Smart Agricultural Technology 2 (2022) 100048

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.elsevier.com/locate/atech

Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries

Rubby Aworka[a], Lontsi Saadio Cedric[a], Wilfried Yves Hamilton Adoni[b,*], Jérémie Thouakesseh Zoueu[c,d], Franck Kalala Mutombo[a,e], Charles Lebon Mberi Kimpolo[a], Tarik Nahhal[f], Moez Krichen[g,h]

[a] African Institute for Mathematical Sciences, Senegal - Ghana - Rwanda
[b] International University of Casablanca, Casablanca, Morocco
[c] University of San Pedro, San Pedro, Côte d'Ivoire
[d] National Polytechnic Institute - Felix Houphouet Boigny, Yamoussoukro, Côte d'Ivoire
[e] University of Lubumbashi, Lubumbashi, Democratic Republic of Congo
[f] Hassan II University of Casablanca, Casablanca, Morocco
[g] Albaha University, Al Baha, Saudi Arabia
[h] REDCAD Research Unit, Sfax, Tunisia

ARTICLE INFO

Keywords:
Smart farming
Yield prediction
Machine learning
Gradient boosting machine
Support vector machine
Random forest

ABSTRACT

Food security has become a real challenge for some organizations in charge of the food program and for the majority of countries, especially African countries. The United Nations Organizations' has recently defined the end of hunger and the improvement of food security in 2030 as its primary goal. Improving food security could also pass through the handling of agricultural yield. Agricultural yield is affected by climate changes since this latest decade. Climate change is considered one of the major threats to agricultural development in Africa. Decision-making level and farmers need efficient analytical tools to help them in decision making. Machine learning has become an impressive predictive analytical tool for large volume of data. It has been used in many domains such as medicine, finance, sport, and recently in agriculture. In this work, we propose three crop prediction models : Crop Random Forest, Crop Gradient Boosting Machine and Crop Support Vector Machine. We combine climate data, crop production data, and pesticides data to develop a decision system based on advanced machine learning models. Despite the poor availability of data related to agriculture in Africa, we were able to propose a decision system able to predict the crop yield at the country level in fourteen East African countries. Our experimental results show that the three proposed machine learning models fit well the crop data with a high accuracy R^2. The Root Mean Square Error ($RMSE$) and Mean Absolute Percentage Error ($MAPE$) associated to our models are very minimal because the agricultural prediction values are very close to reality. Our proposed models are reliable and generalize well the agricultural predictions in East Africa.

1. Introduction

Most African countries depend on rainfed agriculture for the cultivation of crops [19], which is a source of freshwater usage in agriculture. Recent changes in climate and rainfall variabilities have led to a decrease in food production. Additionally, poor farm management, misuse of natural resources inputs and knowledge deficiency of the use of agrochemicals are causing more harm to the environment increasing ecological footprint, which has distorted food security.

The demand for food is at an increasing rate due to population growth, which is expected to reach 10 billion in 2050, Africa and Asia are regions expected to be populated in the coming years [8]. In the past land area expansion was related to increases in agricultural produce in Africa [20]. However, recent population growth has led to a limited area for agriculture and competition of natural resources beneficiary to agriculture . This competition has led to a decrease in agricultural production hence putting the world at higher risk of hunger and malnutrition if measures are not taken to increase quality food production.

* Corresponding author.
E-mail addresses: rubby@aims.edu.gh (R. Aworka), cedric.l.saadio@aims-senegal.org (L.S. Cedric), adoniwilfried@gmail.com (W.Y.H. Adoni), jeremie.zoueu@inphb.ci (J.T. Zoueu), franckm@aims.ac.za (F.K. Mutombo), charles.kimpolo@nexteinstein.org (C.L.M. Kimpolo), t.nahhal@fsac.ac.ma (T. Nahhal).

https://doi.org/10.1016/j.atech.2022.100048
Received 5 December 2021; Received in revised form 31 January 2022; Accepted 27 March 2022

Africa tends to suffer more from food insecurity and malnutrition as in the shadow of COVID-19 about 21% of its population was affected by hunger in the year 2020 [14].

The aforementioned climate variabilities which are linked to climate change is one of the major threats to agricultural development in Africa [20]. Conversely, the agricultural sector is one of the causes of climate change as it generates 52% and 84% of global anthropogenic methane and nitrous oxide emissions [17].

Measures and policies are being employed to improved African agriculture to increase food production while mitigating and adapting to climate change and its effect [20]. Sustainable development in this sector will help improve the economy and livelihoods of Africa as it employs about 60% of the workforce [3]. In addressing the challenges of food and nutrition security and climate change mitigation and adaptation, the Food and Agriculture Organization (FAO) has come up with a concept known as climate-smart agriculture [12] using digital (smart) technologies. Such technologies include the internet of things (IoT), satellites, drones, sensors and robotics. Most of these smart technologies generate huge data and advanced analysis techniques allow more accurate exploitation of these data for an improvement of the performance. Machine Learning (ML) is an advanced predictive analytics tool that has been widely used in many areas such as medicine, finance, marketing and recently in agriculture to create a decision support system.

The use of ML in agriculture is promising as it assists farmers, policymakers and other stakeholders of agriculture in making intelligent decisions. Machine learning applications in agriculture will enhance the optimized use of resources for the cultivation and harvesting of crops and the production of livestock. Proper management of pests and diseases on-farm can lead to an increase in quality farm produce. Image processing was used to detect diseases and spread of disease on leaf and fruits, and weight of mango [10]. Additionally, use of ML has been employed to detect and classify laurel wilt disease from healthy leaves for an effective disease management [1]. Another use of ML in agriculture is in crop yield prediction. Forecasting crop yields enhances crop management, irrigation scheduling, and labor requirements for harvesting and storage [2].

Contributions In this paper, we develop a decision system using machine learning models to predict crop yield in East African countries. We used three machine learning models to predict at a country level four crops namely potatoes, beans, tea, and coffee in Burundi, Comoros, Kenya, Madagascar, Malawi, Mauritius, Mozambique, Rwanda, Seychelles, Uganda, Zambia, Zimbabwe, Eritrea, and Ethiopia. We proposed three advanced predicting algorithms based on machine learning (Crop Random Forest, Crop Gradient Boosting Machine and Crop Support Vector Machine) to predict crop yield in East African countries. We proposed a system decision that will assist farmers in making intelligent decisions regarding their farms and also organizations in decision-making. Hence, based on climatic data, crop production and the quantity of pesticide used in a specific year, farmers will be able to get an idea of the annual crop yield.

