Cascading Ensemble Machine Learning Algorithms for Maize Yield Level Prediction

Hayam R. Seireg, Yasser M. Omar, Fathi E. Abd El-Samie, Adel El-Fishawy, and Ahmed Elmahalawy

***Abstract*—Climate change is destroying many crops around the world. This paper aims to anticipate maize yield levels based on climatic conditions, which would aid in making proper decisions regarding the connected sectors for business planning and yield level prediction. This paper presents two novel models that combine five machine learning algorithms with different techniques. Selecting six months of the climate features for the four regions in China. The first proposed model (FPM) consists of K Nearest Neighbors, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Decision Tree and Quadratic Discriminant Analysis (KMBDQ) that come together in a cascading topology (CT) to feed each other by taking the new prediction and removing the old previous prediction from the input features at each stage. The second proposed model (SPM) also uses the same five machine learning algorithms with different approaches. In this model, the prediction of each machine learning algorithm is used as a feeder to each other in the form of CT without removing any prediction. The performance evaluation of the proposed models was demonstrated and compared with several classifiers using the same dataset. The evaluation was based on metrics such as accuracy, sensitivity, precision, and F1 score. The results showed that the SPM had the highest prediction accuracy of 79.6%, which was a 29.6% increase compared to the first classifier in the model. The SPM also had an 11.1% improvement compared to the FPM and a 10.2% increase compared to the best among the many techniques used. In addition, computation time comparisons were conducted.**

***Keywords*—K Nearest Neighbors Classifier, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Decision Tree Classifier, Quadratic Discriminant Analysis.**

I. INTRODUCTION

Agricultural advances have made rapid improvements, but crop losses are still increasing at a surprising rate [1]. Agriculture is the most important field that meets the requirements of people and governments. Studies have already shown that climate change is affecting both crop productivity [2] and food security [3]. Anomalous alterations in temperature have a negative effect on agriculture, and the most critical stage in crop production is the flowering stage. Sudden changes in temperature and heat waves can impact this stage, leading to a reduction in crop yields [4-5]. In the rural planning process, agricultural financial specialists require simple and accurate methods to predict crop yield based on input features such as maximum and minimum temperature, and precipitation for optimal planting times [6]. China has implemented exceptional methodologies to contribute to rural development and the food industry.

Machine learning methods have the potential to save various crops from damage in the future. China is focusing on utilizing machine learning algorithms [7] for crop yield forecasts to achieve the Food and Agriculture Organization (FAO) goals.

One of the most important objectives of the FAO is to prevent hunger and malnutrition worldwide. Countries should focus on stepping up their efforts to address the economic and climatic challenges. Machine learning techniques are widely used in the agricultural system to predict the yield of each crop [8]. Several approaches have been developed over a long period of time, with varying degrees of success, for assessing and modeling agricultural yields [9-10]. Machine learning is valuable for making informed decisions in farming. These algorithms have the potential to extract the complex relationships found within agricultural field datasets [11].

Researchers use ensemble machine learning algorithms to improve the accuracy of a data-driven model [12-15]. Although machine learning algorithms have made significant advances in numerous areas, they still face certain barriers when it comes to data [16]. They can only satisfy exact expectations based on the quality of the data and model selection [17]. Several issues within the accumulated dataset, such as missing values, the presence of irrelevant features, a significant amount of noise, and the existence of outliers, may limit the predictive power of models [18-19].

Researchers have conducted several studies on predicting crop yield using various strategies [20-21]. Most researchers have shown that an ensemble of machine learning classifiers is more accurate in prediction than individual classifiers [22].

Manuscript received [01 Sept. 2022]; revised [09 Feb. 2023]; accepted [07 May 2023]. Date of publication [25 May 2023].

Hayam R. Seireg is with the Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt (e-mail: seireghayam9@gmail.com).

Yasser M. Omar is with the Department of Computer Science, Faculty of Computing and Information Technology, AASTMT, Cairo, Egypt (e-mail: dr_yaser_omar @yahoo.com).

Fathi E. Abd El-Samie is with the Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt (e-mail: fathi_sayed@yahoo.com).

Adel El-Fishawy is with the Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt (e-mail: aelfishawy@hotmail.com).

Ahmed Elmahalawy is with the Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt (e-mail: a_elmhalaway@hotmail.com).

For this reason, it is essential to focus on building an improved ensemble classifier to achieve the highest level of accuracy [15]. Bagging and boosting are the two main techniques in ensemble machine learning algorithms. Boosting combines multiple weak models sequentially to create a strong model that can better predict outcomes. Bagging, on the other hand, trains multiple models in parallel on randomly selected subsets of the training data and averages their outputs to create the final prediction. The choice between boosting and bagging depends on the specific problem being solved and the data being used [23].

Scientists have developed several experimental and mathematical yield modelling methods for various crops [9, 24]. Large datasets were collected over several years from the site for the purpose of precision agriculture, and a variety of data analysis techniques were used, including agronomic methods [25].

Ensemble machine learning is the ideal approach for producing a single predictive model that achieves powerful results. The majority voting technique, which follows democratic principles by selecting the highest number of votes, does not require any prior knowledge of the problem and is used in ensemble machine learning [26]. Predicting corn yield in the United States using ensemble machine learning with different techniques has been shown to achieve accurate results [27]. An ensemble classifier is used to handle challenging issues in real-world applications that a single classifier cannot solve [28-29].

In this research paper, we developed two novel proposed models that could improve maize yield level prediction using temperature and rainfall data only, without considering soil and fertilizer data, as soil can be replaced with artificial soil or hydroponics (soilless plant culture). Additionally, we tested multiple classifiers in our study. To predict maize yield levels in China from climate data only, we implemented two novel proposed models that combined five machine learning algorithms in CT. These two architectures were combined in different ways to improve accuracy, and the proposed models were compared with many other techniques. The SPM achieved the highest overall accuracy compared to the other techniques used and FPM. The main contribution of this work is the novel combination of five machine learning algorithms. The SPM was implemented by feeding the prediction of each machine learning algorithm in a series form, where each algorithm predicts the whole output of the dataset and adds its prediction as a feature to the dataset in each stage. The order of the machine learning algorithms is crucial and includes KNN, MNB, BNB, DTC, and QDA, with the final classifier being Quadratic Discriminant Analysis.

This paper is organized as follows: Section 2 presents related studies. Section 3 discusses the data used and preprocessing, machine learning algorithms, and the two novel proposed models. Section 4 evaluates performance metrics for the two proposed models and the many techniques used. Section 5 explains the experiments and results of the two proposed models compared to the many techniques used. Section 6 discusses the findings and research gaps. Finally, Section 7 briefly concludes the paper.

II. Literature Review

Previous researches forecast crop yields and soil types using a diversity of classification and regression algorithms.

For evaluation accuracy, five machine learning algorithms were tested: LASSO regression, elastic net (EN), extreme gradient boosting (XGBoost), ridge regression and random forest (RF). From observing the results, the RF models achieved the highest accuracy compared to the other machine learning algorithms for predicting maize yield [30].

S. Motia et al. [31] use the agricultural soils dataset to test the accuracy of three well-known classification models, such as K-Nearest Neighbors (KNN), Naive Bayes (NB) and Decision Tree (DT). Evaluate the ensemble classifier (EC) that is proposed by fusing the above three classifiers. The results showed that EC has the highest accuracy of 84% compared to NB (72.90%), KNN(73.56%) and DT (80.84%).

The multiple linear regression algorithm was used to predict maize yield, but it achieved the lowest R-squared values of 0.0089, 0.0223, 0.0209, and 0.0207 for the four regions of Southwest China, Huanhuaihai, North China, and Northeast China, respectively [32].

A soil-based machine learning comparative analysis framework (SMLF) [33] predicts crop yield, assesses the impact of soil properties and climatic factors and identifies class designations (high, low and medium) for crop yield prediction. Comparing the classifiers including soil properties and climatic factors, the results showed that the fusion of both feature vectors significantly improved the performance of the crop yield prediction system.

The prediction of palm oil yield has been studied in [34] to determine the best technique for forecasting palm oil yields. Numerous machine learning algorithms based on regression techniques, such as RF, SVR, and ANN, have been found to be highly useful in predicting palm oil yield. Additionally, ensemble methods have been used instead of a single algorithm [35-37]. To increase efficiency, a variety of strategies should be studied.

Most of the studies are based on predicting the crop yield using individual machine learning but, in our study, we implemented a novel combination of machine learning using two approaches for maize yield level prediction.

