Deep Science Publishing
https://doi.org/10.70593/978-81-988918-6-0_3

DeepScience
Open Access Books

Chapter 3: Utilizing machine learning models for accurate crop prediction and yield optimization

3.1. Introduction

Precision agriculture (PA) represents one of the most promising applications of Artificial Intelligence (AI) and Internet of Things (IoT) technologies, and its practical implementations often rely on data collected from sensors such as weather stations, cameras, or leaf- and soil-sampling devices. Accurate prediction of crop yield is essential for agricultural producers, companies, and nations that are interested in food supply chain and agricultural planning, not just to optimize profit margins for producers but also to ensure adequate food security for the population. During the past three decades, machine learning models have become the primary way to address crop yield prediction tasks, owing to their efficient leveraging of large historical crop yield datasets in modeling complex nonlinear functions between exogenous factors and yield. In this paper, we examine this large body of literature on traditional and recently proposed machine learning techniques and how they succeed in predicting crop yield evaluated on different crops and countries (Jeong et al., 2016; Chlingaryan et al., 2018; Kamilaris & Prenafeta-Boldú, 2018).

Predicting the yield of agricultural crops is essential for various agriculture applications such as worth excess production, food supply chain, and agricultural investment, agricultural balance sheets. Harmful effects on the agriculture economy derived from yield unpredictability could be alleviated using crop yield prediction. In addition, crop yield prediction is a topic of vital importance for assuring food security. With the constantly increased global population, it is of utmost importance that agricultural activities maximize production, while minimizing land resource inputs. In this sense, crop yield prediction is important so that early warnings of crop failure could be issued, contributing to the effective management of food shortage as well as the development of informed decision-making systems (Khaki & Wang, 2019; Mishra et al., 2021).

3.1.1. Purpose and Scope of the Study

In the early 1900s, America experienced staggering losses as the Dust Bowl drove farmers from their land. In today's world, there are still cyclical patterns that affect global food supply and demand. Widespread use of drought resistant crops made the dust storms a thing of the past. In more modern history, in the spring of 2020, panic buying and quarantine caused consumers to worry about food supply issues as well. COVID-19 posed challenges with global travel and its impact on the supply of food. Farmers faced empty grocery store shelves after crops had been planted. On the demand side, consumers rushed and emptied shelves of basic food supply items. In 2021, the world watched the crippling problem of farmers dumping milk as processing plants lay dormant.

Fig 3 . 1 : Precision Agriculture's Future

The pandemic posed an interesting challenge because some of the biggest suppliers of our food source were suddenly forced to quarantine under health mandates. Food waste seen in the spring of 2020 and dropping commodity prices in the summer of 2020 helped to shed light on the cyclical nature of global food supply and demand and the inefficiencies in how we raise and grow our food supply. Questions arose surrounding the efficiency of global food sourcing through farming. How can technology play a role in modernizing outdated data collection for helping farmers increase yield and become more productive? How can machine learning models help in addressing some of the challenges of federal policies on agriculture that take years to enact? The key question this dissertation will seek to address is: Are machine learning models suitable for crop yield prediction, with the research objective being to explore the applicability of machine

learning models to accurately predict crop yields and lessen the growing cycle of the prediction through the automation of the data collection process? This research will explore hyperspectral remote sensing and the use of satellite data to source near real-time data with the goal of crop yield prediction that is timely and accurate.

3.2. Overview of Crop Prediction

Crop prediction is a crucial aspect of agriculture that involves forecasting the quantity and quality of crops to be harvested in a specific area. Accurate crop prediction enables effective planning for the sale and distribution of crops, as well as the management of input and resource allocation prior to planting. With the rise of the internet and its associated technologies, soil data, monitoring systems, and image processing techniques have become easily accessible to farmers, who can leverage these tools to improve the quality of their agricultural production. For instance, web cameras are being used to record crop growth on a daily basis, while aerial photographs obtained from unmanned aerial vehicles provide detailed snapshots of fields from above. It is necessary to make timely decisions on which crops are to be grown and which inputs are to be applied, in order to ensure optimal yield. Yield prediction of a crop is important for many reasons, including planning for food demand and management of crop production and distribution.

The increasing demand for agricultural products, coupled with climate change and its effect on yield, heightens the risk of crop failure. Successful prediction of crop yield can help the government in the management of resources in response to the actual agricultural situation. In India, resources are made available after the prediction of the crop yield data for the year. Reliable estimates of crop acreage and yield at the right time are essential for effective planning, policy formulation, policy evaluation, and implementation of duty policies. Accurate information helps to check black marketing and artificial price inflation. An extensive crop yield prediction model is essential for the assessment of controls over production, marketing, and consumption of agricultural products. Crop yield prediction is essential in food policy formulation, framework for food insecurity, and estimating the impact of climate change on food security in most underdeveloped countries.

