Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction

Hayam R. Seireg¹, Yasser M.K. Omar², Fathi E. Abd El-Samie³, Adel S.El-Fishawy⁴, Ahmed Elmahalawy⁵

ABSTRACT Precision agriculture is a challenging task to achieve. Several studies have been conducted to forecast agricultural yields using machine learning algorithms (MLA), but few studies have used ensemble machine learning algorithms (EMLA). In the current study, we used a dataset generated by a computer simulation program, and meteorological data obtained over 30 years ago from Maine, United States (USA). The primary goal of this research is to increase the forecast accuracy of the best characteristics for overcoming hunger challenges. We designed stacking regression (SR) and cascading regression (CR) with a novel combination of MLA based on the wild blueberry dataset. We used features that indicated the best regulation for wild blueberry agroecosystems. The four feature engineering selection techniques are applied variance inflation factor (VIF), sequential forward feature selection (SFFS), sequential backward elimination feature selection (SBEFS), and extreme gradient boosting based on feature importance (XFI). We applied Bayesian optimization on popular MLA to obtain the best hyperparameters to achieve accurate wild blueberry yield prediction. The SR used a two-layer structure: level-0 contained light gradient boosting machine (LGBM), gradient boost regression (GBR), and extreme gradient boosting (XGBoost); level-1 provided the output prediction using a Ridge. The (CR) topology is the same MLA used in SR, but in a series form that takes the new prediction as a feeder to each MLA and removes the previous prediction in each stage. We assessed many techniques, CR, and SR outcomes regarding the root mean square error (RMSE) and coefficient of determination (R²). In the results, the proposed SR showed the best performance 0.984 R² and 179.898 RMSE compared with another study that published 0.938 R² and 343.026 RMSE on the seven features selected by XFI. The SR achieved the highest 0.985 R² on all features and the features that were selected by SBEFS. Our SR outperformed CR, many other techniques, and another study on wild blueberry yield prediction.

INDEX TERMS Bayesian optimization, cascading technique, GBR, LGBM, Ridge, stacking technique, XGBoost, EMLA, wild blueberry yield.

I. INTRODUCTION

The wild blueberry crop is Maine’s most important fruit crop, growing in upland acidic sandy soils. The wild blueberry crop (also known as lowbush blueberries) is the most important crop for most people and is considered the largest producer among the other crops in the USA [1]. The main advantage of wild blueberries over cultivated blueberries is that they are not planted but grow naturally in rocky hills and forests. Cultivated blueberries, on the other hand, must be planted [2]. Wild blueberries produce diverse crops, as they are protected from selective breeding and genetic engineering. Hand harvesters use rakes to scrape berries from plants [3] because of the rough terrain and altitude of wild blueberries. The best time to pick wild blueberries in Maine is from the end of July to the beginning of September [4]. Insect pollination species and weather conditions are the major factors affecting wild blueberry yields [5]. The number of insect pollinators, the optimal time for wild blueberry pollination (around 6 am – 5 pm), bee species (honeybee (HB), bumblebee (BB), Andrena
(AD), and Osmia(OS)), and weather conditions (cold temperatures, wet weather, strong winds, frost, and warmer temperatures) are factors influencing the wild blueberry yield [6]. Obtaining reliable crop yield predictions to meet the Food and Agriculture Organization (FAO) targets is difficult. The primary objective of FAO is to eradicate global hunger. Many scientists have focused on precision agriculture, forecasting crop yields using machine learning (ML) methodologies [7-10]. However, few studies have used EMLA with diverse topologies [11-12]. MLA require large and high-quality datasets to accurately estimate crop yield [13]. Modeling agricultural production systems is essential, especially when collecting large-scale temporal and geographical data is prohibitively expensive, complex, or impossible in some cases. In many studies, quality data used during training insufficient, may affect experimental results [14], and this degrades crop yield output forecast accuracy [15]. To address these issues, computer simulation modeling approaches that simulate measurements based on scenes to create data have been used [16]. To validate a computer simulation model, model’s the assumptions are tested. Sensitivity analysis is used for data assumptions. Sensitivity analysis has been used to demonstrate the interactive impact of bee community dominance and meteorological variables on wild blueberry output [17]. The Naval Postgraduate Schools Simulation, Experiments, and Efficient Design Center for Data Farming defines data farming as the practice of accumulating datasets through simulations and computer modelling. Thus, data farming aims to give decision-makers insights into complicated challenges through simulations, generating data [18].

