Journal of Computer Science

Research Article

Crop Yield Prediction in Niger Using Machine Learning Techniques

[1]Mahaman Lawali Inoussa Garba, [1]Naroua Harouna, [1]Chaibou Kadri, [2]Maman Aminou Ali and [3]Hannatou Moussa

[1]*Department of Mathematics and Computer Science, Faculty of Science and Technology, Abdou Moumouni University (UAM), Niamey, Niger*
[2]*FUMA Gaskiya, Maradi, Niger*
[3]*National Agricultural Research Institute of Niger (INRAN), Maradi, Niger*

Article history
Received: 11-07-2024
Revised: 14-09-2024
Accepted: 01-10-2024

Corresponding Author:
Mahaman Lawali Inoussa Garba
Department of Mathematics and Computer Science, Faculty of Science and Technology, Abdou Moumouni University (UAM), Niamey, Niger
Email: mahamanlawali09@gmail.com

Abstract: Agriculture is the most impactful sector on Niger's economic growth. Stresses related to climate change and the environment are leading to a continuous decline in crop yield. Accurate crop yield forecasting is of paramount importance for farmers, government, policy makers, and other stakeholders to make most appropriate decisions regarding resource allocation and food security. This study examines the power of Machine Learning (ML) algorithms for crop yield prediction in Niger by introducing a combined machine learning and multi-level stacking ensemble (CML-MSE) model as a solution to improving the accuracy of crop forecasts using real farming data. Parameters such as agricultural, environmental and soil conditions are used to develop the CML- MSE approach to help farmers in making appropriate decisions for crop selection. The CML-MSE model consists of nine ML algorithms and three layers. The first layer uses K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), XGBoost (XGboost), Lightgbm and Catboost algorithms as base learners; the second layer acts as a meta-learner and use Support Vector Machine (SVM), Neural Network (NN) and Linear Regression (LR); and the third layer which uses LR algorithm. The prediction performance is assessed by using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the coefficient of determination (R^2), allowing comparisons to be made with the prediction of its individual models. The results obtained show that the proposed CML-MSE model offers the best performance compared with other models, with an MAE of 177.14 kg/ha, an RMSE of 243.71 kg/ha and an $R^2$of 65%. This work aims to provide Niger farmers with means to make appropriate decisions for improved food security.

Keywords: Artificial Intelligence, Machine Learning, Stacking Ensemble, Prediction, Agriculture

Introduction

Niger is a landlocked Sahelo-Saharan country whose economy is essentially based on agro-sylvo-pastoralism. The agricultural sector represents more than 40% of the national Gross Domestic Product (GDP) and remains the main sector providing employment (Habou *et al.*, 2016). However, the sector faces many challenges among which the most crucial is the need to feed a rapidly growing population of over 24 million. These challenges are accompanied by significant constraints such as climate change, water scarcity, shrinking arable land, soil degradation, etc. This situation poses a permanent threat to food security, sustainable development and economic growth. Understanding crop yield is essential in improving food security, mitigating the effects of climate change and meeting the growing need for agricultural products. Accurate forecast of crop yield is extremely important in agriculture. However, it poses a major challenge for farmers, policymakers and other stakeholders (Ed-Daoudi *et al.*, 2023; Jhajharia *et al.*, 2023). Indeed, crop yield is influenced by various factors, such as soil conditions, weather conditions (temperature, rain, humidity, etc.) and fertilizers. This makes accurate prediction difficult. It is therefore important to have accurate knowledge of crop yield variations when making decisions regarding agricultural risk management and yield forecasting (Jhajharia *et al.*, 2023).

Our traditional agricultural systems need to be analyzed and improved to increase sustainability and crop production. These systems assume that the field parameters are consistent, thereby discarding existing crop conditions. However, it is necessary to know how sensitive crop growth is to factors such as rainfall, soil, crop pests, etc. Unfortunately, producers, in Niger, face several challenges due to insufficient knowledge of soil management and obsolete farming practices. They most often favor the oldest, best-known or most fashionable crops/varieties without considering environmental and climatic factors. This can lead to underproduction. Therefore, they need to develop rapid and accurate crop yield forecasting systems to help them with appropriate decision-making, building resilience and savings. This is done through awareness of producers and communities in taking advantage of climatic conditions.