III. Materials and methods

A. Dataset

Thirty years of maize yield and weather data were collected from two databases for this investigation. They have been combined into a single dataset. Twenty-four years of maize crop data were gathered from the National Agricultural Statistics (NAS) by County and six years from the Provincial Statistical Yearbook, which is issued each year. The China Meteorological Data Service System provided the climate data, while the statistical data with land cover information for China's Northeast, North, Southwest, and Huanghuaihai regions were used in this study. These regions were selected to account for the varying latitudes and climates found within China. The database contains 73,000 census observations in 2,463 counties over three decades. These data were used to interpret yield data at three variable spatial levels: county, province/district/municipal, and farming system. The Maize agriculture dataset of China consists of climate data, including maximum (tenths of degrees C) and minimum temperature (tenths of degrees C) and monthly rainfall (tenths of mm), as well as annual maize yield (ton/hectare). It covers 15.5 million hectares of agricultural maize land [32].

B. Preprocessing

The actual values of maize yield were transformed into ordinal data to obtain the maize yield level (High or Low). The yield levels were determined based on the median yield and used as the output target. The high yield ranged from 4.58 t/ha to the maximum yield, while the low yield varied from 0.037 to 4.579 t/ha. The machine learning algorithm was used to predict the label of the maize yield level, which has two levels (High, Low). The years were arranged in ascending order to help distribute the high and low yield levels in the dataset. Outlier yield data, station, year, and maize yield were removed from the dataset. Data Cleaning is also known as scrubbing. This task involves filling in missing values and smoothing or removing noisy data and outliers. Missing data values were common in some counties and years. For missing data in the 30-year dataset, the yearly values from the preceding and following years were averaged and used to interpolate the missing values. For missing data at the beginning or end of the period, estimated trends from neighboring counties were used for imputation. Data outliers, which were caused by inaccuracies, gaps in survey statistics, weather changes, or pest infestations, with extremely high or low yield values were found and excluded from the dataset for some countries [32].

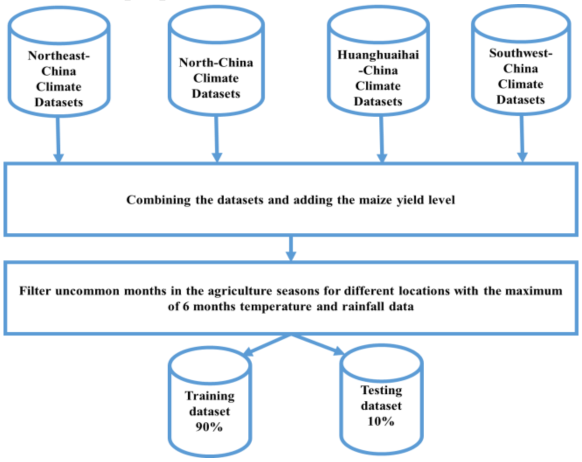
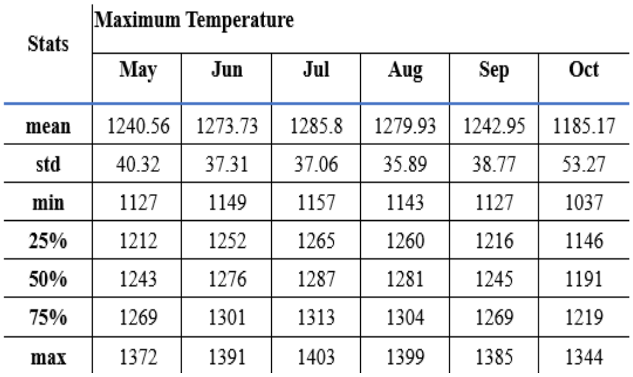

Northeast-China Climate Datasets
North-China Climate Datasets
Huanghuaihai-China Climate Datasets
Southwest-China Climate Datasets
Combining the datasets and adding the maize yield level
Filter uncommon months in the agriculture seasons for different locations with the maximum of 6 months temperature and rainfall data
Training dataset 90%
Testing dataset 10%

Fig. 1. Block diagram for the pre-processing.

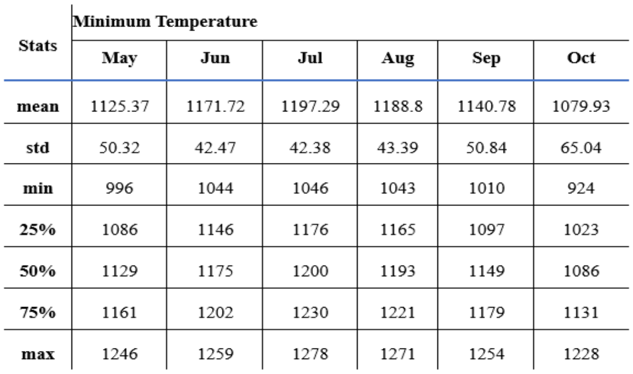
Fig.1 was used to filter out 30 weather features, keeping only the common maize agricultural months across different regions. This resulted in the removal of maximum temperature (Tmax) and minimum temperature (Tmin) as well as rainfall data for the months of February, March, April, and November. The resulting dataset comprised 2253 rows and 18 features, consisting of maximum and minimum temperature, and rainfall data from May to October. The minimum, maximum, mean, and standard deviation values were calculated for all cases from Table I to Table III. Additionally, these parameters were calculated for the high case as shown in Table IV to Table VI, and for the low case as shown in Table VII to Table IX.

TABLE I
Statistical Analysis of the Tmax (May-oct) Features for the Both Classes



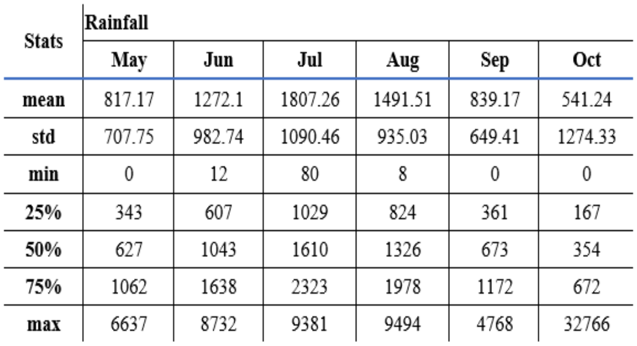
Stats	Maximum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1240.56	1273.73	1285.8	1279.93	1242.95	1185.17
std	40.32	37.31	37.06	35.89	38.77	53.27
min	1127	1149	1157	1143	1127	1037
25%	1212	1252	1265	1260	1216	1146
50%	1243	1276	1287	1281	1245	1191
75%	1269	1301	1313	1304	1269	1219
max	1372	1391	1403	1399	1385	1344

TABLE II
Statistical Analysis of the Tmin (May-oct) Features for the Both Classes



Stats	Minimum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1125.37	1171.72	1197.29	1188.8	1140.78	1079.93
std	50.32	42.47	42.38	43.39	50.84	65.04
min	996	1044	1046	1043	1010	924
25%	1086	1146	1176	1165	1097	1023
50%	1129	1175	1200	1193	1149	1086
75%	1161	1202	1230	1221	1179	1131
max	1246	1259	1278	1271	1254	1228

TABLE III
Statistical Analysis Of The Rainfall (May-oct) Features For The Both Classes



Stats	Rainfall					
	May	Jun	Jul	Aug	Sep	Oct
mean	817.17	1272.1	1807.26	1491.51	839.17	541.24
std	707.75	982.74	1090.46	935.03	649.41	1274.33
min	0	12	80	8	0	0
25%	343	607	1029	824	361	167
50%	627	1043	1610	1326	673	354
75%	1062	1638	2323	1978	1172	672
max	6637	8732	9381	9494	4768	32766

TABLE IV
STATISTICAL ANALYSIS OF THE TMAX (MAY-OCT) FEATURES FOR THE HIGH CLASS

Stats	Maximum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1237.79	1276.37	1289.02	1282.09	1242.89	1177.89
std	36.95	34.10	31.31	29.46	33.51	49.43
min	1130.0	1156.0	1171.0	1158.0	1141.0	1051.0
25%	1212.0	1256.0	1270.0	1264.0	1220.0	1140.0
50%	1240.0	1276.0	1289.0	1282.0	1246.0	1184.0
75%	1265.0	1302.0	1311.0	1302.0	1267.0	1214.0
max	1355.0	1391.0	1377.0	1383.0	1385.0	1334.0

TABLE V
STATISTICAL ANALYSIS OF THE TMIN (MAY-OCT) FEATURES FOR THE HIGH CLASS

Stats	Minimum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1122.23	1172.36	1201.13	1191.40	1137.15	1070.76
std	43.17	34.96	33.73	35.46	46.49	60.61
min	1013.0	1057.0	1062.0	1060.0	1010.0	924.0
25%	1091.0	1149.0	1181.0	1169.0	1098.5	1017.0
50%	1124.0	1174.0	1201.0	1194.0	1145.0	1074.0
75%	1155.0	1198.0	1226.0	1218.0	1174.0	1122.0
max	1235.0	1254.0	1271.0	1271.0	1238.0	1211.0