We have chosen these three models because they present better performance for multicriteria decision making. These learning models are supervised and the prediction approach is based on labelled data where the target is the values of the yield. Firstly, Random Forests leverage many limitations of decision trees and they allow generating a set of predicted crops yields that are going to be very close to reality because they have high variance. In addition, Random Forests will reduce the variance of how much the crop yield prediction will fluctuate during the training process.

Like Random Forests, Gradient Boosting Machine is effective in different ways. Random Forests are well switched for prediction on a dataset with a lot of statistical noise while Gradient Boosting Machine works efficiently when data is uneven. The last proposed model is based on Support Vector Machine, it has high generalization capabilities and can effectively handle non-linear data.

Structure of the paper The rest of this paper is structured as follow. In Section 2, we present some related work about machine learning and agriculture. In Section 3, we present the data analysis aspect of our data by showing the data acquisition process, data prepossessing and the data analysis results. In Section 4, we present our proposed crop yield prediction models followed by experimental results and discussion in Section 5. Finally, we conclude this work with some future directions.

2. Related work

Machine learning is a subfield of artificial intelligence that allows a system to intelligently learn from input data to make decisions, find patterns and relations without explicitly programming. Machine learning, in contrast to traditional statistical-based models, is a “black-box” with sophisticated functions that can handle complex interactions between predictors and the target values. The use of machine learning in agriculture will improve the efficient use of resources for agricultural cultivation and harvesting, as well as livestock production.

Alibabaei et al. [2] proposed a decision support system using two recurrent neural networks (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) with their respective extensions Bidirectional Long Short-Term Memory (BLSTM) and Bidirectional Gated Recurrent Units (BGRU) in estimating of end-of-season crop yields. In this work, BLSTM outperformed the other RNN models with MSE of 0.017 to 0.039. The proposed system will assist farmers to decide when and how to irrigate their farms efficiently.

A Constructed Deep Recurrent *Q*-Network (DRQN) which is a Recurrent Neural Network on top of Deep *Q* Network has been designed to estimate crop yield. The proposed model provides a solution that independently mines the non-linear mapping between the crop yield and the climatic, soil and groundwater parameters. The DRQN model is observed to outperform the other machine learning and deep learning models with an accuracy of 93.7% and improved error measure [7].

Sarijaloo et al. [16] have introduced different learning models including decision tree, gradient boosting machine, random forest, adaptive boosting, XGBoost and neural network were exploited to predict the yield performance of tested corn hybrids. On average XGBoost outperformed the other models with a 0.0524 root mean square error. The decision tree gave the worst performance for a reason that it is a weak learner and prone to overfitting. The selected model XGBoost was applied to untested combinations of inbred and testers, the model was able to identify hybrids with a high predicted yield that can be bred to increase corn production.

The authors in Khaki et al. [11] implemented a hybrid model which combines convolutional neural networks (CNNs), fully connected layer and recurrent neural networks (RNNs) to estimate the yield of corn and soybean. This model outperformed random forest (RF), deep fully connected neural networks (DFNN), and LASSO with a root mean square error of 9% and 8% of the respective average yield of corn and soybean. In this work, the CNNs were used to extract features from weather and soil datasets. The fully connected layer then combines the high-level features from the CNNs into the RNN including the yield data for the prediction analysis.

Predictive learning models have been proposed to classify sugarcane yield grade with input features such as plot characteristics, sugarcane characteristics, plot cultivation scheme and rain volume. The machine learning models used in this work are random forest and gradient boosting trees. The accuracies of both models were compared to two non-machine learning models and they outperformed these models with 71.83% and 71.64% of random forest and gradient boosting tree respectively. Additionally, the authors noticed that both machine and non-machine learning models analyze yield grade 3 incorrectly from the confusion matrices, which they suggested to explore in future and find the cause [6].

As far as Africa is concerned, unfortunately, just a few research works are based on approaches inspired by artificial intelligence for agricul-