3.2.1. Significance of Crop Prediction in Agriculture

The agriculture industry is an exceptional global endeavor that meets the demand for food products worldwide. Moreover, this industry contributes significantly to overall economic growth, particularly in developing countries. Crop yield is the crucial agricultural ratio that reflects the plants' matured phase's performance to determine the

crops' final quantity and quality. However, this criterion is influenced by various factors, such as soil, climate, and environmental conditions, making yield prediction extremely difficult. Therefore, predicting crop yield with sufficient accuracy is an essential requirement of policy-makers and industry leaders to enhance investor interests and mitigate losses. Since ancient times, agriculture has existed as the network of interactions between society and the sun, which promotes a viable and sustainable economy. As a result, agriculture has evolved into the extensive and complex industry that we now see benefitting from technological, structural, and informational improvements, as well as scientific and financial innovations.

Agriculture is the only activity entirely dependent on the environment's climatic characteristics. Thus, crop yield has always been especially dependent on available climate variables, such as temperature and precipitation. Due to an increase in demand, uncertainty surrounding food security, increased incidence of extreme weather events due to climate change, and population growth, crop yield is under severe pressure and the agriculture industry makes the shift towards more sustainable and resilient practices. These industrial, technological, and climatic shifts require systematic reforms, new tools for producers, increased education, and the development of new policies for climate change adaptation for a more resilient real industry. Researchers have focused on crop prediction, with machine learning methods being favored, due to their accuracy and ability to include diverse input parameters.

3.3. Machine Learning Fundamentals

Developing an accurate model to predict the yield of crops can be challenging due to multiple variables, such as uncertainty in weather parameters and climate change. Traditional forecasting methods increasingly show limitations in accuracy, as they typically rely on mechanistic models built on extensive assumptions and expert knowledge of the interactions among different factors affecting crop development. In this study, we utilize machine learning techniques to overcome these limitations and build prediction models that are completely data-driven and rely solely on seasonal and spatial weather predictors. Derived from artificial intelligence, machine learning is a scientific discipline allowing computer systems to gain knowledge from empirical data and improve their performance on a specific task. In the field of machine learning, models are built upon data; if a computer is provided with the appropriate tools and a significant amount of data, it can discover hidden patterns and perform intelligent computations that can be hardly explained with probabilistic or mechanistic models. Moreover, model complexity and large data volumes are two other factors that make machine learning an excellent approach for modeling and forecasting multivariate time series.

According to the different methodologies employed to build the models, machine learning can be split into three main categories. Supervised learning is the most widely studied form of machine learning, where the task is to predict the output associated with the input from a finite set of labeled input-output pairs, assuming that the previously unseen instances are drawn from the same distribution as the training data. On the other hand, in unsupervised learning, the model is unable to access any labeled data, but looks for correlations in the input properties of the data. In particular, the task of clustering groups observations so that internally they are similar, while between different clusters they are distinct. Finally, reinforcement learning can be seen as an extension of supervised learning, where the educator is the environment that provides feedback for the current agent's action by means of reward and punishment.

3.3.1. Types of Machine Learning

Machine Learning is the computational science that utilizes algorithm and statistical model to learn from data and performs a specific task without prior knowledge. Machine learning has numerous applications in agriculture such as crop prediction yield estimation, plant disease detection, soil fertility prediction, and weather prediction which are useful for effective planning of cropping systems. The field of machine learning has gained wide popularity owing to its several applications ranging from advertising to automated driving, from healthcare to NLP. Machine learning can be classified as supervised machine learning, unsupervised machine learning, semi-supervised machine learning, and reinforcement machine learning which are differentiated based on how the algorithm learns and adapts. Briefly, supervised machine learning is the most commonly used machine learning method in which the algorithm learns from labelled inputs. In unsupervised machine learning neither labelled data nor supervised response signal is provided. Reinforcement machine learning is different from supervised or unsupervised learning as it learns based on the consequences of previous actions.

Supervised machine learning contains techniques like regressions, decision tree, support vector machine, artificial neural network and convolutional neural networks, k-nearest neighbor, and boosting algorithms for classification problems. The most commonly used unsupervised learning techniques are suggestion with multilayered graph-clustering based hierarchical algorithm, k-means clustering and its variants, dimensionality reduction by Principal Component Analysis and t-distributed Stochastic Neighbor Embedding, and Gaussian mixture model. The most commonly used reinforcement learning algorithms are Q-Learning, Monte Carlo method, and Temporal Difference Learning. For crop prediction and yield estimation Q-Learning, Monte Carlo method, and Temporal Difference Learning algorithms are frequently used.

3.3.2. Key Algorithms in Machine Learning

Machine learning has turned into a tremendously powerful and successful tool for a variety of applications in areas that cover all walks of known human life. It is also easier than before to make almost any kind of machine learning project. When we look at the present developments and the large and ever growing success of machine learning, we can ask ourselves what has made it possible. Truly, there are many aspects that have made it happen. Some of the more decisive ones are the availability of large amounts of data to feed the learning models, the availability of computing resources that for a long time were required for the learning process, and finally the development of efficient algorithms that for the first time have shown to be successful on many arbitrary problem domains. In the following, we will look a bit closer at the algorithms we are referring to.