On this premise, we used previously validated simulation-based modelling of wild blueberry pollination [16]. Obsie et al [17] utilized computational experiments to generate training data, which may subsequently be used to train and assess MLA.

For empirical datasets, EMLA have outperformed individual MLA in terms of accuracy. The major benefit of using EMLA is that they can solve difficult problems [19]. The fundamental idea behind using EMLA is to convert a weak learner into a powerful one to improve decision support [20]. A few studies have combined various ML techniques to achieve high accuracy. Stacking techniques are used to implement classification models such as PredT4SE-Stack [21] and NeuroPpred-Fuse [22], as well as weighted multiple metamodel stacking methods [23] for regression. Several techniques with different architectures have been used to achieve precision agriculture through crop yield prediction [12]. The stacking technique obtains all prediction outputs from other MLA and feeds them to the metamodel to obtain the best decision. The researcher [17] believes that simulation data combined with climate data may be useful for training MLA because these data are of high quality and produce datasets difficult to collect manually. The major goal is to avoid costly real-time datasets obtained from satellite images, with an average cost of US $1/km² [24].

A previous study has used computer modeling to synthetically generate data for training a wild blueberry yield prediction model, which was then integrated with meteorological data from Maine, USA. A previous study selected seven characteristics [17].

In this study, we used the stacking and cascading techniques to accurately estimate wild blueberry yield prediction based on unique subgroup criteria. Meta-learning is a technique that integrates the predictions of many MLA to develop an EMLA approach. Meta-learning requires an MLA to be trained and tested using data before feeding the output prediction into another MLA.

Our research focuses on predicting wild blueberry yield using EMLA with various methodologies and computer simulation data. Our main goal was to develop a set of MLA using the stacking and cascading approaches to improve prediction accuracy and reduce errors.

The main contributions of this study are summarized.

1. We tuned MLA hyperparameters by Bayesian optimization to obtain the highest accuracy of each prediction model.
2. We developed a novel combination of MLA to forecast wild blueberry yield using the cascading and stacking techniques. The cascading technique is used in a new way that improves the model performance.
3. We integrate feature selection techniques with our proposed SR and CR such as VIF, SFFS, SBEFS and XFI.
4. The performance of the cascading and stacking techniques was evaluated using a different subset of characteristics chosen through feature engineering approaches compared with many MLA.

The remainder of this paper is structured as follows: Section 2 discusses comparable efforts, and Section 3 discusses the data used and the data preprocessing procedure, as well as the benefits of different feature selection strategies. Section 4 introduces the two proposed models and their implementation. The association between the provided attributes and predictions was identified using Shapley additive exPlanations(SHAP). Section 5 presents the MLA used in the two proposed models. Section 6 describes methods for evaluating the performance of each MLA, including an explanation of the benefits of Bayesian optimization. Section 7 evaluates the performance of many ensemble techniques, including CR and SR, and compares the experimental results of the two proposed models with other algorithms. Section 8 outlines the findings and gap study. This study has some
limitations. Finally, Section 9 concludes the study briefly and discusses future work.

II. RELATED WORKS

In developing countries, crop yields are typically manually estimated. However, other countries have focused on using emerging technologies including MLA and deep learning algorithms, to predict crop yields.

Over 30 years on India’s west coast, time-series data on rice yield paired with meteorological characteristics were collected. Four MLA are used for predicting rice yield. Overall, the least absolute shrinkage and selection operator (LASSO) achieved the best result [25].

The hybrid multiple linear regression (MLR) - artificial neural network (ANN) exhibits [26] that, rather than arbitrary weight and bias initialization, input layer weights and bias are initialized using MLR coefficients and bias. The planting area, water system region, fertilizer consumption, and water system points of interest were included in the generated agricultural data. Climate precipitation, highest and lowest temperatures, and solar radiation were among the climatological characteristics included in the dataset. MLR–ANN was applied to these datasets to precisely predict paddy crop yield.

Five MLA, namely, linear regression (LR), LASSO, XGBoost, LGBM, and random forest (RF), were combined using an optimized weighted ensemble to predict corn yield data, including crop management, climate, soil conditions, and historical yields. The five MLA combined with an optimized weighted ensemble outperformed existing MLA [12].