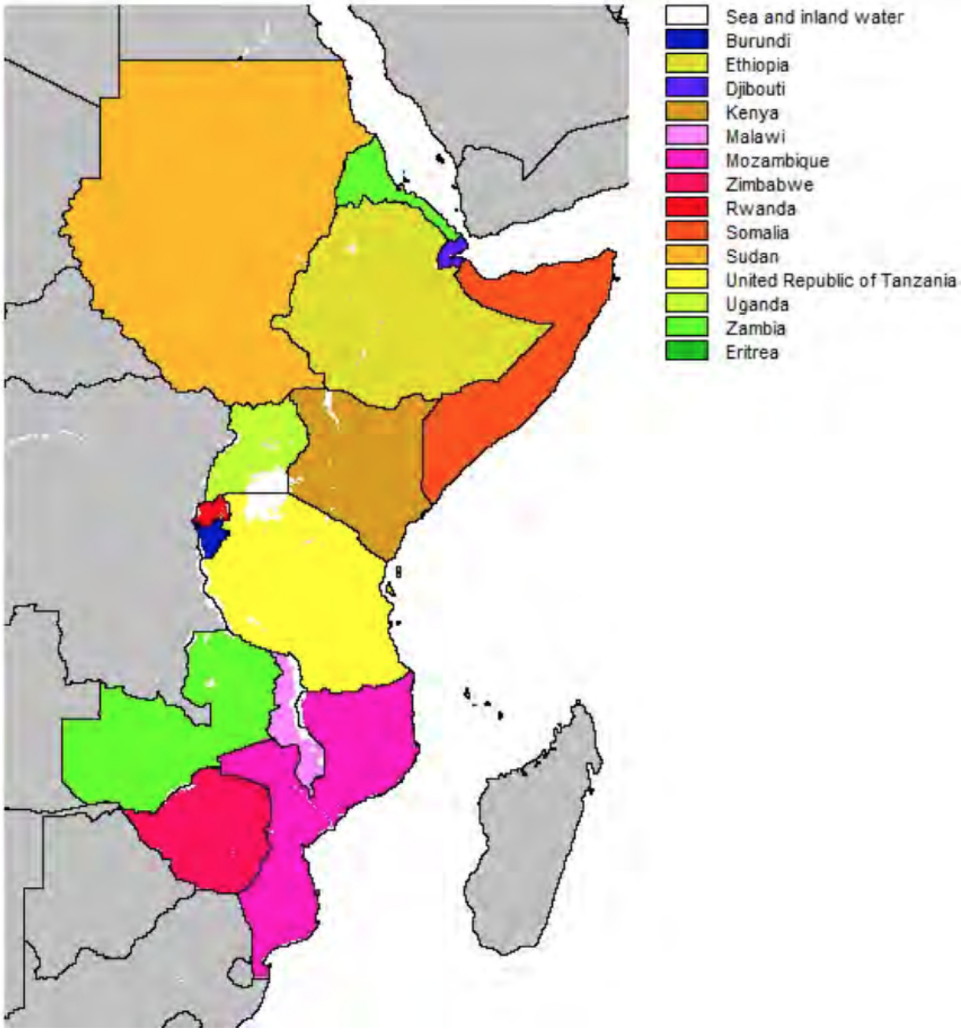


Fig. 1. Geographical area of East Africa, UNSTATS, 2013 Hengsdijk et al. [9].

tural forecasting. In East Africa, agricultural prediction techniques remain traditional because they are based on obsolete prediction models. Climate change is one of the main causes as it biases the results of seasonal trends. However, proper management of pests and diseases on-farm can lead to an increase in quality farm produce. Image processing was used to detect diseases and spread of disease on leaf and fruits, and weight of mango [10]. Additionally, the use of machine learning has been employed to detect and classify laurel wilt disease from healthy leaves for an effective disease management [1]. As a result, fresh ways of estimating agricultural production across East Africa in a timely and reliable manner at a cheap cost must be developed.

3. Material and methods

3.1. Study area

The research area spans fourteen countries of East Africa and contains around 1.8 million km^2. East Africa is the African continent's easternmost area, which is characterized by geographical or geo-political in many ways. Burundi, Comoros, Kenya, Madagascar, Malawi, Mauritius, Mozambique, Rwanda, Seychelles, Uganda, Zambia, Zimbabwe, Eritrea, and Ethiopia make up continental East Africa, according to the UN categorization of geographic areas (see Fig. 1) [9]. Due to geopolitical problems and lack of agricultural data, Sudan and the French overseas territories (Reunion and Mayotte) are not included in this research.

Eastern Africa's overall precipitation climatology is quite diverse. The average annual rainfall in much of the area ranges between 800 mm and 1200 mm [15]. Rainfall is considerably higher in the highlands and significantly lower in northeast Somalia and Kenya. It is concentrated in the north and northwest during the summer, but in the southernmost regions during the winter seasons. The region is affected by unstable seasonal rains that include the prevalence of serious drought regularly and farmers that are strongly reliant on the region's little rainfall. The boreal spring rains are predominant between March and May, whereas the autumn rains are prevalent between October and November. Monthly rainfall averages in these seasons, as well as in the boreal summer. Floods are also a problem in the area. Farmers are harmed by dramatic shifts, especially when they happen regularly [15]. The high consensus that climate change can have a large impact on rainfall, although estimates are somewhat varied, is causing severe concern for the country's future.

In Africa, there are two basic kinds of farming: garden crops are those that are produced largely from the roots and field crops are grown primarily from seeds. Beans and potatoes (sweet potatoes) are part of the family of crops most cultivated and consumed in East Africa. Despite their low nutritional quality, they are a dependable staple crop that may be easily preserved for future use. Cocoa, tea, coffee and grapes are Africa's main beverage crops. The top producers of tea are located in highland areas of East Africa. Climate change and the lack of suitable tools for these types of crops mean that they are heavily imported.

Farmers employ traditional techniques such as chemical fertilizer, particular seed kinds, and irrigation to preserve agricultural yields and

Table 1
Statistical summary of factors affecting crop productivity in East Africa.

	Pesticide (t)	N_2O (kt)	Rainfall (mm)	Temperature (°C)
mean	1009.00	18.97	89.72	22.91
std	1270.13	26.58	33.37	1.91
min	0.00	0.01	16.72	19.03
25%	88.00	3.60	68.15	21.96
50%	383.00	10.47	88.32	22.80
75%	1670.00	21.42	106.79	24.20
max	6753.00	142.56	211.71	28.09

fertility. This cultivation technique is no longer reliable because it is strongly affected by climatic changes to which the soil quality is added. Thus, it becomes difficult for a farmer to have visibility on his agricultural prediction [9]. Indeed, the latter have difficulty in adapting to new seasonal trends caused by climate change.

Table 1 presents a statistical summary of characteristic elements used in traditional agricultural techniques in East Africa. Temperature and rainfall are the variables dependent on climatic conditions while Pesticide and N_2O are strongly related to farmers' actions. In East Africa, temperatures vary between 18°C and 28°C with an average of 23°C. The standard deviation shows that there is no disparity in temperatures. This is due to the fact that the climate remains constantly warm. As for the rainfall, it varies between 17 mm and 212 mm with a remarkable standard deviation from the average. This is explained by the climatic variability. The majority of East Africa has a continental climate. It may therefore be classified into equatorial climatic regions, humid tropical area regions (savannah climate with dry winters) and temperate subtropical climate zones. Droughts and severe rainfall are caused by this fluctuation. Thus, these climate changes, unfortunately, force farmers to use many pesticides to boost their agricultural production.

3.2. Dataset acquisition: ETL process

We give a systematic procedure to how data was acquired till when it was fed into the proposed algorithms for yield prediction. The datasets from different sources were cleaned independently from each other, such that any missing data were removed. After cleaning the data, we merge the individual data into a complete dataset which will be used in the predictive analysis. The complete dataset contains fourteen countries namely Burundi, Comoros, Kenya, Madagascar, Malawi, Mauritius, Mozambique, Rwanda, Seychelles, Uganda, Zambia, Zimbabwe, Eritrea, and Ethiopia.

We used StreamSets data collector engine,[1] because of the size and the format of the data. It easily allows to build a real-time ETL process that consists of Extract, Transform and Load the three-dataset including crop production, agrochemicals and climate data from the Food and Agriculture Organization of the United Nations (FAO)[2] database. Soil data was inadequate hence were discarded during the analysis.