In general, we can sort the learning algorithms according to the kind of function they are trying to learn. The functions are often referred to as hypothesis or ensembling functions, given that a lot of methods are combining the different features on a non-linear level. The first category of functions are classifiers. The input variable can be either a vector or a simple structure with vectors inside. The function of a classifier does a mapping to an unstructured result. The second category are regression functions. They map inputs to real valued outputs. The last type of function to learn are the so-called formative functions, that try to express latent structures or relations that generated the observed training data.

3.4. Data Collection and Preprocessing

Data Collection and Preprocessing is a very vital phase during which data related to climatic, soil, and other crop factors are collected. The collected data is then subjected to the data cleaning process to handle noisy and non-reliable data. The preprocessed data is now ready for analysis and can be used for building machine learning models for the crop prediction and yield optimization tasks. This section describes the types of agricultural data required along with the challenges and solutions in the data preprocessing phase.

A variety of agricultural data including climatic, environmental, soil, and crop information are required to address the crop prediction and yield optimization challenges. Farmer surveys, on-ground field visits, and harvesting data are very tedious and time-consuming methods to collect crop data. Remote Sensing Technology and Satellite Remote Sensing methods have proven to be the best choice for collecting climatic, environmental, soil, and crop data. Although these methods are not completely error-free, they are capable of gathering the required data from vast areas in a very efficient and cost-effective manner. In this study, we consider 28 different features

related to soil, climatic, topographical, and crop management information for 58 crops from states in India. Soil and climatic data are obtained from the Indian government's Soil and Climatic Data established for research and are publicly available. Data related to various crops is obtained from the Indian Council of Agricultural Research. Data deficiency poses huge challenges, for example, in the case of immature crops wherein the crop management data is not available.

3.4.1. Sources of Agricultural Data

Agriculture as a vast area deals with a variety of activities like land preparation, planting, irrigation, harvesting, drying, storage, etc. of crop and livestock for serving food and fiber. Therefore, the data involved in agriculture is real time data which is available from different sources and under various formats. Traditionally, agricultural data are collected at the field level by using remotely sensed technology. However, the big data sources for collecting agricultural data is satellite, UAV, weather stations, and land observatory stations, in addition to the field surveillance using mobile sensors. Therefore, the research on agricultural data for big data analytics is attractive due to the involvement of changing environmental conditions, requirement of real-time data, solving worldwide food production issues, etc.

Some examples of the known top agricultural datasets include various government censuses and collections of international financial data on agriculture. Some economic datasets maintain data about surveys and censuses, whilst their physical science datasets contain detailed information regarding climate, crop growth, and livestock production. There have been heavy investments in the creation of information and data systems, which also produce a subset of the distributed global crop modeling dataset.

3.4.2. Data Cleaning Techniques

In general, the acquired data may not be directly suitable for performing predictive modelling. In order to become usable, the data needs to go through several transformation steps often referred to as preprocessing procedures. The most commonly encountered problems with data include duplicate records, irrelevant features, data leaks, missing data, noisy data, improper formats, and data imbalances. In this section, we will discuss the methods used for solving these problems in the studied data and prepare it for use in the machine learning predictive models. The first major preprocessing step is duplicate removal. Duplicate records can bias results by assigning more weight to the data corresponding to those records. Next, we need to remove any irrelevant features. At least three features – state name, state code, and commodity name are redundant and can be removed. The next considered step is checking for data leaks. A data leak occurs

when the information from the test set is used for the model training leading to prediction accuracies that are not possible in the real world. In this work, there are no clearly defined borders between the training and testing datasets. The entire data is part of the same yearly cyclical process that is effectually continuous. However, the models are still capable of making accurate predictions for the new unseen years and therefore formal data leaks are not present. The next step is to check for missing values in the data and impute them. Although many columns contain values, a quick review of the algorithms used later reveals that they are capable of handling the missing values and we do not need to handle it.

3.4.3. Feature Selection Methods

Choosing the right features is one of the most critical tasks in developing ML models because the right features can reduce model complexity, time involved in training, and costs, while improving performance levels. Conversely, using the wrong features can damage model accuracy, outcome interpretation, speed, and costs. Therefore, in this chapter, dimension reduction techniques are suggested to shrink data created from geospatial, agro-climatic and socio-economic sources and select an optimal subset of features that have the greatest relevance to crop productivity forecasting. Several models can carry out FSD: we can use dimensionality reduction algorithms or even ML classifiers that combine feature importance analysis as a part of their functionality. In our case, we decided to make use of both approaches; we will combine the analysis of dimensionality reduction techniques with that of FSD models. We will use PCA, LDA, t-SNE and UMAP as dimensionality reduction techniques. As models for the selection of important input features, we will use random forests, XGBoost, Lasso regression, ElasticNet and the permutation importance method. PCA is a linear projection trick often used in lieu of exploratory data analysis, LDA is a linear projection model that tries to separate classes by maximizing class variance, t-SNE is a nonlinear projection technique that performs well in visualization scenarios but is very slow, and UMAP is an alternative to t-SNE. Random forests and XGBoost are examples of ensemble predictors that work on the top of less-accurate tree-based algorithms. They build numerous decision trees, combining the results of each of these trees through majority voting or averaging.