TABLE VI
STATISTICAL ANALYSIS OF THE RAINFALL (MAY-OCT) FEATURES FOR THE HIGH CLASS

Stats	Rainfall					
	May	Jun	Jul	Aug	Sep	Oct
mean	717.95	1105.60	1721.78	1379.02	714.237	471.13
std	639.86	871.0	1016.1	831.4	611.44	1432.71
min	1.0	12.0	150.0	39.0	1.0	0.0
25%	311.0	519.5	957.5	743.0	285.5	146.0
50%	565.0	899.0	1522.0	1216.0	544.0	295.0
75%	916.5	1418.0	2221.0	1839.0	944.0	569.0
max	6637.0	7667.0	7307.0	5427.0	4092.0	32766.0

TABLE VII
STATISTICAL ANALYSIS OF THE TMAX (MAY-OCT) FEATURES FOR THE LOW CLASS

Stats	Maximum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1243.21	1271.19	1282.70	1277.87	1243.0	1192.16
std	43.15	40.0	41.62	41.02	43.24	55.84
min	1127.0	1149.0	1157.0	1143.0	1127.0	1037.0
25%	1213.0	1247.0	1259.0	1254.0	1211.0	1153.0
50%	1245.0	1275.0	1284.0	1281.0	1244.0	1197.0
75%	1274.0	1299.0	1315.0	1307.0	1272.0	1226.0
max	1372.0	1386.0	1403.0	1399.0	1364.0	1344.0

TABLE VIII
STATISTICAL ANALYSIS OF THE TMIN (MAY-OCT) FEATURES FOR THE LOW CLASS

Stats	Minimum Temperature					
	May	Jun	Jul	Aug	Sep	Oct
mean	1128.39	1171.1	1193.61	1186.31	1144.26	1088.72
std	56.18	48.61	49.0	49.71	54.48	67.90
min	996.0	1044.0	1046.0	1043.0	1025.0	925.0
25%	1078.0	1141.0	1170.0	1159.0	1095.0	1030.2
50%	1134.0	1177.0	1199.0	1192.0	1153.0	1100.0
75%	1168.0	1206.0	1234.0	1226.7	1185.0	1139.0
max	1246.0	1259.0	1278.0	1270.0	1254.0	1228.0

TABLE IX
STATISTICAL ANALYSIS OF THE RAINFALL (MAY-OCT) FEATURES FOR THE LOW CLASS

Stats	Rainfall					
	May	Jun	Jul	Aug	Sep	Oct
mean	912.35	1431.8	1889.25	1599.41	959	608.49
std	755.36	1055.04	1151.78	1013.31	662.31	1097.66
min	0	27	80	8	0	0
25%	374	703.5	1096.25	908.25	462	202.25
50%	711.5	1217	1664.5	1400.5	815	426
75%	1220	1867.2	2402.25	2083.75	1320.7	791.5
max	5748	8732	9381	9494	4768	32766

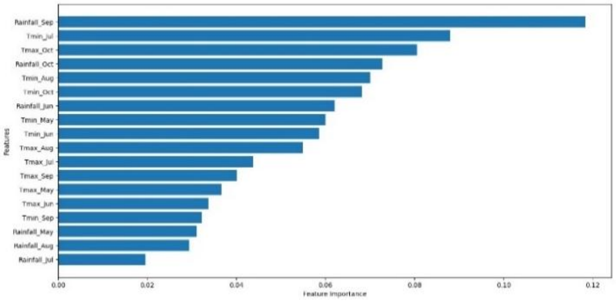


Fig. 2. Feature importance used Random Forest model.

The most important features are presented in the Fig.2 which used the RF model, it showed that ten features are most corresponding to the target maize yield level. Those ten features are Rainfall_Sep, Tmin_Jul,Tmax_Oct, Rainfall_Oct, Tmin_Aug, Tmin_Oct, Rainfall_Jun, Tmin_May, Tmin_Jun, Tmax_Aug. Fig.2 presented the x-axis as a feature importance and the y-axis is the features.

C. *Machine learning for classification*

1) K Nearest Neighbors (KNN)

It is a supervised machine learning method that uses a lazy learner approach to classify new instances based on their similarity to existing data/cases. During the training phase, KNN only stores the data in memory and it performs classification of the yield level of Maize on the dataset at the time of prediction. KNN is used to improve discriminant analysis when a precise parametric estimate of probability densities is uncertain or difficult to evaluate, as well as to classify an unknown sample based on its proximity to previously recognized samples, depending on the distance between them. The prediction of KNN is based on the closest k number of training points to the target location, known as the neighborhood point [35].

2) Naive Bayes (NB)

It is a machine learning algorithm that calculates in a more efficient, accurate, and easier to implement manner, and is usually used for classification problems. It is based on Bayes theory and assumes that the final output has uncorrelated features, even if the dataset features are correlated or related to each other. The main contribution of using Naïve Bayes is that all the features can independently classify correctly. The Naïve Bayes's assumption of conditional independence helps to measure the sample data's class conditional probabilities, and the training data can be directly estimated from the training data instead of evaluating it [36]. We used two models in Naive Bayes for categorization:

a) Multinomial NB (MNB): It is used for discrete counts.

b) Bernoulli NB (BNB): The binomial model is useful if your feature vectors are binary.

3) Decision Tree Classifier (DTC)

DTC [36] is supervised machine learning algorithms used to solve classification problems. A decision tree is represented as a tree with inner nodes, branches, and leaf nodes. Each leaf node represents the final decision, while the inner node represents the features of the dataset. The main advantage of using a decision tree is that it provides all possible solutions for solving a problem, similar to human thinking in decision making.

4) Quadratic Discriminant Analysis (QDA):

QDA [37] is a supervised machine learning model that requires the dataset to be normally distributed and works well when there is not a significant difference between the group covariance matrices. As one of the most popular classifier models, it has the ability to create a quadratic function of categorized data that delivers the greatest mean differences between the various data levels.

D. First Proposed Model (FPM)

The FPM ensemble of five machine learning algorithms (KMBDQ) is presented in Fig.3. It removes the old previous prediction and uses the new prediction of each machine learning as an input feature so that the final machine learning will have only one output prediction as an input feature with the dataset. Each machine learning model divided the dataset into two parts, using 90% of the dataset for the training stage to obtain the parameters of the model, followed by the testing stage to evaluate the model's performance using the remaining 10% of the dataset. The processed data was organized in ascending order by year. The dataset is then separated into training with 90% (2027 samples) and 10% (226 samples) correspondingly. Even though most studies employ a 70:30 or 80:20 split of training and testing samples, this study only used 10% processed data for testing, because the main target is predicting the maize yield level for the current year, which is good enough 10% for testing and it uses a massive training dataset.

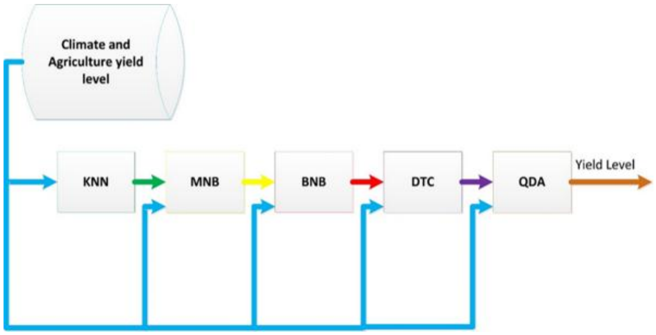


Fig. 3. The block diagram of the FPM.

The dataset is divided into two parts, with 90% (2027 samples) used for training and 10% (226 samples) for testing. Although many studies use a 70:30 or 80:20 split for training and testing, this study only used 10% of the processed data for testing because the main objective is to predict the maize yield level for the current year. Using a large training dataset can improve testing accuracy. The steps of the FPM process are presented in Fig.4.

E. Second Proposed Model (SPM)

The novel architecture shown in Fig.5 ensembles five classification machine learning algorithms (KMBDQ). It takes the forecast output labels (discrete values of 0 or 1) of each machine learning as an input feature with the dataset to learn at each stage with different machine learning in a series form. The final machine learning will have all the prediction output labels of other machine learning as an input feature with the dataset, according to the Maize China agricultural dataset. At each stage, the number of features in the dataset increases in ascending order. Each classifier divides the dataset into 90% for training and 10% for testing.

The examination of several orders for the five-machine learning has been performed and evaluated, and the order of these five machine learning algorithms has been chosen based on the accurate final output. To achieve a more accurate prediction, the SPM process involves five stages: (i) KNN, (ii) MNB, (iii) BNB, (iv) DTC, and (v) QDA. The steps of the SPM process are presented in Fig.6. The goal of the proposed work is to increase accuracy, and the time consumed in prediction is low, so the computation time of the prediction is not a critical point in our model.