The crop production data contains the observed yield data of all crops and corresponding yield measured in hectogram per hectare hg/ha. We selected two major food crops and two cash crops consumed in Eastern Africa. Potatoes and Beans constitute the major crops while Coffee and Tea are the cash crops, between the years 1961 and 2019.

The crop production data contains the observed yield data of all crops measured in hectogram per hectare hg/ha. We selected two major food crops and two cash crops consumed in Eastern Africa between the years 1961 and 2019. Potatoes and Beans constitute the major crops while Coffee and Tea are the cash crops.

The agrochemicals dataset contains the total amount of pesticides used to control pests in the agriculture sector between 1990 to 2019 and is measured in Tonnes t.

[1] https://streamsets.com/.
[2] http://www.fao.org/faostat/en/#data.

The environment dataset contains greenhouse gas, nitrous oxide (N_2O), emission from agriculture soil, which is measured in kilotons from 1961 to 2019. It also contains the mean annual rainfall in mm and temperature in °C from 1901 to 2020. The list of features used in the dataset is summarized in Table 2.

3.3. Data exploration and analysis

We found that the target was containing outliers, to handle the issue, we applied a log transformation method. Fig. 2, shows the distribution of the target variables before and after we applied a log transformation technique. The first row in Fig. 2 shows the distribution of the target and the value of R^2 before the log transformation. We can see that the target variable is skewed to the right of standard deviation (σ), 52855.90 hg/ha. The coefficient of determination (R^2) of its probability plot is 0.6842. The second row in Fig. 2 shows the distribution of the target variable after the log transformation. The data is approximately normally distributed with $\sigma = 1.31$ and $R^2 = 0.9572$.

3.4. Feature engineering

In this step, we applied some techniques to get the data ready for training with machine learning. Categorical variables in Table 2 were encoded into numerical variables before feeding them into the machine learning models. The 'Crop' variables were encoded to numerical variables using the *OneHotEncoding* encoding algorithm. It is a widely encoding algorithm used in machine learning. In the feature selection process, we decided to drop the columns 'Year' and 'Country'. The other kept features were considered important for the training phase. The selected features were scaled to the same level using the MinMaxScaler function from the Scikit Learn library. In the phase of getting training, testing and validation data, we randomly split the dataset into a 70% training set and 30% testing set. The training set is used to train the machine learning models while the test is used to validate and evaluate the performance of the trained models to know how they fit new agricultural data. Three metrics were used to evaluate the performance of the models and they are coefficient of determination (R^2), root mean squared error ($RMSE$) and mean absolute percentage error ($MAPE$).

4. Proposed yield prediction models

In this section, we give a systematic procedure on how data was acquired till when it was fed into the proposed algorithms for yield prediction. Fig. 3 shows the architecture of our yield prediction system. We initially collected crop production and climate data from different sources and we merged them into a centralized database. We applied some prepossessing techniques, followed by exploration and analysis techniques to understand the knowledge hidden in the data. We applied some feature engineering to prepare the data for training. After the feature engineering step, we trained the model and used the tested data to evaluate the three machine learning models to produce a crops decision system. That decision system will be able to generalize new crop yield data for predicting a crop yield at a country-level per year in East Africa.

We propose three non-parametric models for our predictive analysis of crop yield. Two of the models are tree ensembles: Crop Random Forest (CRF) and Crop Gradient Boosting Machine (CGBM). The other model is linear, which is Crop Support Vector Machine (CSVM).

4.1. Crop random forest

The proposed Crop Random Forest (CRF) is a supervised machine learning algorithm based on Breiman [5] work. It is used for both classification and regression of crop yield problems. As shown in Algorithm 1, our model is based on an ensemble of decision trees. This allows us to make a strong prediction while minimizing the loss error between the prediction crop targets and the reality. It takes as input the number of

Table 2
Features description (Cat. = Catogory, Cont. = Continuous).

Name	Type	Detail
Country	Cat.	The selected East African countries, which are 14 in all
Crop	Cat.	The selected crops used which has 4 different crops
Pesticide (t)	Cont.	The amount of pesticides used
N_2O (kt)	Cont.	The value of nitrous oxide emitted from agriculture soil in
Average Rainfall (mm)	Cont.	The annual mean rainfall
Average Temperature (°C)	Cont.	The annual mean temperature
Yield (hg/ha)	Cont.	The target variable which indicates the yield of crops
Yield Log (log(hg/ha))	Cont.	The logarithmic transformation of the yield