Shahhosseini et al. [27] presented a new ensemble convolutional neural network (CNN) – deep neural networks to predict corn yields: a dataset containing planting date, weather data, soil characteristics, and historical corn yields.

Cereal yield was predicted using MLA trained with a dataset containing weather data, remote sensing, climate, and yield data [28], and the dataset was combined with satellite data and historical yield. These four datasets were combined to form a single dataset. One linear MLA and three nonlinear MLA were used. XGBoost outperformed other MLA.

Linear and nonlinear models were used to estimate soybean and corn yields using Landsat and SPOT images. Two models, neural network and MLR, have been examined [29], and the findings revealed that when all images were included, the models created a strong match, exceeding neural networks in terms of accuracy.

Khaki et al. [30] YieldNet is a contemporary CNN model that combines transfer learning between corn and soybean yield prediction by switching spine weights, including the extractor, using a new deep learning method. The dataset was created using remote sensing data, allowing for continuous crop monitoring throughout the growth cycle.

Datasets were created using a wild blueberry pollination simulation model. This information comprises four species of bees, their clone sizes (CSs), and weather data. Four MLA were used to predict wild blueberry yields. XGBoost [17] achieved the best performance across all evaluation measures. Many published research papers have been reviewed and their findings summarized. Few studies have used simulation data to predict crop yield [17],[31], and over 5 years, the Agricultural Production Systems Simulator provided maize yield, soil, and N fertilizer data for the US Midwest [31].

III. DATA SOURCES AND SELECTION

A. EXTRACTED DATA

The simulation model was used to study the influence of wild blueberry placement, bee species and their density, and weather fluctuations on the yield. The dataset used comprised data from an open-source computer simulation model and meteorological information collected from Maine, USA, and the Canadian Maritimes over the last 30 years [16]. Meteorological data for the areas of Bangor, Maine, USA, were collected between 2015 and 2019, and the five-year data were averaged to be feature inputs. Obsie et al [17] deliberately changed the weather data to 125 percent and 75 percent of the average daily temperature and rainy-day data daily from May 1 to June 30 over the past five years. These factors are the main targets to be obtained from model simulations. The number of levels for each factor was provided. Among the characteristics generated by the simulation trials were the following: CS; HB, BB, AD, and OS densities; daily air temperature; and daily precipitation. CS ranges from 10 to 40 m², and different density ranges should be provided for different bee taxa. For field statistics, the levels of each factor were calculated. The levels of (i.e., the number of sample spaces within) each component have been computed, which are as follows: six levels for CS, seven for HB density; ten for BB density, twelve for AD and OS bee densities, and three for weather data. Obsie et al [17] performed three distinct random Latin hypercube sampling (RLHs) [33] with various boundary values to provide excellent space-filling qualities, and RLHs helped in reducing the number of simulations. The total number of simulations performed was 77700. After applying the RLHs, the number of simulations was reduced to 777.

B. MODEL SIMULATION FEATURES

Model simulation is used to generate features that can be utilized in a prediction model. Weather conditions were obtained by simulation experiments that had to be translated.
into features, and the weather conditions were measured as follows: all temperatures are given in degrees Fahrenheit, whereas rain days are measured in inches (""). The upper and lower temperature ranges differed. The maximum upper temperature range (MaxUTR) is from 69.7 °F to 94.6 °F however, the minimum upper-temperature range (MinUTR) varies from 39.0 °F to 57.2 °F, and the average upper-temperature range (AvUTR) varies from 58.2 °F to 79.0 °F during the simulation flowering period. The maximum lower temperature range (MaxLTR) is 50.2 °F to 68.2 °F, the minimum lower temperature range (MinLTR) is 24.3 °F to 33.0 °F, and the average lower temperature range (AvLTR) is 41.2 °F to 55.9 °F during the simulation flowering period. The aggregate counted days throughout the simulation flowering period are represented by raining days (RD), which range from 1 " to 34 ", and average raining day (AvRD) range from 0.06 " to 0.56 " throughout the simulation flowering period. The densities of the four bee species were measured in bees/m2/min, and their initial values were zero. The maximum density for AD and OS was 0.75, that for HB was 18.43 and that for BB was 0.585. The yield of wild blueberries changes between 1637.70 and 8969.40 kg/ha.

C. FEATURE SELECTION

The main target of using feature engineering techniques is as follows:
(1) Some features are unrelated to the output target, which is the fundamental goal of feature selection.
(2) Noisy features are removed.
(3) The accuracy is improved.