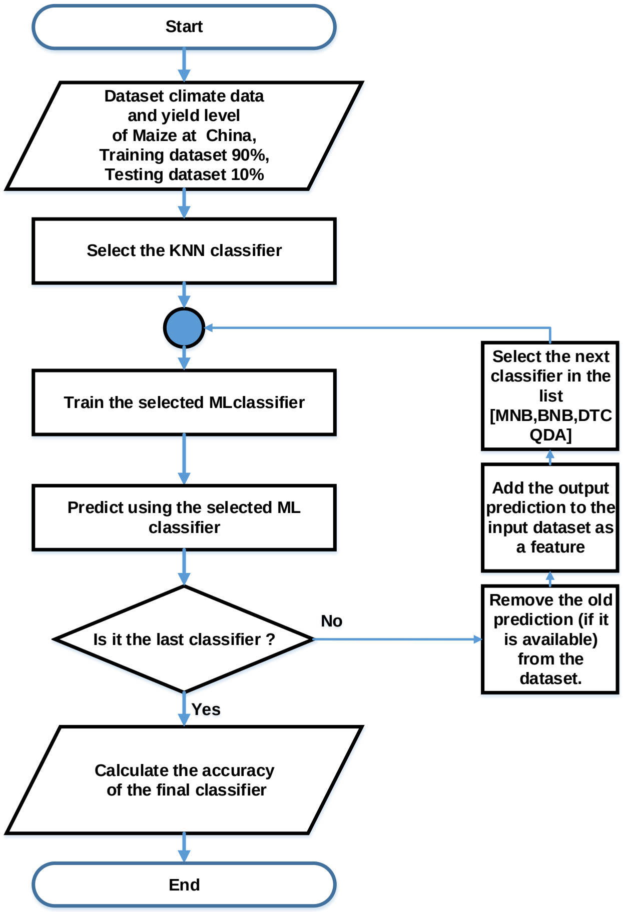

Start
Dataset climate data and yield level of Maize at China, Training dataset 90%, Testing dataset 10%
Select the KNN classifier
Train the selected MLclassifier
Predict using the selected ML classifier
Is it the last classifier ?
No
Yes
Remove the old prediction (if it is available) from the dataset.
Add the output prediction to the input dataset as a feature
Select the next classifier in the list [MNB,BNB,DTC QDA]
Calculate the accuracy of the final classifier
End

Fig. 4. The FPM flowchart.

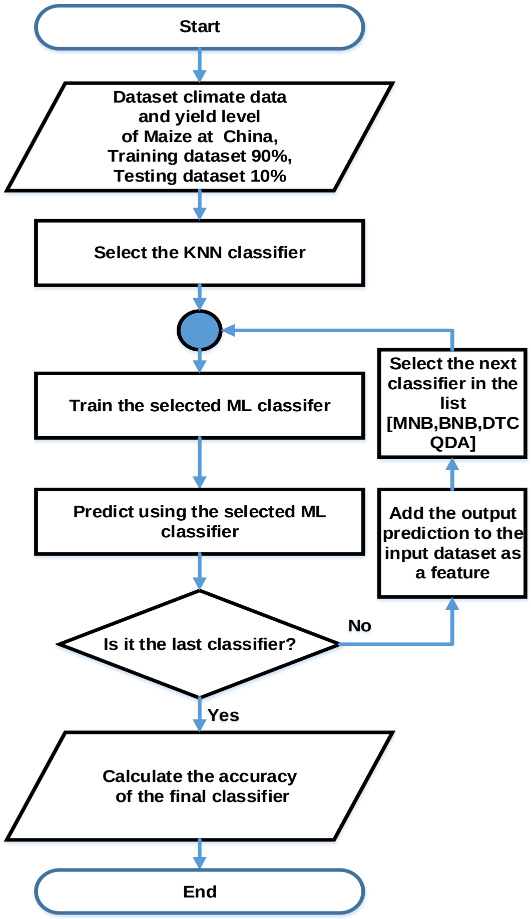

Start
Dataset climate data and yield level of Maize at China, Training dataset 90%, Testing dataset 10%
Select the KNN classifier
Train the selected ML classifer
Predict using the selected ML classifier
Is it the last classifier?
No
Yes
Add the output prediction to the input dataset as a feature
Select the next classifier in the list [MNB,BNB,DTC QDA]
Calculate the accuracy of the final classifier
End

Fig. 6. The SPM flowchart.

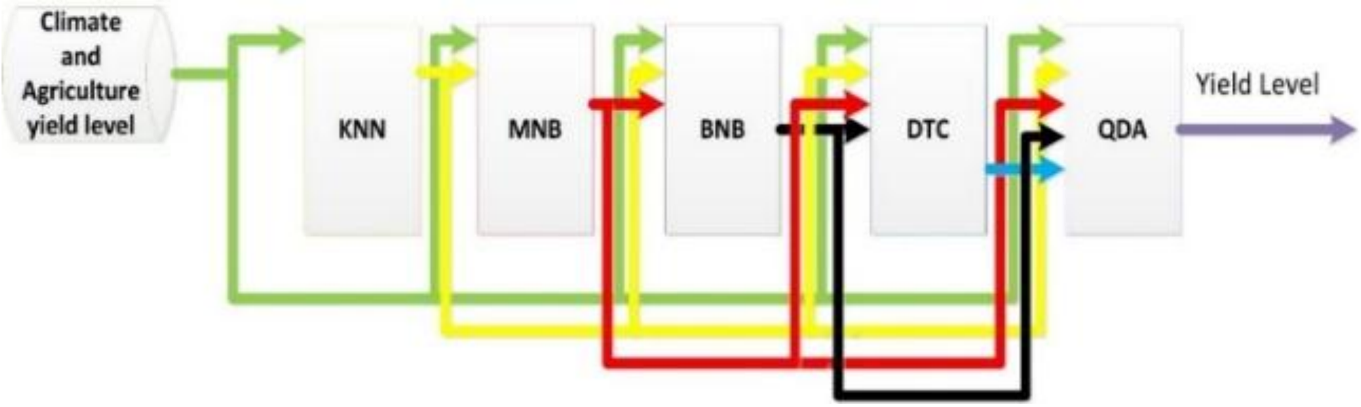


Fig. 5. The block diagram of the SPM.

IV. PERFORMANCE EVALUATION

For evaluating the performance of the two proposed models, some metrics such as accuracy, sensitivity, precision and F1 score [38] were measured as shown in Equations 1-4. The machine specification is CPU Intel® Core™ i3-9100 processor and RAM is 8 GB used in this analysis.

A confusion matrix is one of the measures used to assess the model's performance, along with true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These parameters are used to calculate the accuracy in equation 1, which is one of the most commonly used metrics for assessing model performance. The accuracy is defined as the ratio of correctly classified patterns to the total number of all classified patterns.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Sensitivity (Recall) is shown in equation 2 as the ratio of correctly predicted positive events to all positive events in the actual test dataset.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

Precision is given as in equation 3, which is the ratio of the correctly predicted positive events to the total predicted positive events.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

F1 score [39] is given as in equation 4 is the weighted mean of sensitivity and precision.

$$\text{F1 score} = 2 \times \frac{(\text{Sensitivity} \times \text{Precision})}{(\text{Sensitivity}+\text{Precision})} \quad (4)$$

V. EXPERIMENTS RESULTS

The main important points that should be covered:

- Removing the uncommon features to merge four different regions in China that affect maize yield the most.
- Predicting maize yield level one month earlier before harvesting.
- Compare the standard metrics results of each machine learning, FPM and SPM.

A. Models and machine learning classifiers analysis

The Scikit-learn library [40] and the open-source programming language Python [41] were used to generate predictive models. However, the FPM approach did not achieve high accuracy. Therefore, the SPM approach was applied by trying out all possible orders for each algorithm in the five machine learning algorithms, and the best order was found to be KNN-MNB-BNB-DTC-QDA, which achieved the highest accuracy. The accuracy of the two proposed models and 17 popular machine learning classifiers are shown in Fig.7.

A comparison of the 17 popular machine learning classifiers, as well as the FPM and SPM models, was performed. The Stochastic Gradient Descent Classifier (SGDC) achieved the lowest accuracy of 40.2% among the machine learning classifiers.

The best performing algorithm was the Multinomial Naive Bayes (MNB), achieving an accuracy of 69.4% in the 17 popular machine learning classifiers.

In the FPM, the same algorithms were used with a different approach and achieved an accuracy of 68.5%, while in the SPM approach, it obtained an accuracy of 79.6%, outperforming the other 17 classifiers, including Stochastic Gradient Descent Classifier (SGDC) with the worst accuracy of 40.2%. Other classifiers, such as Adaptive Boosting Classifier (AdaBoost) (64.1%), Light GBM Classifier (LGBM) (65.9%), Extra Trees Classifier (ETC) (66.8%), Logistic Regression (LR) (56.1%), Gradient Boosting Classifier (GBC) (62.8%), Decision Tree Classifier (DTC) (60.1%), Random Forest Classifier (RFC) (56.6%), Support Vector Classification (SVC) (61.9%), Gaussian Naive Bayes (GNB) (67.6%), Bernoulli Naive Bayes (BNB) (57.9%), Linear Discriminant Analysis (LDA) (55.7%), Quadratic Discriminant Analysis (QDA) (67.6%), K Nearest Neighbors (KNN) (50%), Multi-layer perceptron Classifier (MLPC) (63.2%), and Extreme Gradient Boosting (XGB) (64.1%) were also evaluated.