In recent years, with the digital revolution in the agricultural sector, various data sources are regularly collected and stored. These include data on soil conditions, weather stations and satellite data. Exploiting this information using Machine Learning (ML) algorithms has provided the opportunity to improve crop yield forecasts (Ed-Daoudi *et al.*, 2023). ML makes it possible to build models from data samples and gives the possibility of making decisions automatically based on past experiences (Waikar *et al.*, 2020). Machine Learning algorithms are able to analyze large amounts of data, identify trends, and make predictions based on the relationships between different variables. (Ed-Daoudi *et al.*, 2023). Various studies have used Machine Learning algorithms to predict crop yield in different countries and suggest that these algorithms have the capacity to significantly improve crop yield prediction (Ed-Daoudi *et al.*, 2023; Jhajharia *et al.*, 2023; Kuradusenge *et al.*, 2023; Sadenova *et al.*, 2023).

However, there are no studies that have used ML algorithms to predict crop yield in Niger. In fact, although in other countries some studies have investigated the effectiveness of ML techniques to predict crop yield, there are gaps in understanding their potential in the context of local crops in Niger. To fill these research gaps and provide application of Machine Learning techniques for crop yield forecasting in Niger, this paper proposes a model based on Combined Machine Learning (CML) approach with a multi-level stacking ensemble (MSE) for precise and effective forecasting of yield of various crops in Niger. The model uses 9 different ML algorithms, namely K-nearest neighbor (KNN), Decision Tree (DT), Random Forest (RF), XGBoost (XGboost), Lightgbm, Catboost, Support Vector Machine (SVM), Neural Network (NN) and Linear Regression (LR) as base learners and its performance is compared against that of each of its individual models (base learners). The dataset used consists of real data on crop yields in rural areas collected from 2017 to 2023 from the FUMA Gaskiya Maradi, Niger, and climatic data collected on the NASA website. Based on agricultural, environmental and soil conditions in Nigerien areas, our model suggests the most appropriate crop to plant. Therefore, the producer chooses the best crop offering the best yield thereby gaining through this new method.

Related Works

In their study Dhaliwal & Williams (2024) proposed a technique to predict sweet corn crop yield using ML models. They used historical data over a 26-year period (1992 to 2018) on field-level sweet corn yield, obtained from several U.S. vegetable processors. After data preprocessing, they implemented and compared several Machine Learning models, including principal components regression, partial least squares regression, multiple linear regression, regularized regression, multivariate adaptive regression splines and random forests, using 67 variables related to parameters such as weather, crop genetics, soil and management. The results show that the random forests model is better compared to all other models with the lowest RMSE of 3.29 Mt/ha and the highest R^2 of 0.77 between predicted and observed yields.

Raju *et al.* (2024) proposed a ML model based on the stacking ensemble method to increase the accuracy of yield forecasts using agroecological zone data. They used agriculture, fertilizer, rainfall, weather and soil data available for India over 25 years (1990–1991 to 2016–2017). This data includes attributes such as precipitation, humidity, temperature, nitrogen (N), phosphorus (P), potassium (K) and soil pH. After data preprocessing, they implemented a 3-layer staking ensemble model based on the following algorithms: decision tree (DT), random forest (RF), AdaBoost regression (ABR), XGBoost regression (XGBR), SVM and Naïve Bayes (NB). The 1st layer consists of all algorithms, the 2nd layer consists of XGBR, ABR and NB using the weighted stacking technique, and the 3rd layer which makes the final prediction consists of XGBR. They evaluated the model performance using F1 score, precision, recall, and specificity. The results obtained show an accuracy of 97.1%, an F1 score of 97.09%, a precision of 97.03%, a recall of 97.12% and a specificity of 100%.

The study conducted by Jhajharia *et al.* (2023) consists of estimating crop yield in the state of Rajasthan in India using various Machine Learning algorithms namely random forest, gradient descent, SVM, Long Short-Term Memory (LSTM) and LASSO regression. The data used comes from several different sources. Data from 1997 to 2019 on the most widely grown crops, including rapeseed and mustard, wheat, barley, maize, jowar onion and bajra, relating to 33 districts of the state were collected on the Rajasthan government's official website. This data has 3664 rows and 7 columns (state, district, area, season, crop, production and soil type) after removing invalid data and zero values. They then retrieved monthly rainfall data for each district from 1901 to 2002, 2004 to 2010 and 2011 to 2017 from the official documentation of the Rajasthan government. They calculated the rainfall for 2003 using the average rainfall data from 1901 to 2002 and that of 2018 and 2019 using the average rainfall data from 1901 to 2017. They used monthly rainfall data since the different crops do not have the same season. After data combining, preprocessing and standardization, they obtained a final data source composed of 71 columns which they subdivided into 98% for training data and 2% for test data. After training and testing the different algorithms, the results obtained show that the random forest algorithm is better with an R^2 of 0.963, an RMSE of 0.035 and an MAE of 0.0251.