3.5. Model Development

Many kinds of machine learning models have been applied to crop yield prediction. In this study, we used multiple linear regression, quadratic regression, random forest regression, gradient boosting regression, extremely randomized trees, decision trees,

support vector regression, k-nearest neighbors, and artificial neural network regression models. We used standard psychometric techniques for hyperparameter tuning, model selection, and model performance evaluation. For regression models, we used the standard hyperparameters for the applicable models, except we tuned the maximum depth hyperparameter of decision trees using a grid search. For the artificial neural network regression models, we used a multi-layer feed-forward neural network with three hidden layers with 50, 30, and 10 neurons, respectively. The loss function was set to optimized for the mean squared error, and the means of updated weights were implemented using the Adam optimizer. A set of momentum values were used to tune the performance of the model for the number of epochs. Other hyperparameters were set to standard values.

The models were tuned using 60% of the data and evaluated on the other 40%. A 10-fold cross-validation technique was implemented to select the best K-nearest neighbors. The number of splits was set to 10, and the random_state was set to a specific integer. The K-nearest neighbor algorithm was chosen because it does not implicitly learn a function. Instead, it memorizes examples in the training dataset. Therefore, a cross-validation technique is suitable for this type of machine learning model. The hyperparameter was the number of neighbor samples. The maximum depth hyperparameter of the regression tree was tuned by a grid search over the possible values.

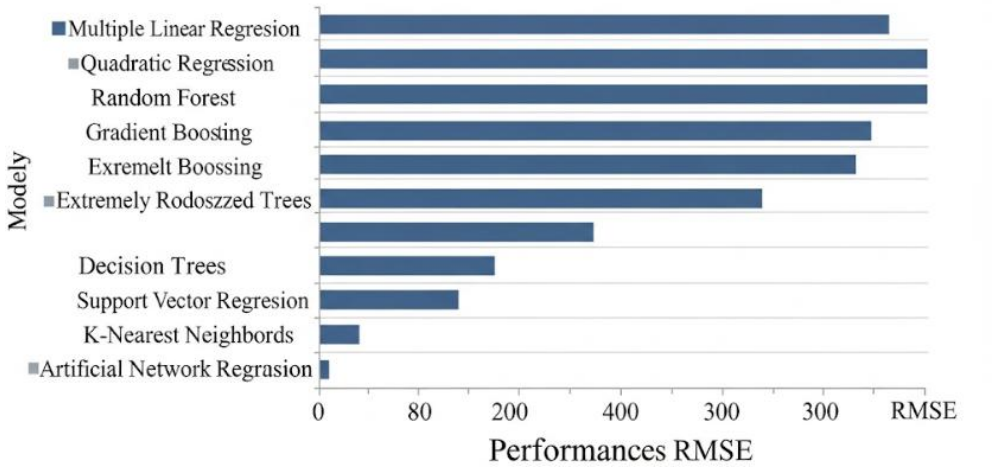

Fig 3 . 2 : Comparative Performance of Machine Learning Models in Crop Yield Prediction

3.5.1. Training Machine Learning Models

We utilize three different machine learning algorithms to create predictive models for the target variables, namely, the crop category for classification and yield for regression.

The predictive model is then trained by using two supervised approaches. The supervised classification algorithm used is decision tree in this study. For regression, we utilized a number of different algorithms – decision tree regression, ensemble learning techniques such as random forest and gradient boosting machine. The models were implemented using popular libraries. The models were tuned to their optimum performance with respect to hyperparameters using random search combined with 5-fold cross-validation. For decision tree, the criterion was tuned to either gini or entropy. For decision tree regression, we maintained minimum number of samples for a leaf and maximal depth of the tree as hyperparameters during tuning. For Random Forest model, the number of trees were tuned to either 10 or 100 while maintaining the number of minimum leaf samples during tuning. For Gradient Boosting model, the number of boosting stages were tuned to either 50 or 100 while maintaining the learning rate during tuning. The decision tree and the other ensemble models were combined using hard and soft voting as well. The essence of hard and soft voting is that in hard voting, the best predicted class of each model is taken to create majority votes. The class with the maximum majority is taken as the final model decision.