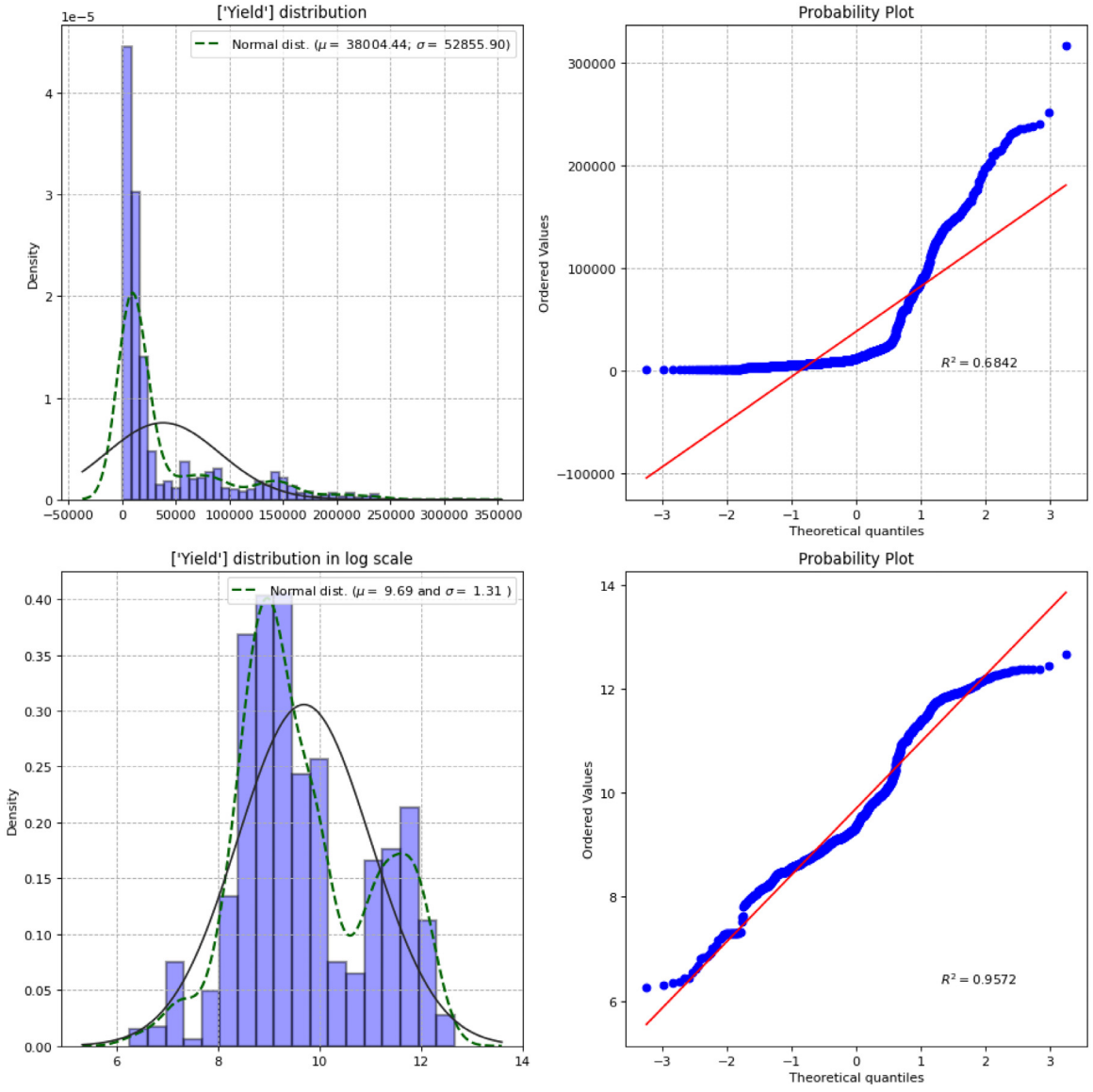


Fig. 2. Distribution plot with corresponding probability plot. The first row plots indicate distribution without log transforming. While the second row plots indicate distribution which has been log transformed.

trees B and the training data $D = \{(x_i, y_i)\}_{i=1}^{N}$ where N is the number of instances.

The accuracy of the prediction of our model depends on the number of crop trees B used in the ensemble. Each crop tree $\{T_b\}$ is built from a random sample of crop training data $D = \{(x_i, y_i)\}_{i=1}^{N}$ with replacement $D^* = (x_b, y_b)$, a technique known as bootstrap. The crop training data D is resampled with replacement into individual subsets D^* for each crop decision tree $\{T_b\}$ to fit on a specific sample data to predict the yield.

Then, decisions made by each crop tree $\{T_b\}$ are combined to make the final crop yield precision. This step is done by taking the average

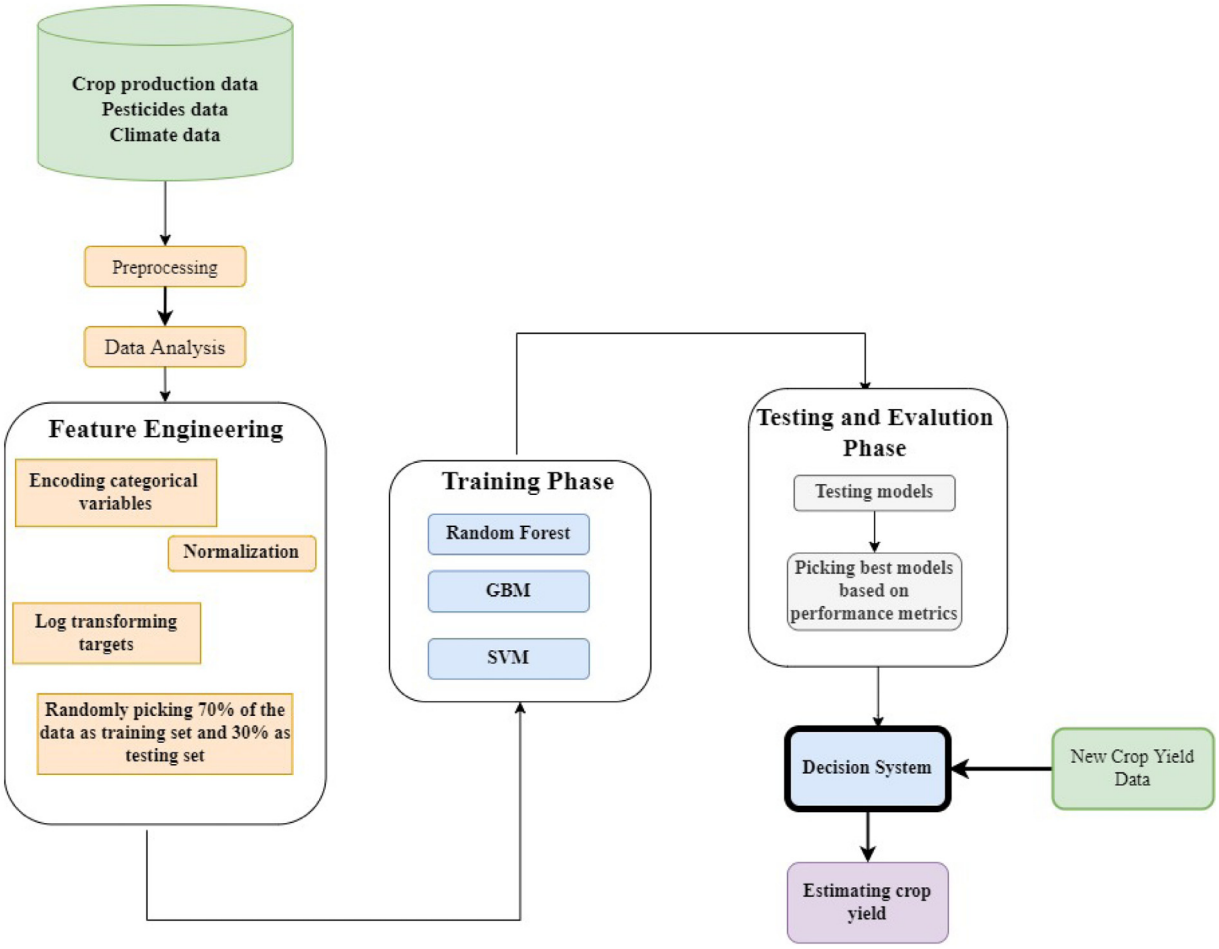


Fig. 3. Overall architecture of our crop yield prediction system.