Saraswat (2023) proposed a study to predict crop yield by comparing various Deep Learning and ML algorithms. The data used was collected on Kaggle. The dataset consisting of 2200 rows, contains information concerning 22 crops and 7 attributes including, precipitation, humidity, temperature (Temp), nitrogen (N), phosphorus (P), potassium (K) and pH. After data preprocessing, he divided the data into 2 parts: one for training made up of 80% of the data and the other for the test made up of the remaining 20% of the data. Then, he applied several classifiers, namely decision tree, random forest, Gradient boosting, Gaussian Naïve Bayes, logistic regression, Artificial Neural Network (ANN) and SVM. After training and testing the algorithms, the comparison was made based on accuracy. The results obtained show that Gaussian Naïve Bayes algorithm performs better with an accuracy of 99.54%.

In their study Kuradusenge *et al.* (2023) proposed a method to predict crop yield using ML models in Rwanda. They mainly used 2 data sources, namely: historical data on the yield of maize and Irish potato crops for the Musanze district concerning the agricultural year 2005/2006 to 2020/2021, and meteorological data concerning precipitation and air temperature. After data preprocessing, they analyzed the data using the random forest algorithm, polynomial regression and SVM. The result obtained shows that the random forest model is better with an MSE of 129.9 and 510.8 for maize and potato, respectively; and an R^2 of 0.817 and 0.875 for the same crops.

Sadenova *et al.* (2023) proposed a study in which they applied ML techniques and neural networks to predict the yield of legumes, cereals, fodder and oilseeds crops in Kazakhstan. They used processed images of experimental farms obtained from 2017 to 2022 from the Landsat-8 (EO Browser) and Sentinel-2 satellites. These data consist of production data comprising 1600 indicators with MSAVI and NDVI indices recorded at a frequency of once a week and weather data. They implemented and compared different algorithms, namely linear regression, Ridge regression, spurious regression, polynomial regression, multilayer perceptron, random forest and SVM. The results show the best average prediction accuracy of all crops of 85 and 82% for multilayer perceptron and polynomial regression respectively.

The study by Elbasi *et al.* (2023) consists of proposing a model for crop forecasting using Machine Learning algorithms. They used data on 22 crops consisting of 2200 records and attributes such as nitrogen (N) content ratio, temperature, soil pH value, precipitation, humidity, phosphorus (K) and potassium (P). They implemented and compared 15 various algorithms, namely the Bayesian network, naïve Bayesian classification, logistic regression, multilayer perceptron, simple logistic regression, IBK, KSTAR, LWL, Ada BoostM1, regression, decision tree, Hoeffding tree, J48, random forest and random tree. The results obtained show that the Bayesian network and Hoeffding tree algorithms offer better accuracies of up to 99.59% and 99.46% respectively.

Patil *et al.* (2023) presented an approach for crop selection and yield forecasting using ML algorithms. They used 3 data sources collected from the Kaggle platform, namely the “India Agriculture Crop Production” data source used for yield forecasting, the “District Wise Rainfall Normal” data source used to extract rainfall data by the district to predict yield and “Crop Recommendation” data source used for crop selection. After data preprocessing, they implemented different models namely naive Bayes, KNN, random forests and logistic regression and for crop selection considering parameters such as precipitation, humidity, temperature, N, P, K and pH; and linear regression, random forests and decision tree regression for crop yield prediction considering parameters such as city, crop, season and annual rainfall (in mm). The results obtained show that the random forest algorithm is the best for yield prediction with an R^2 of 0.96 and an MAE of 0.64, while the Naïve Bayes model is better for crop selection with an accuracy of 99.39%.

