bharati Panigrahi-"A ML-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models:", 2023, Volume 218, 2023, Pages 2684-2693.

[6] Iniyan-Elsivier , "CYP using ML techniques", S -2022.

[7] Abbaszadeh,"Bayesian Multi-modeling of Deep Neural Nets for Probabilistic Crop Yield Prediction" , Peyman -2022, Volume 314, 1 March 2022, 108773, DOI: 10.1016/j.agrformet.2021.108773.

[8] Lontsi Saadio Cedric "Crops yield prediction based on ML models: Case of West African countries", -2022, .

[9] Alexandros Oikonomidisa , "Hybrid Deep Learning-based Models for Crop Yield Prediction",-2022, Vol. 36, No. 1. pp. 1-18.

[10] Swati Vashisht- Taylors & Francis- "CYP Using Improved Extreme Learning Machine for nachine failure multiclass classification ", 2022, 12(16), 3501; https://doi.org/10.3390/electronics12163501, Volume 12 ,Issue 16

[11] Amna Ikram ,"Crop Yield Maximization Using an IoT-Based Smart Decision"- -2022.

[12] Pallavi Kamath , "Crop yield forecasting using data mining", -2021

[13] Anjana, Aishwarya Kedlaya , "An efficient algorithm for predicting crop using historical data and pattern matching technique", -2021

[14] Zheng Chu- ,"An end-to-end model for rice yield prediction using deep learning fusion", 2020, Volume 174, July 2020, 105471, Computers and Electronics in Agriculture.

[15] Akshar Tipathi, Reet Kamal Tiwari, " A Deep Learning Multi-Layer Pecrceptron and Remote Sensing approach for Soil Health based Crop Yield Etimation",International journal of Applied earth observations and Geoinformation, 2022, https://doi.org/10.1016/j.jag.2022.102959.

[16] Nihar Patel, Deep Patel, Samir Patel, Vibha Patel "Crop Yield Estimation Using ML", International conference, Springer, 2021, Soft Computing and its Engineering Applications.

[17] E. Kanimozhi, D. Akila, "An Empirical Study on Neuroevolutional Algorithm Based on ML for Crop Yield Prediction", 2020

[18] Ms. Kavith, " Crop Yield Estimation in India Using ML" International Conference ,IEEE ,2020, DOI:10.1109/ICCCA49541.2020.9250915.

[19] Gunkirat Kaur, "Soil Nutrients Prediction Using Remote Sensing Data in Western India", International Conference ,IEEE International Geoscience and Remote Sensing Symposium 2020, DOI:10.1109/IGARSS39084.2020.9324201.

[20] Mayank Champaneri, "CYPusing ML", 2020Thomas van Klompenburga, Ayalew Kassahuna, Cagatay Catal "CYPusing ML: A systematic literature review", Elsivier 2020.

[21] Mohamad M. Awad, "An innovative intelligent system based on remote sensing and mathematical models for improving crop yield estimation", International Conference 2019, DOI:10.1016/j.inpa.2019.04.001

*[22] Pushpa Mohan and Kian Kumari Patil, "Crop production rate estimation using parallel layer regression with deep belief network", International Conference 2018, December 201, **DOI:** 10.1109/ICEECCOT.2017.8284659.*

[23] Jayantrao Mohite "Spatialization of rice crop yield using Sentinel-1 SAR and Oryza Crop Groate", International Conference Reseach gate, July 2019,DOI:10.1109/Agro-Geoinformatics.2019.8820245

[24] Yao Chunjing a , Zhang Yueyao, "Application Of Convolutional Neural Network In Classification Of High Resolution Agricultural Remote Sensing Images", ISPRS,2017, , Volume XLII-2/W7, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

[25] Miss.Snehal S.Dahikar1 , Dr.Sandeep V.Rode has illustrated "Agricultural CYPUsing Artificial Neural Network Approach" International Journal Of Innovative Research In Electrical, 2014.

[26] Shunping Ji, " 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images,Issue1 ,10.3390/rs10010075 , https://doi.org/10.3390/rs10010075, 2018

[27] Saeed Khaki presented "A CNN-RNN Framework for Crop Yield Prediction" plant science, 2019, Volume 10 - 2019 | https://doi.org/10.3389/fpls.2019.01750.

[28] 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.

[29] A. Suresh, N. Manjunathan, P. Rajesh and E. Thangadurai, "CYPUsing Linear Support VectorMachine", European Journal of Molecular & Clinical Medicine, vol. 7, no. 6, pp. 2189-2195, 2020,

[30] Alagurajan and C. Vijayakumaran, "ML Methods for CYPand Estimation: An Exploration", International Journal of Engineering and Advanced Technology, vol. 9, no. 3, 2020

[31] P. Sivanandhini and J. Prakash, "CYPAnalysis using Feed Forward and Recurrent Neural Network", International Journal of Innovative Science and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020.

[32] N. Nandhini and J. G. Shankar, "Prediction of crop growth using ML based on seed", Ictact journal on soft computing, vol. 11, no. 01, 2020.

[33] Husam Lahza, K.R Naveen Kumar, Optimization of Crop Recommendations Using Novel ML Techniques, 2020

[34] Peijuan,W.,Jiahua, Z., Donghui, X., Yuyu, Z., Rui, S., "Yield estimation of winter wheat in north china plain using RS-P-YEC model", IEEE International Geoscience and Remote Sensing Symposium, pp, 378-381, 2009

.