Algorithm 1 Crop random forest.

Input:

For given crop training data set $D = \{(x_i, y_i)\}_{i=1}^{N}$

B = Number of crop prediction trees

initialization:

$\forall b, \{T_b\} = \emptyset$

$D^* = \emptyset$

building crop prediction trees:

For $b = 1, \cdots, B$

- Sample, with replacement from D, $D^* = (x_b, y_b)$
- Build crop regression tree $\{T_b\}$ on D^*

crop prediction:

$\forall$ new x, the crop prediction is given by averaging the predictions from all the individual regression trees on x:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

outputs of all crop trees associated with the model. For new agricultural input data x, the yield prediction is given by calculating the average yield prediction values from all the individual crop regression trees $\{T_b\}$ using the following function $f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$. Finally, the last step consists of validating the performance of the yield prediction model by computing the accuracy R^2, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) on the agricultural testing data.

4.2. Crop gradient boosting machine

The proposed agriculture crops yield Gradient Boosting Machine (GBM) model is a supervised learning algorithm that consecutively fits new crop models to provide a more accurate estimate of the yield prediction target. Our crop prediction model is among the variants of ensemble methods where one creates multiple weak crop models and combines them to get better performance as a whole. During the training process on agricultural data, we build iteratively new crop base-learners (weak-learners) to be maximally correlated with the negative gradient of the loss function associated with the whole ensemble of the crop training set [13].

The weak crop learners are fit in such a way that each new crop learner is trained on the residuals or the errors of the previous crop learner till the residuals are zero as the model improves. The final crop model aggregates the result of each step and thus a strong crop learner is achieved.

Fig. 4 shows an idea of how our crop gradient boosting works on an agricultural training dataset. We kept the default number of estimators (weak crop model) to 100. Each gray box represents a crop weak model which is a decision tree model. The prediction $P_i, i \in \{1, \ldots, 99\}$ is used as the input of the next weak crop model. The prediction process works sequentially and the current weak crop model P_i takes as input the yield prediction result of the previous weak crop model P_{i-1} until we reach

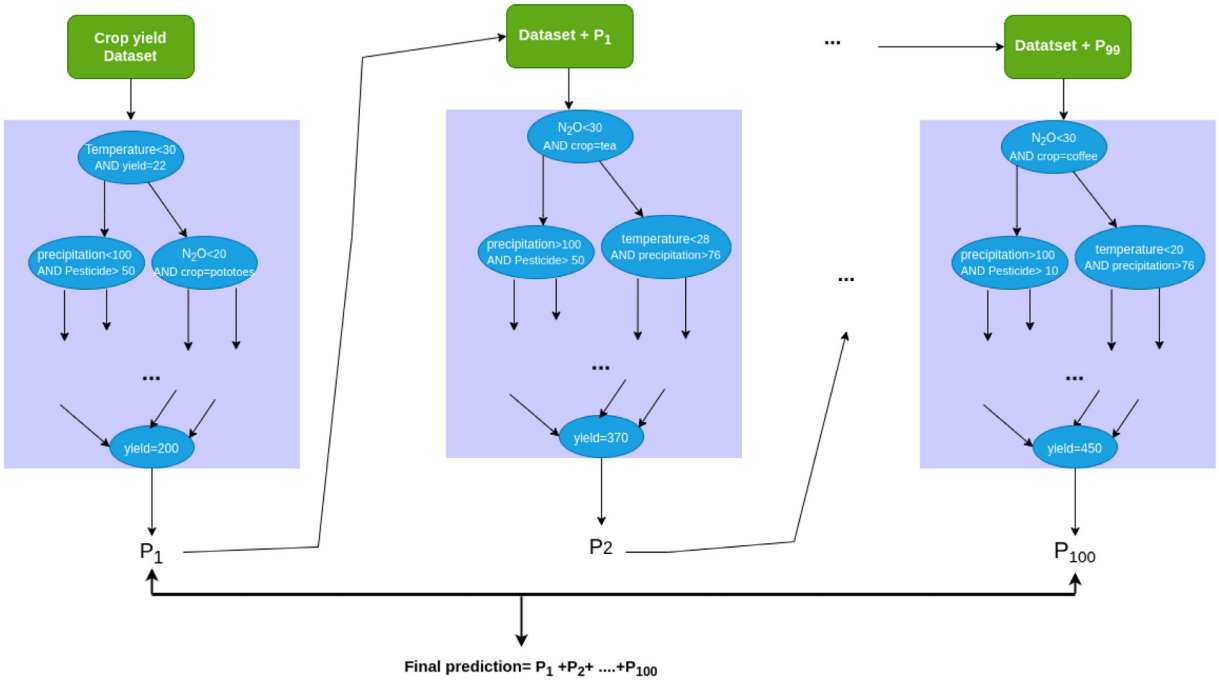


Fig. 4. Crop gradient boosting machine process.

the number of estimators. This technique helps to reduce the bias while handling the trade-off bias-variance and the risk of overfitting.

Finally, the last step consists of aggregating the intermediate yield predictions of each weak learner P_i in order to have the final yield prediction P such as $P = P_1 + P_2 + \ldots + P_{99}$.

4.3. Crop support vector machine

The third crop yield prediction model is based on Support Vector Machine (CSVM). It is a supervised learning model, which searches for hyperplanes in m dimensional space that distinctly classifies crop data points [4]. In this work, m is the number of independent features in the crop training data. The hyperplane is used to predict crop yield. Our goal is to build a hyperplane such that data points closest to the hyperplane are within the decision boundaries (hyper tube) a distance ε from the hyperplane. Hence the equation that satisfies our SVM in a linear situation is given by:

$$-\varepsilon \leq y_i - \boldsymbol{\omega}^T \cdot \boldsymbol{x_i} - b \leq \varepsilon, \tag{1}$$

where:

- ϵ is the user-defined error.
- x_i is the vector of crop features of the ith yield training instance.
- y_i: is the crop yield target of the ith training instance.
- ω is a vector of learning parameters that will fit to the crop yield prediction.