The approach proposed by Hasan *et al.* (2023) is based on ensemble ML methods for the prediction of suitable agricultural cultivation in Bangladesh. This method combines the algorithms of random forests, k-nearest neighbors and ridge regression to predict crop yield. The data used is from four agricultural organizations (BMD, BADC, BRRI and BBS) in Bangladesh for different seasons from 1969 to 2021. This data has parameters such as crop production, crop area, precipitation, max temperature, min temperature, cloud cover, wind speed and sunshine. After data preprocessing, they implemented different models namely SVM, naive Bayes, ridge regression, random forest and CatBoost and compared the results against their method. The results obtained show that their approach is better for forecasting the production of aus (MSE = 0.009 MSE and R^2 = 99%), aman (MSE = 0.92

and R^2 = 90%), boro (MSE = 0.246 and R^2 = 99%), wheat (MSE = 0.062 and R^2 = 99%) and potato (MSE = 0.016 and R^2 = 99%).

In their study Ed-Daoudi *et al.* (2023) proposed a model based on ML algorithms to improve crop yield predictions in Morocco. The data used include information on crop yields, precipitation, weather conditions and soil moisture captured from sensors. After data preprocessing, for crop yield prediction they evaluated the performance of various models, namely neural networks, random forests and decision trees, and compared them with traditional statistical models. The results show that the neural network offers the best results with an MSE = 0.10 and an R^2 = 90%.

Rao *et al.* (2022) conducted a study to find the best model in predicting crop yield based on soil nutrients and climatic conditions. Data used was collected from the Kaggle website and included 22 different crops and 7 attributes such as precipitation, humidity, temperature, N, P and K. After data preprocessing, various classification algorithms were compared, including KNN, random forests and decision trees. The results obtained show that random forests offer the best accuracy of up to 99.32%.

Materials and Methods

This work uses CML-MSE for crop prediction in Niger. The work examines pre-sowing crop production and yields based on real historical field data collected in the farming environment. First, the agricultural, soil and climatic data from FUMA Gaskiya and NASA website are given for preprocessing, cleaning and noise removal. The pre-processed data is split into two parts, namely training and testing containing 80% and 20% of the data, respectively. On the training and testing data, the proposed CML-MSE model is used for predicting suitable crops yield in the farming environment. This model consists of 3 layers and employs stacking ensemble learning. The first layer employs six ML regressors (KNN, DT, RF, XGboost, Lightgbm, and Catboost) as base learners. The average prediction results from the first layer are used as inputs to the second layer, which employs three ML regressors (SVM, ANN and LR) as meta-learner. The outcomes from the second layer are used as inputs to the third layer, which employs LR regressor for final prediction. The overall architecture of the model is shown in Figure 1.

The performance of various regressors and the proposed CML-MSE model is evaluated by using performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2).

Data Collection

Two datasets were used in this study. The yield dataset used comes from the FUMA Gaskiya of Maradi (through the women's field I, II, and III projects financed by McKnight Foundation) in Niger, and covers the period from 2017 to 2023. Real data was collected regularly in the farmers' fields who conduct demonstration trials for the project during the winter season using mobile phones and tablets. A Farmer Research Network Application (FRNA), developed by FUMA, has been used in collaboration with facilitators, technicians, and researchers in the collection and cleaning of the data. This dataset contains a total of 42,159 records and 10 attributes. The attributes are year, site, village, fertiliser, soil type, field longitude, field latitude, crops, variety and yield (in kg/ha).

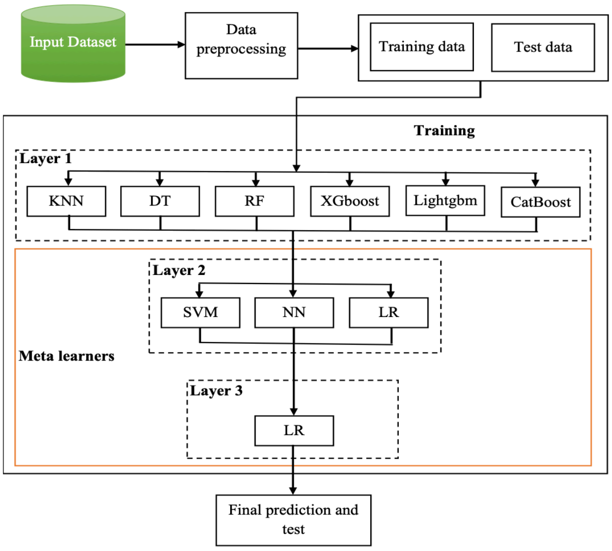

Fig. 1: System architecture diagram

The second dataset consists of annual climatology data specific to the different fields, collected between 2017 and 2023. The data was obtained from the NASA website (https://power.larc.nasa.gov) through monthly API calls and using the geographic coordinates (field longitude and field latitude) of the field as parameters. The dataset consists of 24,780 rows and 9 columns (Year, field longitude, field latitude, Average Temperature, Maximum (MAX) Temperature, Minimum (MIN) Temperature, Precipitation, Relative Humidity and Wind Speed).