The VotingClassifier has combined hard voting and soft voting. The final label is predicted by weighting the predicted classes using the class probabilities of the algorithms combined with the class decision using the predicted method. The class with the maximum value in the weighted array is taken as the final label. Order of models is very important in soft voting. In this study, we utilized decision tree, random forests, and XGBoost to create the soft voting predictive model. The order of the models utilized for soft voting is Decision Tree — Random Forest — XGBoost.

3.5.2. Hyperparameter Tuning

Machine learning has a concept known as hyperparameters, which are parameters that control the learning process. Hyperparameters are set before the learning process begins and their values influence the speed and quality of the learning. Unlike other parameters in machine learning, hyperparameters are not learned from the training data, but are set before the learning process. Finding an optimal combination of values for hyperparameters can be very costly. Automated hyperparameter tuning can save both time and computation power, as opposed to manually adjusting the hyperparameters of the model. The best results during training often come from grid search on the hyperparameters, though this method is expensive to run.

There are two different types of hyperparameters: algorithm hyperparameters and model hyperparameters. Algorithm hyperparameters relate to the algorithm itself, while model hyperparameters can enhance the predictive performance of a machine learning algorithm. Hyperparameter tuning can drastically improve a model's predictions, though

it can be expensive to run. With many more advanced algorithms built on top of the classical algorithms, there are many more hyperparameters to tune. The standard methods of hyperparameter tuning include grid search, random search, and hyperband, though recent advancements have created methods that can tune even deeper algorithms more efficiently.

3.5.3. Cross-Validation Techniques

In some cases, it is desirable to decrease the value of the train/test split (or increase the ratio of examples used in training) to increase the chances of our models being in a locally optimal state. In such cases, k-fold cross-validation is beneficial. In k-fold cross-validation, instead of splitting the dataset into a validation set once, we take k different split points and create a model for validation using k-1 segments of the training set. Then we average the score of KPI value for these models (the validation score). Cross-validation reduces the risk of overfitting the models and helps to reduce the bias in our estimation and get a better understanding of how the selected model will generalize on unseen data. In some datasets, the split using random stratification techniques can occur in a way that makes the dataset different enough to impact the results severely, and models can perform poorly on unseen data.

Considering the low amount of data and the possibility of underfitting, a variant of this method is preferred (stratified k-fold). In stratified k-fold cross-validation, we use stratification techniques, which deliver a distribution from the full dataset by segmenting it using the dataset's most representative independent variables. The problem with k-fold cross-validation is that the models are trained k times on a non-overlapping dataset. This increases the k times of the model's time complexity without a large-scale change in the time complexity distribution. Hence for large datasets and models that take a long time to train, especially deep learning models, stratified k-fold cross-validation would not be practical in a production scenario. For these kinds of data and tasks, holdout validation is the way to go.

3.6. Crop Yield Prediction Models

Crop yield prediction models can be classified based upon two main ideas which include: prediction target and prediction methodology. Based upon the prediction target concept, the crop yield prediction models can be further classified into two main types which include: regression models and classification models.

1. Regression Models

Regression models are used to estimate the output crop yield in terms of its numerical value, and such models include polynomial regression model, linear regression model, multiple linear regression model and support vector regression model. Polynomial regression models are simple models that have high predictive power of data, but they suffer from the problem of increasing the number of features. Linear regression models and its variation multiple linear regression models are one of the simplest yield prediction models. These models are also easy to implement. These models seek to establish a linear relationship among the crop yield and input features. Support vector regression (SVR) method relies on the principle of structural risk minimization, it has the ability to avoid over-fitting. SVR can use kernel functions to manage non-linear mapping, and it has fewer free parameters than neural network does.

2. Classification Models

Classification models estimate the target crop yield through predicting the class labels, thus they have the ability to capture the underlying nonlinear mapping relationships among the input data features and set of crop yield. Crop yield prediction models falling in this group include decision trees, artificial neural networks, random forests and so on. Decision tree is a tree structured prediction model that provides a graphical representation of possible solution paths and options. Decision tree has the advantage of having an intuitive approach, it can be used for both classification and regression problems; it can handle multidimensional data and its performance does not depend on the structure of the data. In addition to the above characteristics, decision tree can also be modified to form its ensemble version. The ensemble version of decision tree takes an ensemble of many decision trees, and it has shown the promise in increasing the predictive performance of predicting models.