The execution time for predicting a single yield level for the two proposed models and the most popular classifier was calculated and presented in Fig.8.

According to Fig.7 and Fig.8, the SPM model demonstrated high predictive accuracy and fast run time compared to other classifiers such as AdaBoost, ETC, and RFC.

The performance of the two proposed model approaches will be demonstrated by combining two, three, four, and five machine learning algorithms, which will be shown in Fig.9 and Fig.10.

The first proposed technique achieved accuracy of 69.4%, 69.4%, 63.7%, and 68.5%, respectively, when two, three, four, and five combined machine learning algorithms were used. However, the FPM combination approach was not effective and resulted in reduced accuracy in the final machine learning model. This technique was the worst combination of machine learning algorithms applied to the Maize dataset of China, as shown in Fig.9.

The accuracy of each combined machine learning algorithm was measured using the SPM approach, and each combination achieved a different accuracy, as shown in Fig.10. In the second proposed technique, the accuracy of the two, three, four, and five combined machine learning algorithms achieved 69.4%, 50%, 50.8%, and 79.6%, respectively.

The classification performance of the two proposed models that were implemented in the current research was evaluated by combining five machine learning algorithms using different approaches. The same dataset was applied to various benchmark machine learning classification algorithms such as AdaBoost, LGBM, ETC, LR, GBC, DTC, RFC, SVC, GNB, BNB, MNB, LDA, QDA, KNN, MLPC, SGDC, and XGB.

To evaluate the performance of each machine learning algorithm and the two proposed models, several metrics were used, including confusion matrix, precision, sensitivity, F1 score, and execution time for predicting the yield level. The classification accuracy was calculated as a percentage and is presented in Table X to Table XIII. The evaluation metrics for the popular machine learning algorithms are shown in Table X and Table XI.

The parameters of the confusion matrix are arranged according to the high maize yield level, and the sensitivity, precision, and F1 score are calculated for the high-yield level. These metrics will be presented from Table X to Table XIII.

The observations taken from the results shown in Table X and Table XI indicate that the 17 classifiers were measured based on their F1 scores, achieving a range from 0.41 to 0.80. The sensitivity was achieved from a range of 0.27 to 0.79. The best classifier, in terms of both F1 score and sensitivity, was MNB, while the worst one was SGDC. Precision was calculated for LGBM, which achieved an impressive result of 0.89, while MNB and KNN were also very good at achieving a precision of 0.81.

It can be observed from the results presented in Table XII and Table XIII that the evaluation metrics, namely F1 score, sensitivity and precision, were calculated for the two, three, four and five combined machine learning algorithms in the FPM and SPM. Each combination was evaluated based on the F1 score, which achieved a range of 0.58 to 0.80 in the FPM and a range of 0.58 to 0.88 in the SPM.

Regarding the FPM, the best combination was KNN, MNB and KNN, MNB, BNB, which achieved the highest F1 score, but KNN was not efficient. On the other hand, in the SPM, the combination of KNN, MNB, BNB, DTC, QDA achieved the highest F1 score, whereas the KNN, MNB, BNB combinations and KNN alone had the lowest F1 score.

The sensitivity and precision for each combination were calculated using the parameters of the confusion matrix separately for the FPM and SPM approaches. In the FPM, the sensitivity ranged from 0.45 to 0.79, while in the SPM it ranged from 0.45 to 0.99. Precision was also calculated for each combination, with the FPM ranging from 0.81 to 0.86 and the SPM ranging from 0.81 to 0.83. The best values for the F1 score, sensitivity, and precision are close to one, while the worst values are close to zero.

It can be seen from Tables X to XIII that AdaBoost, ETC, and RFC take longer execution time compared to the two proposed models and the other popular classifiers. Furthermore, XGB is faster than the other popular machine learning algorithms, while AdaBoost is the slowest. MNB and SGDC finish the task at the same time. The SPM achieved the highest prediction accuracy, sensitivity, and F1 score, as shown in Tables XII and XIII.

VI. DISCUSSION

The current research introduces two novel proposed models for improving the prediction of maize yield levels using the Scikit-learn library in Python programming.

The aim is to predict higher maize yield levels in suitable climates for maize and lower maize yield levels in unsuitable environments [24]. This prediction is important to inform farmers about the best time to harvest in order to achieve higher yields. The main objective of using the SPM is to assist farmers in making informed decisions regarding crop management policies and practices in order to achieve sustained maize productivity.

The majority of researchers suggest that building better ensemble machine learning algorithms for the available datasets is the most effective way to improve accuracy [20-21]. However, the FPM in our research achieved less accuracy than the best classifier machine learning. To overcome this limitation, the SPM was implemented, which combined five machine learning algorithms (KNN, MNB, BNB, DTC, and QDA) with different approaches.

The comparison with other works on the same dataset is difficult due to the use of multiple linear regression algorithm in the other work, which did not yield good results. The R-squared values were closer to zero, indicating incorrect predictions [32]. For instance, the R-squared values for multiple linear regression were 0.0089, 0.0223, 0.0209, and 0.0207 for the four regions Southwest China, Huanhuaihai, North China, and Northeast China, respectively [32]. Therefore, it was necessary to change the target of the output to the ordinal (High-low) maize yield level and combine the four regions into a single dataset. Using classifier machine learning algorithms resulted in better prediction results.

Among 30 weather features, the filter approach is used to get the common months to feature in maize agriculture. The common features selected were 18 weather features (max, min temperature and rainfall) from May to October, reducing the number of features achieved at high speed in the runtime.

The maize China dataset is divided into 90 % training and 10% testing phases. In the training phase, the machine learning works on the updated datasets at every stage to obtain the parameters for each classifier to be tested. In the testing phase, the performance of each combination of machine learning has been evaluated.

The experimental results have shown that the SPM outperformed the first classifier KNN by 29.6% in terms of improvement. The SPM achieved high sensitivity, great accuracy, and an impressive F1 score. Furthermore, the feature filtering method successfully reduced the number of features, which helped to speed up the classification process. The filtering approach was achieved by removing the uncommon features in the four different regions of the China datasets and combining them into a single dataset, thus reducing the complexity of feature selection.

The study aimed to predict recent maize yield by arranging the years in ascending order and dividing them into 90% for training and 10% for testing. The purpose of this split was to simulate the prediction of future yield based on recent data. In this context, K-fold cross validation was not deemed necessary as the objective was to evaluate the model's ability to make predictions on recent data, rather than to estimate its generalization performance. But other evaluation approach FPM was used by the train-test split to compare it with SPM.

A single algorithm may not give a perfect prediction for a given data set. Machine learning algorithms have their limitations and creating a model with high accuracy is challenging. multiple models have been combined together to increase the overall accuracy. The study focused on four regions in China and the results might not be generalizable to other regions or countries with different climate patterns.

The study was limited in its scope by only considering climatic factors for predicting maize yield level and not taking into account other crucial factors such as types of insects, water availability and agricultural practices that have the potential to greatly influence the yield. The drawback of using the FPM is that it decreased accuracy by 0.9% compared to the two combined machine learning algorithms (KNN-MNB) in the model. However, the advantage of using the FPM is that it achieved an impressive result in precision compared to the SPM. The best MNB classifier achieved acceptable accuracy, great F1 score, and good sensitivity compared to other machine learning algorithms.

Comparing the experimental results to previous studies predicting maize yield levels in China was challenging because those studies targeted different yield levels of prediction and used different factors that impact yield levels. Some studies depended on soil parameters combined with climate data to predict yield levels [42-43]. However, those studies neglected modern agricultural approaches such as the use of artificial soil or hydroponics (soilless plant culture). Additionally, each study used a different dataset. The use of soil properties and climatic data in predicting crop yield has been a common approach in many studies, as seen in [43].

In our study, we aimed to cover a larger area of agricultural maize land by combining data from four different regions in China, totaling 15.5 million hectares. Rather than relying on soil data, our focus was on utilizing modern agricultural techniques such as artificial soil or hydroponics.

Maize crops are highly sensitive to changes in climate, with various climatic parameters such as maximum and minimum temperatures, and rainfall affecting their growth. The average temperature has increased by 1.2°C since 1961, indicating the occurrence of climate changes [44]. In the dry North, rainfall has decreased while it has increased in the wet South [45]. In relation to maize yield, an inverse proportional relationship exists between Tmax and yield, while a direct proportional relationship exists between Tmin and yield. An increase in Tmax would result in a decrease in maize yield, whereas an increase in Tmin would lead to a rise in yield, particularly in Northeast China [46]. Furthermore, changes in rainfall (ΔRf) have a negative effect on maize yield in the Northeast and Huanghuaihai regions [32].