Data Pre-Processing

The two datasets were merged based on field longitude, field latitude and year columns to create a single dataset which contains a total of 42,159 rows and 16 columns (year, site, village, fertiliser, soil type, field longitude, field latitude, crops, variety, yield, Average Temperature, Maximum (MAX) Temperature, Minimum (MIN) Temperature, Precipitation, Relative Humidity and Wind Speed). This dataset will be used for analysis.

The dataset contains unwanted data. Pre-processing is a procedure used to remove or correct data. The values of the attributes that are of string type (site, village, soil type, and fertiliser) cannot be read by some Machine Learning algorithms. Therefore, they are coded with numbers starting with zero. The labelling was done using the label encoding technique, which simply converts each value in the column into a number. Missing values and outliers can also have a negative impact on the accuracy of the model. Though there are several techniques in Machine Learning for dealing with missing values, the corresponding rows are simply deleted. This will certainly have an impact on the amount of data to be used to train the model, but the advantage is that it ensures clean and reliable data. The protocol column is not important, so it has been removed, and the yield column is used as target.

CML-MSE Model for Crop Prediction

Stacked generalization, is an integration technique that combines multiple base models via meta-models (Li *et al.*, 2023). Unlike algorithms such as clustering or traditional boosting methods, the stacking technique combines different base learners for model fusion. In the primary stage of this technique, the cross-validation method is used to convert the original features into secondary features, and then the transformed secondary features are routinely trained and fitted by meta-learners (Li *et al.*, 2023).

The generation process of the CML-MSE method proposed in this study (Figure 2) is described as follows. Initially, multiple base learners (layer 1) are trained simultaneously using the stacking technique on the dataset through 6 machine learning algorithms, namely KNN, DT, RF, XGBoost (XGboost), Lightgbm, and Catboost. Next, the predictions generated by these base learners are then used to create new attributes for the layer 2 meta-model. Here, 3 ML algorithms, namely SVM, ANN and LR. Finally, the output value of the layer 2 meta-model is taken as input to the layer 3 meta-model consisting of one ML algorithm (LR) for final prediction.

Regression Model Evaluation

When building a model, one tries to reduce the error of an algorithm. This is done by selecting an error measurement function, also called a cost function. In the case of a regression problem, the number of errors is not an appropriate criterion for evaluating performance, rather metrics designed to analyse continuous values are used. Thus, one would prefer to quantify the performance of a regression model in terms of the difference between predictions and actual values (Azencott, 2022). Three widely used measures to evaluate the performance of a regression model are used.

Mean Absolute Error

The Mean Absolute Error (MAE) is the absolute difference between the target value and the value predicted by the model. It is particularly interesting because of its robustness to outliers (Pedregosa *et al.*, 2012). MAE is a linear score, which means that all individual differences are equally weighted. It is not suitable for applications where more attention to outliers is desired. Mathematically it is defined by (Pedregosa *et al.*, 2012):

$$MAE\left(y,\hat{y}\right)=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples-1}}\left|y_i-\hat{y}_i\right| \tag{1}$$

where $\hat{y}_i$ is the predicted value of the *i*th sample, y_i is the corresponding true value and n_{samples} is the number of samples.

Root Mean Squared Error

The Root Mean Squared Error (RMSE) is the most widely used metric for regression tasks. It is defined as the square root of the mean square difference between the target value and the value predicted by the model, as shown in equation 2. It is preferable in some cases because the errors are first squared before the mean is calculated. This results in a strong penalty for large errors and implies that RMSE is useful when large errors are not desired. It indicates how close the predicted values are to the actual values. Therefore, a lower RMSE value means that the performance of the model is good. One of the key properties of RMSE is that the unit will be the same as the target variable (Swamynathan, 2019).

$$RMSE\left(y,\hat{y}\right)=\sqrt{\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples-1}}\left(y_i-\hat{y}_i\right)^2} \tag{2}$$

where $\hat{y}_i$ is the predicted value of the i^{th} sample, y_i is the corresponding true value and n_{samples} is the number of samples.