3.6.1. Regression Models

Recently, crop yield modeling has profoundly affected the field of agriculture, which is increasingly becoming data-intensive. Crop modeling enables evaluation of both the potential levels of productivity as well as the impact of environmental management practices on future yields. Consequently, it has wide-ranging consequences for food supply forecasts, food security issues, and global changes. During the last couple of years, new crop yield prediction models have emerged, enabled primarily by the exponential growth in satellite remote-sensing data and the increasing availability and stability of Internet connections around the world. Sensors mounted on satellites provide reconnaissance data that are available at low incremental costs, enabling global coverage and fine temporal resolutions. Remote sensing models are able to track key indicators of crop growth during the growing season, including plant vigor, nutritional status, phenology, and water use. The remote sensing response from satellites illustrates ground

brightness variations as a function of specular reflection and back scatter. Consequently, it electrically encodes information on the land surface-atmosphere interactions in the form of land radiance. This brightness variation is observed to follow a smooth curve, which relates the peak response to energetic radiance properties of land surface material. Many studies have reported the use of satellite sensors to predict area yield over small regions. Generally speaking, crop yield estimation models based on remote-sensing observations can be classified into two categories: empirical models and crop growth simulation models. For example, regression algorithms have been built to automatically estimate corn yields from satellite reflectance measurements over the central United States. Crop yield maps have been generated from the reflectance data as well as from farmer-provided yield data.

3.6.2. Classification Models

Machine learning has been extensively used and proposed in various types of applications, machine learning models for crop type prediction have been proposed on many levels. These model types include Fuzzy Decision Trees, a Random Forest, a Support Vector Machine, a Naive Bayes classifier, a linear discriminant analysis model, and a greedy learning classifier system. With its high classification accuracy, the greedy learning classifier system is considered the best solution for pixel-wise classification of multispectral remote sensing data. However, the implementation is slow and not suitable for applications needing results in shorter binary times, such as crop monitoring. Instead of using a greedy learning classifier system for such applications, researchers advocate using Random Forest, a boosted Random Forest algorithm, or a Support Vector Machine as the crop classifiers due to their more rapid performance. In addition, there have been various studies on crop classification using Random Forest, Support Vector Machine, and boosted Random Forest or other combinations of the two.

Decision trees have been widely used due to their fast processing speed, robustness, and classification performance. However, a single decision tree often works badly when classifying a subspace with a different field structure and relationships. In contrast, a Forest of Randomized Trees improves the weak performance concern of single Random Trees by training the classifiers based on multiple instances. They combine the responses from many trees to make classification decisions and rely on a simple average instead of a majority vote, unlike traditional Random Trees. The outputs of individual trees are averaged for prediction of the total population as well as being calculated on grasslands.

3.6.3. Ensemble Methods

The ensemble methods refers to a group of machine learning models that attempt to achieve better prediction accuracy and/or better generalization performance by combining the predictions from multiple models than simply using one predictive model. Any machine learning algorithm can suffer from high bias or high variance depending upon the applied data. Therefore there is a need to carefully choose machine learning algorithm type per data set.

Ensemble algorithms build a model which consists of multiple sub-models. These models can be both heterogeneous or homogenous set of models. The synthesis of the models are carried out by methods such as bagging, boosting, stacking. Bagging refers to bootstrap aggregated models. In this method models are trained independently. While fitting these models, a random set of data is chosen using bootstrap sampling with replacement from the training data way bigger than the original set of data. Bagging helps in reducing overfitting. Random Forest, an ensemble method uses Bagging on Decision Trees. Bagging is a model averaging ensemble method. The most common type of model averaging is majority voting, which gives some weight to each model and selects the one that achieved majority votes with the sum of weights being one. Stacking also known as stacked generalization algorithm is a heterogeneous ensemble technique where multiple different types of models are applied on the same training data set. These trained models are combined with a meta model, as a second stage classifier which acts on the outputs given by the sub-models. The meta model is also trained on a fraction of the training data marked for validation of the sub-models. Stacking helps in variance reduction.

3.7. Challenges in Crop Prediction

While machine learning offers exciting prospects for improving crop prediction, it also presents a range of challenges. Understanding these challenges is essential for both researchers utilizing ML for crop prediction and policymakers seeking to facilitate progress through supportive policy measures. This section outlines three of the most significant challenges currently faced by researchers utilizing ML to improve crop prediction.

1. Data Quality Issues

One major challenge in deploying ML models to improve accuracy of crop predictions is that time-series datasets are notoriously problematic. Data are often missing for extended periods of time, which can present complications for the training process. Moreover, data may contain large numbers of outliers due to various recording issues such as sensor failures or faulty sensor hardware. While adding extra data points to

training or prediction datasets may mitigate some of these issues, ML models operate by recognizing patterns in the data, and extraneous data points can lead to model overfitting. Researchers using time-series datasets for prediction would be well-served by investing significant time during preliminary data processing to clean the dataset to the extent possible.

2. Model Overfitting

Another issue that researchers deploying ML solutions for crop prediction need to be aware of is overfitting. The models discussed in this paper all require significant tuning on labeled datasets of sufficient size in order to avoid overfitting. While tuning hyperparameters for models with large numbers of hyperparameters is a well-developed field containing several common strategies, the need for labeled data on which to tune these hyperparameters may negate some of the benefit of using data-driven methods over traditional methods that rely less on labels.