The proposed future work is to develop and implement a machine learning and deep learning hybrid model for predicting crop yield. This will involve performing various experiments on the dataset and evaluating the performance of different deep learning algorithms [47] with an increased dataset. Ensemble techniques will also be explored to further improve the accuracy of the model.

In the current study, climate data was used as the feature and maize yield level as the target variable. The SPM achieved an accuracy improvement of 10.2% over the best performing machine learning classifier.

VII. CONCLUSIONS

The study proposed two novel models to predict accurate maize yield levels in China. The models consisted of five machine learning classifiers (KMBDQ) that used different methods of combination resulting in two architectures. Standard performance evaluation metrics were used to evaluate the models.

The SPM achieved higher accuracy compared to the FPM and the other 17 popular machine learning classifiers. The SPM also achieved the best sensitivity and F1 score. It was evident that the SPM was efficient and yielded impressive improvements. On the other hand, the FPM combination method was not as efficient as the SPM, as it gave a lower accuracy compared to the best individual machine learning and SPM results. The computation time for a single predicted maize yield level was calculated within microseconds.

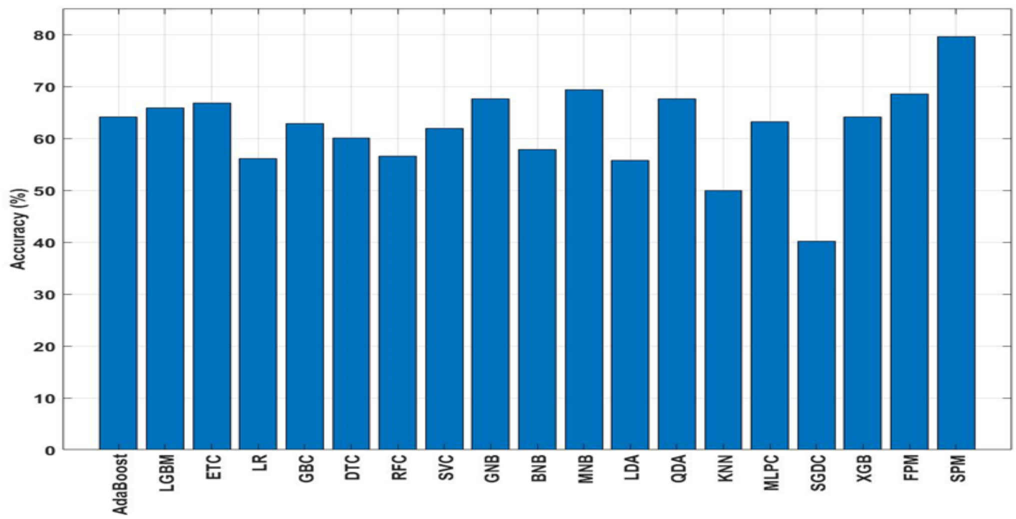


Fig. 7. The accuracy of the two proposed models and the most popular classifier machine learning algorithms.

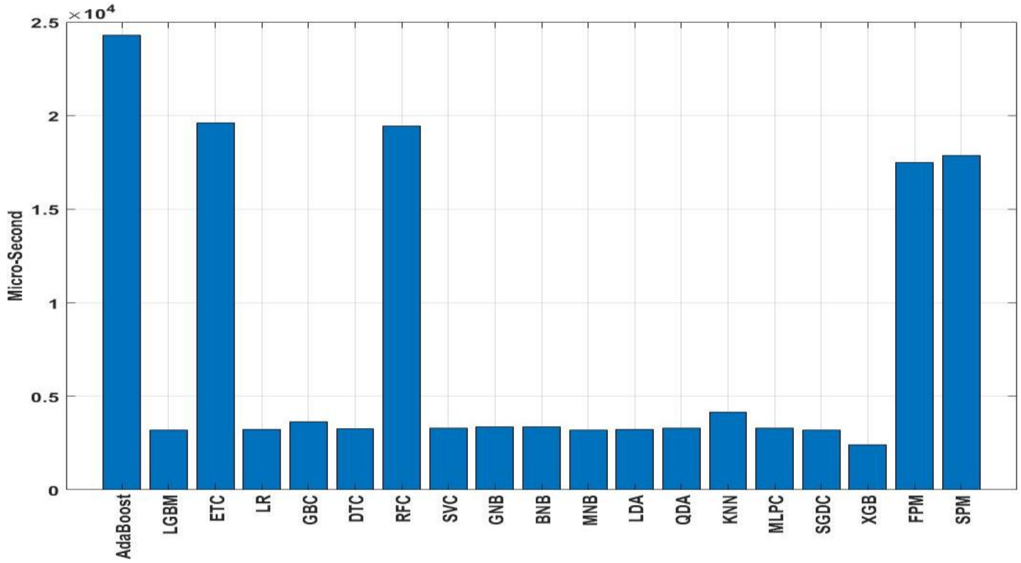


Fig. 8. Execution time for prediction of the two proposed models and the most popular classifier machine learning.

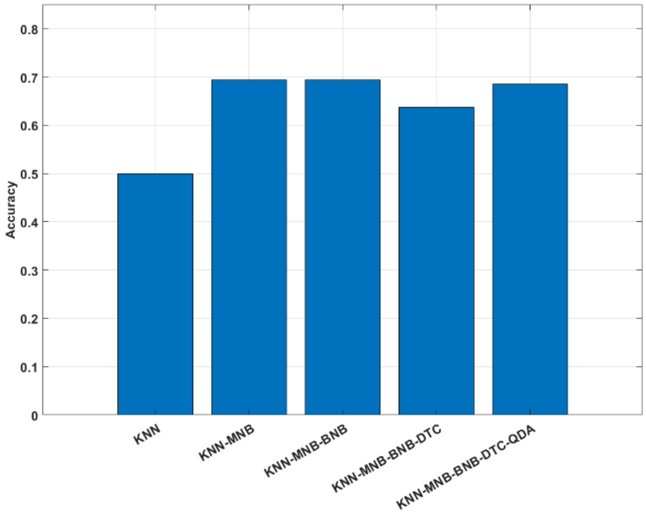


Fig. 9. The accuracy of the two, three, four and five combined machine learning algorithms that used the first proposed algorithms.

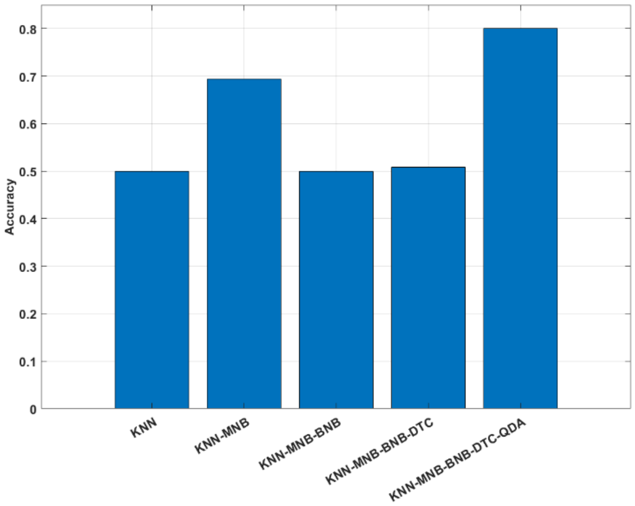
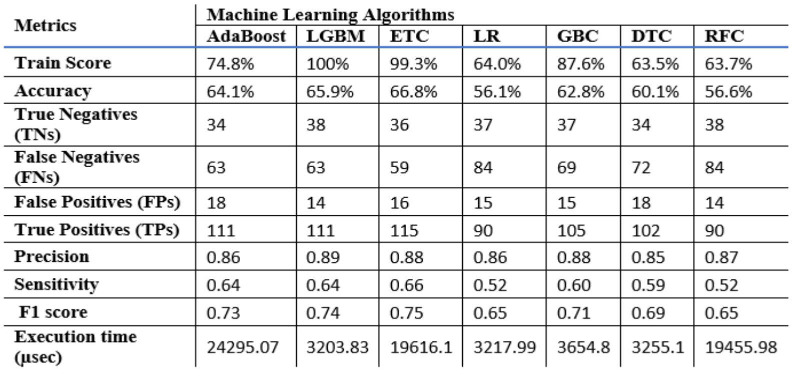


Fig.10. The accuracy of the two, three, four and five combined machine learning algorithms that used the second proposed algorithms.