Coefficient of determination (R^2)

Another metric used to assess the performance of a regression model is the coefficient of determination or R^2. It helps compare the current model with a constant baseline and tells how much better the model is. It represents the proportion of variance explained by the independent variables in the model (Pedregosa *et al.*, 2012). It provides an indication of the goodness of fit, and therefore, a measure of how well new samples could be predicted by the model. Mathematically, it is defined by (Pedregosa *et al.*, 2012):

$$R^2\left(y,\hat{y}\right)=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{3}$$

Where $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ and $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\in_i^2$.

Hyper-Parameters Tuning

To assure good prediction results, parameters have been finetuned using the GridSearchCV tool in python, which performs hyperparameter tuning to determine the optimal values for a given model.

Cross Validation

To ensure the reliability of the results and avoid the risk of over-learning (overfitting), k-fold cross validation (k = 10) is performed on the training. We chose k = 10 because it is less biased than k = 5, although it has a higher computational cost.

Standardization

The dataset contains numerical values that are different in scale. They are Standardized to have a common scale while building models using StandardScaler tool in python.

Results and Discussion

The testing was conducted on hardware comprising a PC running Windows 11 with an Intel Core i7-13700H CPU and 16 Gigabytes of Random Access Memory.

Based on the CML-MSE model, the crop yields in Niger were estimated. The proposed CML-MSE model was trained and tested using crop recommendation dataset consisting of agricultural and climatic data. The agricultural data was collected from local farmers, by a non-governmental organization called FUMA Gaskiya which is operating in the rural areas of Maradi in the republic of Niger, from 2017 to 2027 and the climatic data was collected from the NASA website for the same period. The performance of the CML-MSE was compared with those of single learner models. The metrics used to measure the performance of the models are mainly MAE, RMSE and R^2. The results of the evaluation metrics of the crop yield estimations by the KNN, DT, RF, XGboost, Lightgbm, Catboost, SVM, NN, LR and CML-MSE models are shown in Table 1.

Table 1: Model performance

Model	MAE	RMSE	R^2
KNN	198.15	271.06	0.57
DT	188.25	258.46	0.61
RF	185.50	253.81	0.62
XGboost	182.10	248.33	0.64
Lightgbm	180.12	247.02	0.64
Catboost	182.69	248.17	0.64
SVM	211.96	293.21	0.50
NN	241.87	311.72	0.43
LR	333.38	407.34	0.03
CML-MSE	177.14	243.71	0.65

Table 2 shows the average train and test time for each model. Since the time varies with each program execution, we trained and tested each model 5 times with their optimal parameters values and calculated the respective average time. It can be noticed in Figure 2 that the CML-MSE model training time is greater than others models because it is the combination of 9 models. The CML-MSE model makes a final prediction using the predictions of the models that constitute it, which increases the training time. Figure 3 shows that the CML-MSE and SVM models have higher test time than the others.

Considering the three evaluations metrics (MAE, RMSE and R^2), Table 1 shows that the overall performance of the CML-MSE model is significantly better than those of other models. The MAE of CML-MSE is 177.14, which is 21.01, 11.11, 8.36, 4.96, 2.98, 5.55, 34.82, 64.73 and 156.24 lower than those of KNN, DT, RF, XGboost, Lightgbm, Catboost, SVM, NN and LR, respectively, for an average reduction of 34.42. The RMSE of CML-MSE is 243.71, which is 27.35, 14.75, 10.1, 4.62, 3.31, 4.46, 49.5, 68.01 and 163.63 lower than those of KNN, DT, RF, XGboost, Lightgbm, Catboost, SVM, NN and LR, respectively, for an average reduction of 38.42. The R^2 of CML-MSE is 0.65, which is 0.08, 0.04, 0.03, 0.01, 0.01, 0.01, 0.15, 0.22 and 0.62 higher than those of KNN, DT, RF, XGboost, Lightgbm, Catboost, SVM, NN and LR, respectively, for an average increase of 0.13.