3. Interpretability of Models

Finally, models such as deep learning techniques have the underlying structures and behaviors that are extremely difficult for human research analysts to interpret. This is particularly troubling when modeling growing processes that are not well-established, or when unbiased input features are not readily available. Intensive manual exploration is often required to select input features, a process that diminishes some of the benefits of using large datasets to capture complex relationships in crop growth.

3.7.1. Data Quality Issues

Utilizing advanced Machine Learning models can be a powerful tool for predicting crop yields and optimizing production in Precision Agriculture. The efficiency of ML models strongly depends on the quantity and quality of their input data. Likewise, the prediction of agricultural yield using ML models suffers from similar data quality issues that impact several data-driven models across other fields. High ML model accuracy depends on clean and reliably represented patterns in the data as generalizable predictors. Accurate detection of these predictive features in turn is dependent on the clear associations of yield with input factors over a longer period. Different types of data issues related to the yield prediction domain can lead to unreliability in ML models, thus creating challenges in crop yield prediction and use for PA.

Using manual methods, crop yield predictions are bound to be dropping in performance as realization of high quality data around input factors becomes suspicious. Unmanned Aerial Vehicles, satellite aerial photography, sensor networks to capture real-time environmental data, technology for soil testing and application of moisture-monitoring

data loggers can tackle the demand of reliable data generation to be used in ML applications. The introduction of UAVs in an agricultural production setting allows near real-time collection of information relevant to both vegetation health and crop yields over large areas. With such data gathering tools, the storage and transportation of high quality data to be used for ML models can be addressed. However, the reliability of data storage and transfer can be subjected to manual efforts which can have a cascading impact.

3.7.2. Model Overfitting

In the case of data scarcity, deep models exhibit a major issue of overfitting. The cause of this issue can be explored by considering the approximation theory of deep networks. Deep networks are universal approximators with suitable activation functions; but not only that, they are equipped with the capability of locally approximating arbitrary nonlinear functions within a region of arbitrary dimension. To accomplish that, deep networks must operate with sufficient hidden neurons and parameters. The problem is that this is also what makes them prone to overfitting. When enough data is available, deep models are capable of learning from the data distributions the features that should be exploited to capture only the most relevant variations in the multi-dimensional data space; and for this reason, they are able to achieve worse generalization compared to traditional models.

The number of hidden neurons is critical. It should be higher compared to that of traditional models, possibly much higher; yet, the more hidden neurons there are, the more susceptible to overfitting deep models become. In addition, most models are indeed based on the same activation function. For many datasets, well-tuned image classifiers have less capacity than other traditional algorithms that employ fewer parameters. Certain deep models are supposed to provide more coarse-grained approximations rather than very detailed ones. More detailed approximations are not always better. When predicting yield, it may not be beneficial to predict the yield at a vertex of the grid induced by the resolution over which data is gathered. More coarse-grained approximations may be needed in a practical context, because detailed predictions would be very unstable, and moreover, for fine-grained predictions to be useful in a practical sense, the difference between fine-grained yield predictions should reflect differences in sowing density or nutritive contributions.

3.7.3. Interpretability of Models

While we can derive the mapping of the input feature space X onto the yield class space Y from trained models using an abstract representation such as a neural network, that

mapping may be abstract to the point of incomprehensibility. Thus, model output does not in itself allow us to really understand how changes in input features would change the yield probabilistic class distribution. One of the main drawbacks for many machine learning models is the fact that they are black-box models. Once the model has been trained, we can query it to get the leaf nodes for specific input in a random forest or the local regression coefficients for some input in a locally weighted linear regression model. Each of the input-output mappings has some kind of probabilistic consensus about the neighborhoods of the queried model. How much that local mapping is trusted affects the reliability of predicted future debugging, i.e., the future behaviors of the model. There are some local interpretable model-agnostic methods for assessing black-box models using answers to queries of the model. The explainability of all of these methods is never greater than the explainability of the black-box model being queried. Because yield prediction is inherently a predictive rather than explanatory problem, much of the research in recent neural network work on modeling image data assumes that the complexity of this function can be characterized by its approximate Lipschitz continuity.

In yield prediction applications, black-box models have predicted future yield distributions more fruitful than low-complexity modeling methods like linear regression methods. The field of explainable AI has yielded methods for producing explanations for black-box models. An explanation attempts to specify which input features really determined a response. The most significant interpretable black-box model is the explainable boosting machine, a generalized additive model that supports interactions. This model uses feature selection to pare down the input feature set before random forest building methods create an explainable boosting model from a regularized method that penalizes complexity. Each of those learning methods supports restricted output functions so that their predictions depend longitudinally on only a small set of input features.

3.8. Conclusion

Information and communication technologies (ICT) have witnessed a rapid expansion recently. Research into agriculture's diverse technological features has been spurred by ICT development. Machine learning has expanded into every field of economics research. The future of development relies on smart agriculture, and research in this area has become the focus of economists. Essays on agriculture generally lack the use of modern data analysis technologies. In this paper, we explore facilitating agriculture research in two ways: choosing machine learning to provide more accurate predictions and policy implications; and combining various regression settings and machine learning with varied data science syntax to produce high accuracy.