TABLE X

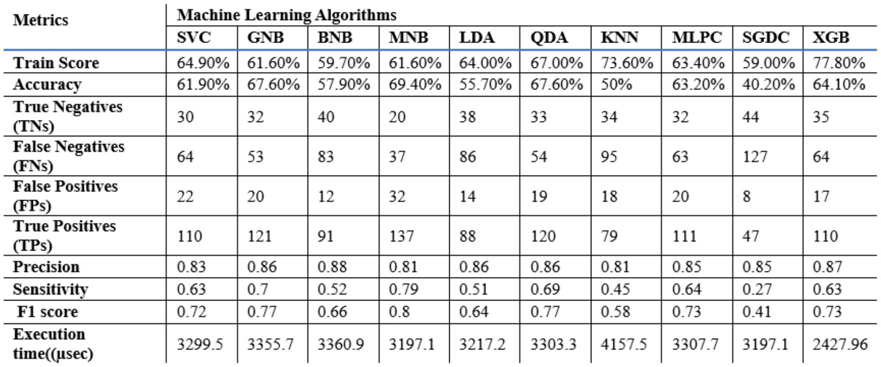
Performance Evaluation Of The Most Popular Classifiers Adaboost, Lgbm, Etc, Lr, Gbc, Dtc And Rfc



Metrics	Machine Learning Algorithms						
	AdaBoost	LGBM	ETC	LR	GBC	DTC	RFC
Train Score	74.8%	100%	99.3%	64.0%	87.6%	63.5%	63.7%
Accuracy	64.1%	65.9%	66.8%	56.1%	62.8%	60.1%	56.6%
True Negatives (TNs)	34	38	36	37	37	34	38
False Negatives (FNs)	63	63	59	84	69	72	84
False Positives (FPs)	18	14	16	15	15	18	14
True Positives (TPs)	111	111	115	90	105	102	90
Precision	0.86	0.89	0.88	0.86	0.88	0.85	0.87
Sensitivity	0.64	0.64	0.66	0.52	0.60	0.59	0.52
F1 score	0.73	0.74	0.75	0.65	0.71	0.69	0.65
Execution time (µsec)	24295.07	3203.83	19616.1	3217.99	3654.8	3255.1	19455.98

TABLE XI

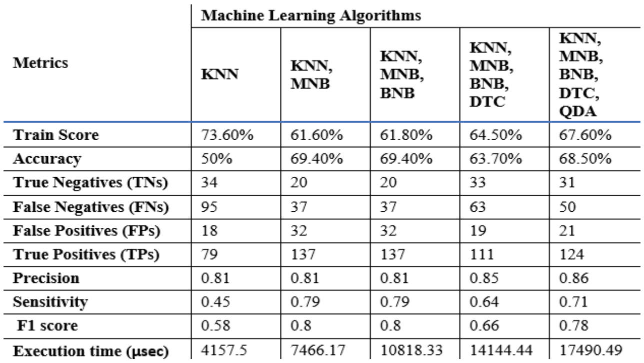
Performance Evaluation Of The Most Popular Classifiers, Svc, Gnb, Bnb, Mnb, Lda, Qda, Knn, Mlpc, Sgdc And Xgb



Metrics	Machine Learning Algorithms									
	SVC	GNB	BNB	MNB	LDA	QDA	KNN	MLPC	SGDC	XGB
Train Score	64.90%	61.60%	59.70%	61.60%	64.00%	67.00%	73.60%	63.40%	59.00%	77.80%
Accuracy	61.90%	67.60%	57.90%	69.40%	55.70%	67.60%	50%	63.20%	40.20%	64.10%
True Negatives (TNs)	30	32	40	20	38	33	34	32	44	35
False Negatives (FNs)	64	53	83	37	86	54	95	63	127	64
False Positives (FPs)	22	20	12	32	14	19	18	20	8	17
True Positives (TPs)	110	121	91	137	88	120	79	111	47	110
Precision	0.83	0.86	0.88	0.81	0.86	0.86	0.81	0.85	0.85	0.87
Sensitivity	0.63	0.7	0.52	0.79	0.51	0.69	0.45	0.64	0.27	0.63
F1 score	0.72	0.77	0.66	0.8	0.64	0.77	0.58	0.73	0.41	0.73
Execution time((µsec)	3299.5	3355.7	3360.9	3197.1	3217.2	3303.3	4157.5	3307.7	3197.1	2427.96

TABLE XII

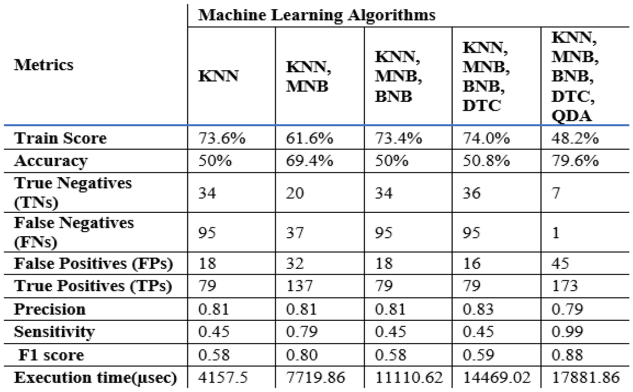
Performance Evaluation Of The Fpm Algorithm Applied On The Two, Three, Four And Five Combinations



Metrics	Machine Learning Algorithms				
	KNN	KNN, MNB	KNN, MNB, BNB	KNN, MNB, BNB, DTC	KNN, MNB, BNB, DTC, QDA
Train Score	73.60%	61.60%	61.80%	64.50%	67.60%
Accuracy	50%	69.40%	69.40%	63.70%	68.50%
True Negatives (TNs)	34	20	20	33	31
False Negatives (FNs)	95	37	37	63	50
False Positives (FPs)	18	32	32	19	21
True Positives (TPs)	79	137	137	111	124
Precision	0.81	0.81	0.81	0.85	0.86
Sensitivity	0.45	0.79	0.79	0.64	0.71
F1 score	0.58	0.8	0.8	0.66	0.78
Execution time (µsec)	4157.5	7466.17	10818.33	14144.44	17490.49

TABLE XIII

Performance Evaluation Of The Spm Algorithm Applied On Two, Three, Four And Five Combinations



Metrics	Machine Learning Algorithms				
	KNN	KNN, MNB	KNN, MNB, BNB	KNN, MNB, BNB, DTC	KNN, MNB, BNB, DTC, QDA
Train Score	73.6%	61.6%	73.4%	74.0%	48.2%
Accuracy	50%	69.4%	50%	50.8%	79.6%
True Negatives (TNs)	34	20	34	36	7
False Negatives (FNs)	95	37	95	95	1
False Positives (FPs)	18	32	18	16	45
True Positives (TPs)	79	137	79	79	173
Precision	0.81	0.81	0.81	0.83	0.79
Sensitivity	0.45	0.79	0.45	0.45	0.99
F1 score	0.58	0.80	0.58	0.59	0.88
Execution time(µsec)	4157.5	7719.86	11110.62	14469.02	17881.86

REFERENCES

[1] Y. Xie, Y. Zhang, H. Lan, L. Mao, S. Zeng and Y. Chen, "Investigating long-term trends of climate change and their spatial variations caused by regional and local environments through data mining," *Journal of Geographical Sciences*, vol. 28, no. 6, pp. 802–818, 2018.

[2] X. Han, F. Liu, X. He and F. Ling, "Research on rice yield prediction model based on Deep Learning," Computational Intelligence and Neuroscience, vol. 2022, pp. 1-9, 2022. DOI: 10.1155/2022/1922561.

[3] O. P. Dhankher and C. H. Foyer, "Climate resilient crops for improving global food security and safety," Plant, Cell and Environment, vol. 41, no. 5, pp. 877–884, 2018.

[4] C. A. Silva Junior, A. H. Leonel-Junior, F. S. Rossi, W. L. Correia Filho, D. de Santiago, J. F. Oliveira-Júnior, P. E. Teodoro, M. Lima and G. F. Capristo-Silva, "Mapping soybean planting area in Midwest Brazil with remotely sensed images and phenology-based algorithm using the Google Earth Engine Platform," Computers and Electronics in Agriculture, vol. 169, 2020. DOI:10.1016/j.compag.2019.105194.

[5] S. Veenadhari, B. Misra and C. D. Singh, "Machine Learning Approach for forecasting crop yield based on climatic parameters," Jan. 3, 2014 International Conference on Computer Communication and Informatics, 2014. DOI:10.1109/iccci.2014.6921718.

[6] M. K. Behera, "Crop yield prediction using machine learning techniques," Protected Cultivation and Smart Agriculture, 2020. DOI:10.30954/ndp-pcsa.2020.35.

[7] L. S. Cedric, W. Y. Adoni, R. Aworka, J. T. Zoueu, F. K. Mutombo, M. Krichen and C. L. Kimpolo, "Crops yield prediction based on machine learning models: Case of West African countries," Smart Agricultural Technology, vol. 2, 2022. DOI: 10.1016/j.atech.2022.100049

[8] A. Chlingaryan, S. Sukkarieh and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in Precision Agriculture: A Review," Computers and Electronics in Agriculture, vol. 151, pp. 61–69, 2018.