Table 2: Train and test times for models

Model	Train time (second)	Test time (second)
KNN	1.762	0.127
DT	1.131	0.002
RF	198.242	0.492
XGboost	3.531	0.083
Lightgbm	3.399	0.031
Catboost	267.189	0.053
SVM	2505.364	36.291
NN	3479.676	0.045
LR	0.225	0.016
CML-MSE	6929.597	34.478

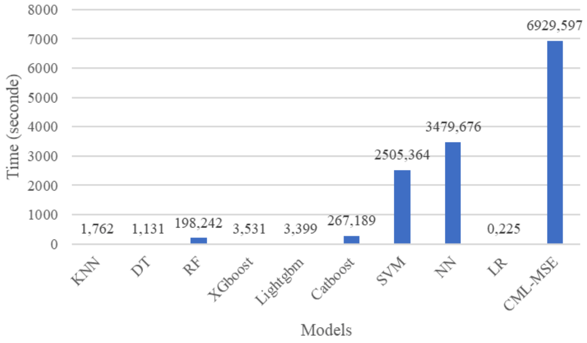

Fig. 2: Train time comparison of all models

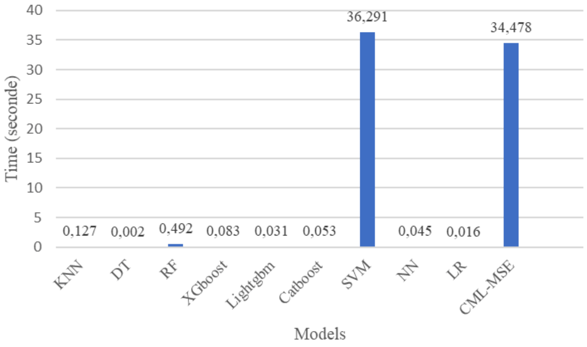

Fig. 3: Test time comparison of all models

In summary, it can be seen in Figures 4, 5 and 6 that the proposed method (CML-MSE) performs better than other single models, with lowest MAE and RMSE and highest R^2, respectively. Moreover, it reduces bias and variation, increases model variety, and improves the interpretability of the final forecast by merging the predictions of 9 base models. Therefore, it is the best model for predicting the crop yield based on our data with high prediction accuracy and strong generalizability.

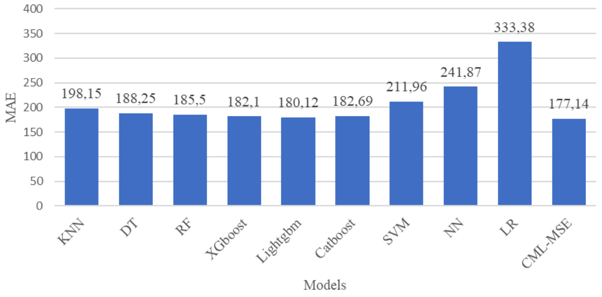

Fig. 4: MAE comparison of all models

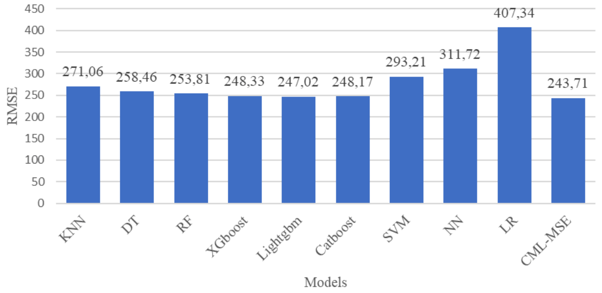

Fig. 5: RMSE comparison of all models

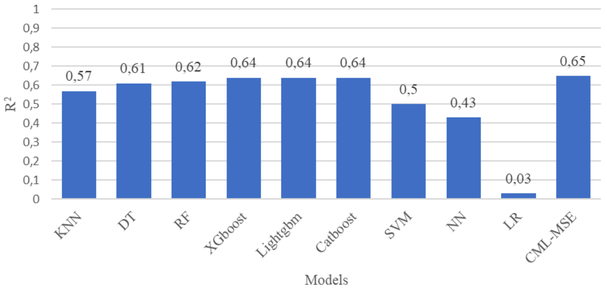

Fig. 6: R^2 comparison of all models

Conclusion

These results constitute a contribution to the body of literature regarding the use of machine learning techniques for crop yield forecasting to address challenges related to resource allocation and food security in Niger and other countries. Moreover, in terms of innovation, this is one of the first study evaluating the performance of several Machine Learning algorithms for predicting local crop yield in Niger and having proposed a combined machine learning with a multi-level stacking ensemble (CML-MSE) model to improve the quality of prediction. Therefore, this study can be used as a basis for further research in this area, which may lead to more accurate and efficient methods for crop yield prediction in Niger and other countries.