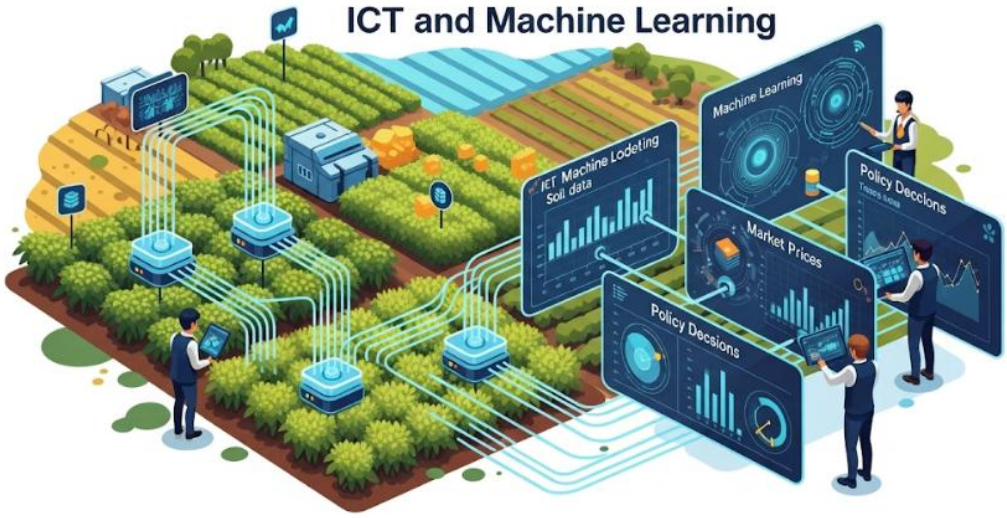

Fig 3 . 3 : ICT and Machine Learning for a Sustainable Future

The importance of utilizing advanced information technologies and novel econometric nonparametric tools to predict what is most significantly impacting the food and agriculture sector has been identified in the paragraph above. More agricultural policy implications about optimizing crop yield and crop production ratio using crop prices, seasonal factors, inputs, modern technology adoption, climate change, and diversity of soil types across land at the second stage, and the impacts of diversification of soil across land types for crop price prediction at the first stage, should be explored in the future. Enabling policy decisions to be data driven is the future of our research. In particular, we hope to revisit and enrich these models when more new soil data arise from their increasing availability for major agricultural ecologies and world cereals about climate-smart innovations of different crop varieties, trenching, different fertilizer innovations, climate insurance, different planting practices, and managed drainage.

3.8.1. Summary and Implications for Future Research

Crop yield prediction is one of the most challenging and difficult tasks due to the involvements of so many publicly available specific and non-specific parameters. To this date, very few models have been able to predict crop yield using all publicly available specific and non-specific parameters. This is important because managers would want to use simple applications to upload all specific and non-specific parameters needed for crop yield prediction without going through so many steps and learn the complexities of crop yield prediction. By aggregating crop yield data at geographical levels, specific and non-specific parameters can be inferred and crop yield prediction models can be trained and used for prediction. Furthermore, these models can be used to

optimize crop management practices, soil amendments and fertilizers needed to apply to get targeted crop yield predictions. This could, in effect, help avert food shortages and malnutrition in parts of the world that suffer from them.

In this paper, we presented deep artificial neural network models to predict maize, rice, and wheat yield with publicly available databases. The correlations between observation and predicted crop yields are in high-level and reflect the capabilities of the models developed. Simple applications with user-friendly interface have been created to enable managers and stakeholders to access the trained models for real-time predictions. Finally, we elaborate on the advantages, limitations, and implications of this research, the directions for future research are also mentioned. The proposed work could not just be used to predict crop yield, but also to prescribe soil amendments, fertilizers and optimal crop management practices needed to get targeted crop yield prediction which could, in effect, help avert food shortages and malnutrition in parts of the world that suffer from them.

References

Jeong J.H., Resop J.P., Mueller N.D., Fleisher D.H., Yun K., Butler E.E., Timlin D.J., Reddy V.R., Kim S.H. (2016). Random Forests for Global and Regional Crop Yield Predictions. PLOS ONE, 11(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

Khaki S., Wang L. (2019). Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science, 10, 621. https://doi.org/10.3389/fpls.2019.00621

Mishra A., Singh V.P., Pandey R.P. (2021). Application of Machine Learning in Crop Yield Forecasting: A Comprehensive Review. Agricultural Reviews, 42(3), 218–227. https://doi.org/10.18805/ag.R-2106

Kamilaris A., Prenafeta-Boldú F.X. (2018). Deep Learning in Agriculture: A Survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Chlingaryan A., Sukkarieh S., Whelan B. (2018). Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review. Computers and Electronics in Agriculture, 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012