In practicality, not all data points are within the hyper tube. The crop yield dataset is not be linearly separable, hence slack variables ξ_i, ξ_i^* for each crop datapoint i are introduced to increase the number of errors to be committed [18]. The slack variables allow the violation of certain limitations. Hence the optimization problem can be defined as a minimization of the points outside the hyper tube while relaxing some constraints. We would like the amount of crop points inside the margin to be as little as feasible, and we also expect them to have as little penetration of the margin as possible.

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \\ &\text{subject to} && \begin{cases} y_i - \boldsymbol{\omega}\cdot\boldsymbol{x_i} - b \leq \varepsilon + \xi \\ \boldsymbol{\omega}\cdot\boldsymbol{x_i} + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{2}$$

In the optimization constraints, the slack variables ξ_i, ξ_i^* are introduced in two ways. Firstly, the slack variables specified the extent whereby the constraint on the x_i crop datapoint can be broken. Secondly, by including them, we hope to reduce the use of the slack variables at the same time.

5. Results and discussion

We performed our experimentation using a Linux computing system, consisting of Intel Core i5 CPU 16.0 and 8.0 GB of RAM. Anaconda environment was used to implement the models using Python 3.8.8. We randomly split the crop datasets into training and testing data. The training dataset was chosen as 70% of all datasets and was used to train the three proposed machine learning models. The testing dataset was used to evaluate the performance of the models and we chose 30% of all crop datasets. The imported libraries used for the analysis of the agriculture data include NumPy, Pandas, Scikitlearn, Matplotlib which are embedded packages in the Anaconda environment.

In the training process, we fixed the default hyperparameters of our models. For the Crop Random Forest (CRF) model we have set the number of trees in the forest to n_estimators = 10. This means that CRF includes ten crop decision trees in the forest. Concerning the Crop Gradient Boosting Machine (CGBM) model, we set the number of estimators to n_estimators = 100. This means that the model includes one hundred weak-learners involved in the yield prediction process. Finally, for the Crop Support Vector Machine (CSVM) we have fixed dimensional space to $m = 6$ since we have six features used for the prediction of the yields. Then, we fixed the value of the penalty for the prediction to $C = 1$ and the kernel function used in the training process to kernel = 'rbf'.

Table 3
Performance metrics for evaluating the crop models.

Models	R^2 (%)	RMSE	MAPE (%)	Runtime (s)
CRF	92.272	0.343	2.314	0.2754
CGBM	90.186	0.400	3.198	0.0826
CSVM	86.377	0.474	3.504	0.0310

We evaluate the performance of the three crop models based on the aforementioned metrics R^2, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The accuracy R^2 shows how well crop data fits the yield prediction, regression model. RMSE is based on Euclidean distance and is used to measure the gap between the prediction yields and the expected yield results. The metric MAPE is also known as mean absolute percentage deviation (MAPD); it measures the gap ratio between the obtained yield prediction and the expected yields.

Table 3 gives the evaluation results of each crop yield prediction model. We found that CRF has the highest $R^2 = 92.272\%$ and outperforms both CGBM and CSVM which have respectively an accuracy of 90.186% and 86.377%. This result means the CRF model fits the crop data better than CSVM and CGBM.

We can also see that CRF has the smallest root mean squared error compared to CGBM and CSVM. The RMSE that highlights the gap between predicted yields and the expected yields is respectively 0.342 for CRF, 0.4 for CGBM and 0.474 for CSVM. These results are not surprising since they are correlated with the score R^2 and the mean absolute percentage error on the whole crop data set.

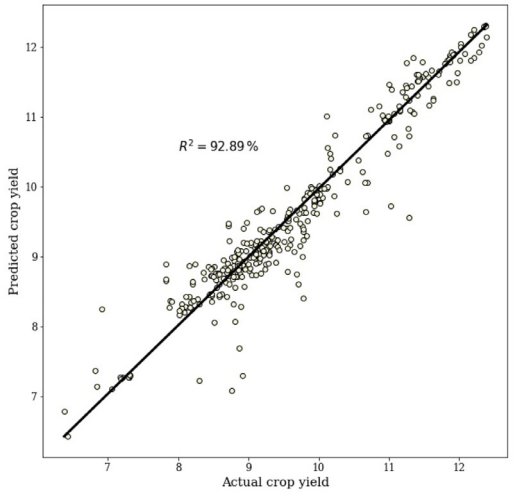
In terms of time complexity, CRF takes a little longer to execute. CSVM is 9x faster than CRF and 3x faster than CGBM. CSVM performs a yield prediction in 0.031 s whereas CRF and CSVM respectively make yield prediction in 0.0826 s and 0.2754 s.

Fig. 5 shows a comparison between the predicted and actual crop yield of each model. Fig. 5a shows that CRF results fit well the regression line. We also found that the tradeoff variance/bias was well handled with CRF. CGBM had quite comparable values but took into consideration the number of input features x_i. Fig. 5b shows that CGBM results did not fit as well the regression line as the CRF. Fig. 5c shows that CSVM prediction results do not fit well on the regression line like CGBM and CRF. These results also help to justify why CRF has the highest R^2 score.

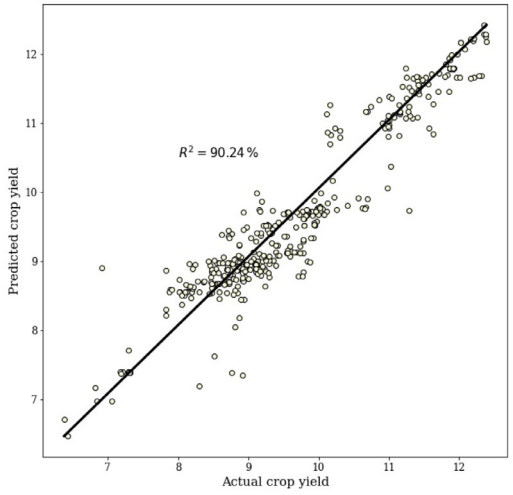
It is true that our three models present good results on the agricultural data used. Their complexities remain to be discussed because their performances depend on the turning of the learning parameters w and the hyperparameters such as the number of crop threes, the default number of estimators and the kernel function. Concerning CSVM, if C is big, there will be a considerable penalty for training prediction, resulting in a limited margin. If C is little, the penalty is low, there will be less penalty and the margin is big. By turning this hyperparameter, we can improve the crop yield prediction results. Having a larger number of crop trees and estimator improves the efficiency of CRF prediction but causes the increasing of the time complexity and the use of the highest hardware resources such as the CPU and RAM.