This article is a contribution to the effort of improving crop yields in Niger through the application of Machine Learning techniques. The prediction of various crops in rural areas using the proposed CML-MSE model is successfully implemented with an increase in accuracy rate. The evaluation of the CML-MSE, KNN, DT, RF, XGboost, Lightgbm, Catboost, SVM, NN models using MAE, RMSE and R^2 metrics, showed that the proposed CML-MSE model was the best among the ten models, with the lowest MAE and MSE, and the highest R^2. Based on agriculture, soil and climatic factors, the proposed model successfully provides orientation to farmers on the type of crop to be grown on a given field.

In order to improve the performance of the proposed CML-MSE model, future works can consider larger amount of data, various food crops and other important factors such as crop pests. Furthermore, optimization techniques can also be used.

Acknowledgment

This research is supported by FUMA Gaskiya Maradi, Niger, through the women's field project.

Data was obtained from the FUMA Gaskiya Maradi, Niger.

References

Azencott, C. A. (2022). *Introduction au Machine Learning-2e ed*.

Dhaliwal, D. S., & Williams, M. M. (2024). Sweet Corn Yield Prediction Using Machine Learning Models and Field-Level Data. *Precision Agriculture*, *25*(1), 51-64. https://doi.org/10.1007/s11119-023-10057-1

Ed-Daoudi, R., Alaoui, A., Ettaki, B., & Zerouaoui, J. (2023). Improving Crop Yield Predictions in Morocco Using Machine Learning Algorithms. *Journal of Ecological Engineering*, *24*(6), 392-400. https://doi.org/10.12911/22998993/162769

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop Prediction Model Using Machine Learning Algorithms. *Applied Sciences*, *13*(16), 9288. https://doi.org/10.3390/app13169288

Habou, Z. A., Boubacar, M. K., & Adam, T. (2016). Les Systèmes De Productions Agricoles Du Niger Face Au Changement Climatique: Défis Et Perspectives. *International Journal of Biological and Chemical Sciences*, *10*(3), 1262-1272. https://doi.org/10.4314/ijbcs.v10i3.28

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble Machine Learning-Based Recommendation System for Effective Prediction of Suitable Agricultural Crop Cultivation. *Frontiers in Plant Science*, *14*, 1-18. https://doi.org/10.3389/fpls.2023.1234555

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop Yield Prediction Using Machine Learning and Deep Learning Techniques. *Procedia Computer Science*, *218*, 406-417. https://doi.org/10.1016/j.procs.2023.01.023

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture*, *13*(1), 225. https://doi.org/10.3390/agriculture13010225

Li, Q., Xu, S., Zhuang, J., Liu, J., Zhou, Y., & Zhang, Z. (2023). Ensemble Learning Prediction of Soybean Yields in China Based on Meteorological Data. *Journal of Integrative Agriculture*, *22*(6), 1909-1927. https://doi.org/10.1016/j.jia.2023.02.011

Patil, P., Athavale, P., Bothara, M., Tambolkar, S., & More, A. (2023). Crop Selection and Yield Prediction using Machine Learning Approach. *Current Agriculture Research Journal*, *11*(3), 968-980. https://doi.org/10.12944/carj.11.3.26

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Raju, C., D.V., A., & B.V., A. P. (2024). CropCast: Harvesting the Future with Interfused Machine Learning and Advanced Stacking Ensemble for Precise Crop Prediction. *Kuwait Journal of Science*, *51*(1), 100160. https://doi.org/10.1016/j.kjs.2023.11.009

Rao, M. S., Singh, A., Reddy, N. V. S., & Acharya, D. U. (2022). Crop Prediction Using Machine Learning. *Journal of Physics: Conference Series*, *2161*(1), 012033. https://doi.org/10.1088/1742-6596/2161/1/012033

Sadenova, M., Beisekenov, N., Varbanov, P. S., & Pan, T. (2023). Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan. *Agriculture*, *13*(6), 1195. https://doi.org/10.3390/agriculture13061195

Saraswat, T. (2023). Crop Prediction Using Machine Learning and Artificial Neural Network. *Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)*, 561-568. https://doi.org/10.2991/978-94-6463-196-8_43

Swamynathan, M. (2019). *Mastering Machine Learning with Python in Six Steps*. https://doi.org/10.1007/978-1-4842-4947-5

Waikar, V. C., Thorat, S. Y., Ghute, A. A., Rajput, P. P., & Shinde, M. S. (2020). Crop prediction based on soil classification using machine learning with classifier ensembling. *International Research Journal of Engineering and Technology (IRJET)*, *7*(05), 4857-4861.