[9] H. Shah, P. Hellegers and C. Siderius, "Climate risk to agriculture: A synthesis to define different types of critical moments," Climate Risk Management, vol. 34, 2021. DOI: 10.1016/j.crm.2021.100378.

[10] J. Zhang, H. Tian, J. Yang and S. Pan, "Improving representation of crop growth and yield in the dynamic land ecosystem model and its application to China," Journal of Advances in Modeling Earth Systems, vol. 10, no. 7, pp. 1680–1707, 2018.

[11] J. Bongaarts, "FAO, IFAD, UNICEF, WFP and whothe state of food security and nutrition in the World 2020. Transforming Food Systems for affordable healthy DIETSFAO, 2020, 320 p.," Population and Development Review, vol. 47, no. 2, pp. 558–558, 2021.

[12] S. Karlos, G. Kostopoulos and S. Kotsiantis, "A soft-voting ensemble based co-training scheme using static selection for binary classification problems," Algorithms, vol. 13, no. 1, 2020. DOI: 10.3390/a13010026

[13] N. Wang and Z. Li, "A stacking-based short-term wind power forecasting method by CBLSTM and ensemble learning," Journal of Renewable and Sustainable Energy, vol. 14, no. 4, 2022. DOI: 10.1063/5.0097757

[14] S. S. Kale, P. Patil and G. Mamatha, "Application of stacking ensemble technique in agriculture for prediction of crop yield," International journal of health sciences, pp. 14750–14755, 2022.

[15] S. Thomas and J. Thomas, "Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning," Artificial Intelligence in Agriculture, vol. 6, pp. 100–110, 2022.

[16] R. Yamparla, H. S. Shaik, N. S. Guntaka, P. Marri and S. Nallamothu, "Crop yield prediction using random forest algorithm," 7th International Conference on Communication and Electronics Systems (ICCES), 2022. DOI: 10.1109/icces54183.2022.9835756

[17] Y.-H. Kuo, Z. Li and D. Kifer, "Detecting outliers in data with correlated measures," Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 287-296, Oct 17 2018.

[18] P. Saraswat and S. Raj, "Data pre-processing techniques in data mining: A Review," International Journal of Innovative Research in Computer Science &amp; Technology, pp. 122–125, 2022.

[19] M. D. Anto Praveena and B. Bharathi, "Removal of outliers and missing values in diabetes dataset using ensemble method," Advances in Data Science and Management, pp. 335–342, 2022.

[20] H. K. Butler, M. A. Friend, K. W. Bauer and T. J. Bihl, "The effectiveness of using diversity to select multiple classifier systems with varying classification thresholds," Journal of Algorithms &amp; Computational Technology, vol. 12, no. 3, pp. 187–199, 2018.

[21] M. Keerthana, K. J. Meghana, S. Pravallika and M. Kavitha, "An ensemble algorithm for crop yield prediction," Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 04-06 February 2021. DOI: 10.1109/icicv50876.2021.9388479

[22] J. Zou, X. Fu, L. Guo, C. Ju and J. Chen, "Creating ensemble classifiers with Information Entropy Diversity Measure," Security and Communication Networks, vol. 2021, pp. 1–11, 2021.

[23] A. O. Akyuz, M. Uysal, B. A. Bulbul and M. O. Uysal, "Ensemble Approach for Time Series Analysis in Demand Forecasting: Ensemble learning," IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), 03-05 July 2017. DOI: 10.1109/inista.2017.8001123.

[24] J. Zhao, X. Yang and S. Sun, "Constraints on maize yield and yield stability in the main cropping regions in China," European Journal of Agronomy, vol. 99, pp. 106–115, 2018.

[25] K. Saito, J. Six, S. Komatsu, S. Snapp, T. Rosenstock, A. Arouna, S. Cole, G. Taulya and B. Vanlauwe, "Agronomic gain: Definition, approach and application," Field Crops Research, vol. 270, pp. 1-13, 2021.

[26] E. M. M. van der Heide, C. Kamphuis, R. F. Veerkamp, I. N. Athanasiadis, G. Azzopardi, M. L. van Pelt and B. J. Ducro, "Improving predictive performance on survival in dairy cattle using an ensemble learning approach," Computers and Electronics in Agriculture, vol. 177, pp. 1-10,2020.

[27] M. Shahhosseini, G. Hu and S. V. Archontoulis, "Forecasting corn yield with machine learning ensembles," Frontiers in Plant Science, vol. 11, pp.1-16, 2020.

[28] T. Wu, W. Zhang, X. Jiao, W. Guo and Y. Alhaj Hamoud, "Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration," Computers and Electronics in Agriculture, vol. 184, pp. 26-42, 2021.

[29] Y. Ren, L. Zhang and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions [review article]," IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 41–53, 2016.

[30] M. Shahhosseini, R. A. Martinez-Feria, G. Hu and S. V. Archontoulis, "Maize yield and nitrate loss prediction with machine learning algorithms," Environmental Research Letters, vol. 14, no. 12, 2019. DOI: 10.1088/1748-9326/ab5268

[31] S. Motia and S. R. N. Reddy, "Ensemble classifier to support decisions on Soil Classification," IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, 2021. DOI:10.1088/1757-899X/1022/1/012044

[32] X. Li, N. Liu, L. You, X. Ke, H. Liu, M. Huang and S. R. Waddington, "Patterns of cereal yield growth across China from 1980 to 2010 and their implications for food production and food security," PLOS ONE, vol. 11, no. 7, 2016. DOI:10.1371/journal.pone.0159061.

[33] K. Jain and N. Choudhary, "Comparative analysis of machine learning techniques for predicting production capability of crop yield," International Journal of System Assurance Engineering and Management, vol. 13, no. S1, pp. 583–593, 2022.

[34] R. Chapman, S. Cook, C. Donough, Y. L. Lim, P. Vun Vui Ho, K. W. Lo and T. Oberthür, "Using Bayesian networks to predict future yield functions with data from commercial oil palm plantations: A proof of concept analysis," Computers and Electronics in Agriculture, vol. 151, pp. 338–348, 2018.

[35] S. Jing, Y. Wang and L. Yang, "Selective Ensemble of Uncertain Extreme Learning Machine for pattern classification with missing features," Artificial Intelligence Review, vol. 53, no. 8, pp. 5881–5905, 2020.

[36] R. Mouhssine, A. Otman and E. khatir, "Performance Analysis of Machine Learning Techniques for smart agriculture: Comparison of supervised classification approaches," International Journal of Advanced Computer Science and Applications, vol. 11, no. 3, pp. 610-619, 2020.

[37] A. Adebanji, M. Asamoah-Boaheng and O. Osei-Tutu, "Robustness of the quadratic discriminant function to correlated and uncorrelated

normal training samples," SpringerPlus, vol. 5, no. 1, 2016. DOI 10.1186/s40064-016-1718-3.

[38] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz and J. H. Moore, "PMLB: A large benchmark suite for machine learning evaluation and comparison," BioData Mining, vol. 10, no. 1, 2017. DOI 10.1186/s13040-017-0154-4.

[39] A. Tasnim, M. Saiduzzaman, M. A. Rahman, J. Akhter, and A. S. Rahaman, "Performance evaluation of multiple classifiers for predicting fake news," Journal of Computer and Communications, vol. 10, no. 09, pp. 1–21, 2022.

[40] A. Pajankar and A. Joshi, "Introduction to machine learning with Scikit-Learn," Hands-on Machine Learning with Python, pp. 65–77, 2022.

[41] S. Lynch, "A tutorial introduction python," Dynamical Systems with Applications using Python, pp. 1–31, 2018.

[42] J. R. Romero, P. F. Roncallo, P. C. Akkiraju, I. Ponzoni, V. C. Echenique and J. A. Carballido, "Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires," Computers and Electronics in Agriculture, vol. 96, pp. 173–179, 2013.

[43] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," IEEE Access, vol. 9, pp. 63406–63439, 2021.

[44] S. S. Dhaliwal, V. Sharma and G. Verma, "Agronomic strategies for improving micronutrient use efficiency in crops for nutritional and food security," Input Use Efficiency for Food and Environmental Security, pp. 123–156, 2021.

[45] M. Zhang, X. Yuan and J. A. Otkin, "Remote Sensing of the impact of flash drought events on terrestrial carbon dynamics over China," Carbon Balance and Management, vol. 15, no. 1, 2020. DOI:10.1186/s13021-020-00156-1.

[46] Y. Chen, X. Han, W. Si, Z. Wu, H. Chien and K. Okamoto, "An assessment of climate change impacts on maize yields in Hebei province of China," Science of The Total Environment, vol. 581-582, pp. 507–517, 2017.

[47] X. Han, F. Liu, X. He and F. Ling, "Research on rice yield prediction model based on Deep Learning," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–9, 2022.