To have good decision making, we have to take into consideration the prediction value of the three proposed models. The main advantage of this decision-making proposal is that we will have a good prediction range with adapted evaluation metrics. We take the mean value of the three prediction values and the evaluation metrics. The three models were well trained on the East African agricultural dataset. The experimental results show that they are generalizable to the fourteen East African countries.

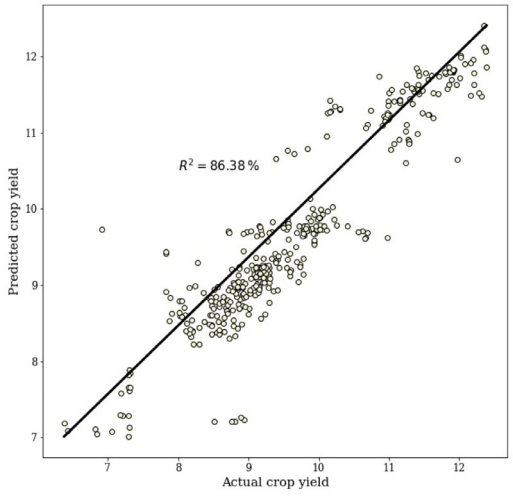
However, the performance of the models can be further improved by adding other more relevant features such as the availability of water, in quantity and quality (agricultural water), the climate and its meteorological variations (drought, hail, frost and other climatic calamities), the soil and its various characteristics, in particular its fertility, the domesticated plant and animal species.



(a) CRF



(b) CGBM



(c) CSVM

Fig. 5. Comparison between predicted and actual crop yield of each models.

6. Conclusion and future directions

In this work, we proposed a new agricultural decision system based on machine learning. Our main idea was to see how climate changes can impact the agricultural crop yield in East African Countries. We merged different data sources from climate, crop production and pesticides to build a crop yield prediction system for farmers and decision-makers. The prediction system help to predict an annual crop yield at the country-level of four crops in fourteen East African countries. After merging the different data sources, we applied some data processing techniques like data cleaning and missing value handling. We applied some exploratory and statistical analysis techniques to understand the behavior of the data. We also applied some feature engineering techniques like feature encoding, and normalization to prepare the data for the machine learning model. We trained three machine learning models: Crop Random Forest (CRF) model, Crop Gradient Boosting Machine (CGBM) model and Crop Support Vector Machine (CSVM) model.

The experimental results are conclusive. Our three agricultural prediction models are generalizable in the East African region. From the perspective of this work, adding others features such as 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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Abdulridha, Y. Ampatzidis, R. Ehsani, A.I. de Castro, Evaluating the performance of spectral features and multivariate analysis tools to detect laurel wilt disease and nutritional deficiency in avocado, Comput. Electron. Agric. 155 (2018) 203–211.

[2] K. Alibabaei, P.D. Gaspar, T.M. Lima, Crop yield estimation using deep learning based on climate big data and irrigation scheduling, Energies 14 (11) (2021) 3004.

[3] W. Bank, Policy brief: Opportunities and challenges for climate-smart agriculture in Africa, 2013.

[4] D. Boswell, 2002. Introduction to support vector machines. Department of Computer Science and Engineering University of California San Diego.

[5] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[6] P. Charoen-Ung, P. Mittrapiyanuruk, Sugarcane yield grade prediction using random forest and gradient boosting tree techniques, in: 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2018, pp. 1–6.

[7] D. Elavarasan, P.D. Vincent, Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications, IEEE Access 8 (2020) 86886–86901.

[8] Food FAO, et al., The future of food and agriculture–trends and challenges, Annu. Rep. 296 (2017) 1.

[9] H. Hengsdijk, A. Smit, J. Conijn, B. Rutgers, H. Biemans. 2014. Agricultural crop potentials and water use in East-Africa.

[10] M. Jhuria, A. Kumar, R. Borse, Image processing for smart farming: detection of disease and fruit grading, in: 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), IEEE, 2013, pp. 521–526.

[11] S. Khaki, L. Wang, S.V. Archontoulis, A CNN-RNN framework for crop yield prediction, Front. Plant Sci. 10 (2020) 1750.

[12] A. Meybeck, V. Gitz, Why Climate-Smart Agriculture, Forestry and Fisheries, Climate-Smart Agriculture Sourcebook, Food and Agriculture Organization of the United Nations. Retrieved on, 2014, pp. 14–103.

[13] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobotics 7 (2013) 21.

[14] F. Nations, I. Development, U. Fund, W. Programme, W. Organization, The state of food security and nutrition in the world 2021: transforming food systems for food security, improved nutrition and affordable healthy diets for all, The State of Food Security and Nutrition in the World (SOFI), Food & Agriculture Orgn, 2021. https://books.google.it/books?id=CnE5EAAAQBAJ.

[15] S.E. Nicholson, Climate and climatic variability of rainfall over eastern Africa, Rev. Geophys. 55 (3) (2017) 590–635.

[16] F.B. Sarijaloo, M. Porta, B. Taslimi, P.M. Pardalos, Yield performance estimation of corn hybrids using machine learning algorithms, Artif. Intell. Agric. 5 (2021) 82–89.

[17] P. Smith, D. Martino, Z. Cai, D. Gwary, H. Janzen, P. Kumar, B. McCarl, S. Ogle, F. O'Mara, C. Rice, et al., Greenhouse gas mitigation in agriculture, Philos. Trans. R. Soc. B 363 (1492) (2008) 789–813.

[18] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.

[19] S.P. Wani, T. Sreedevi, J. Rockström, Y. Ramakrishna, et al., Rainfed agriculture–past trends and future prospects, Rainfed Agric. 7 (2009) 1–33.

[20] T.O. Williams, M.L. Mul, O.O. Cofie, J. Kinyangi, R.B. Zougmoré, G. Wamukoya, M. Nyasimi, P. Mapfumo, C.I. Speranza, D. Amwata. et al., 2015. Climate smart agriculture in the African context.