The major contribution of feature selection approaches is the removal of characteristics significantly degrade performance. M1 represents all features in the dataset, as shown in Figure 1; However, many approaches and methodologies, including filters, wrappers, and embedding methods, should be used to obtain a subset of features to achieve the highest prediction accuracy and speed up the various MLA. Filtering techniques were used in the data preparation process. MLA was used to choose the key features based on the accuracy of the target predicted feature, and statistical tests were conducted to identify the correlation between the input features and the anticipated output target.

![Feature Selection Techniques Diagram](image)

Figure. 1: Represents the features that are selected by different techniques.

The VIF was used to detect collinearity and multicollinearity. Higher values indicate that precisely assessing the contribution of predictors to a model is difficult or impossible. VIF yielded nine features represented in M4, SFFS yielded 10 features represented in M2, and SBEFS yielded 10 features represented in M3 as shown in Figure 1. Both feature selection techniques were used as wrapper techniques. An ML approach based on wrapper techniques was used to select the best characteristics. Embedded techniques combine the benefits of filtering and wrapper methods. Based on feature significance, XGBoost based on XFI yielded seven features represented in M5. Obsie et al. [17] used various techniques for feature engineering selection. The optimal features that may improve performance were obtained.
IV. TWO PROPOSED MODELS

The workflow of the two models implemented is as follows:

(1) Important features were selected using the VIF, SFFS, SBEFS, and XFI techniques.

(2) Bayesian optimization was applied to the MLA to tune hyperparameters for boosting the EMLA.

(3) The two proposed models were examined for all features and the different subgroups of features chosen using different feature engineering techniques.

(4) The target was compared and validated by verifying the model performance.

A. STACKING REGRESSION TECHNIQUE

Stacking is the process of training several MLA to forecast the target variable while simultaneously learning how to use the predictions of each MLA to anticipate the target variable’s value. Stacking is a popular EMLA technique. Stacking methods employed a hierarchical approach, with MLA separated into two levels, level-0 and level-1. We performed several iteration combinations with different MLA using the stacking technique and the best combination was chosen based on the wild blueberry yield dataset.

LGBM, GBR, and XGBoost generate predictions for each MLA separately, and the three new features are added to the dataset. After training Ridge with the new dataset, Ridge obtains the final prediction. The LGBM, GBR, and XGBoost should be at level-0, and the output prediction of those MLA should feed Ridge level-1 before obtaining the final output prediction from Ridge as shown in Figure 2. EMLA should improve the accuracy and reduce the RMSE; the MLA are chosen based on their advantages in solving real-world problems.

Figure 2. Our proposed stacking technique with (LGBM), (GBR), (XGBoost), (Ridge).

Figure 3. Flowchart of our proposed stacking technique.
The stacking technique depends on increasing the number of input features in the last MLA, which is called a metamodel for improving performance. First, the wild blueberry dataset consists of 13 features and 777 rows. To minimize the number of features in the dataset, feature engineering approaches are utilized, and various strategies are used to the dataset. Each technique selects different subgroup characteristic features. Datasets containing that had all features and every subset of features were divided into 80% training and 20% test sets. The Bayesian optimization algorithm is applied. Stacking was established with LGBM, GBR, XGBoost, and Ridge was applied to every subgroup of features selected, and all features are shown in Figure 3. We evaluated the performance of SR and other techniques according to popular metrics.

B. CASCADING REGRESSION TECHNIQUE

The cascading technique employed MLA at each stage of the MLA while sequentially feeding the other MLA. Each MLA’s prediction is the feeder and at each step, the MLA is trained on the dataset with the newly predicted feature, while removing the previously predicted feature. Consequently, in each stage, the MLA included only one new predicted feature in the dataset. Finally, the MLA forecasts the outcomes as shown in Figure 4.

Figure 4. Our proposed Cascading technique with (LGBM), (GBR), (XGBoost), (Ridge).

Figure 5. Flowchart of our proposed cascading technique.
The cascading combination uses the same MLA as the stacking techniques used for comparison. Datasets are split into 80% training and 20% test sets. Then Bayesian optimization applied to each MLA. After these procedures, the cascading technique is applied to different subsets of features and 13 features as shown in Figure 5. Finally, the performance of the cascading technique and many other techniques was compared.

C. Advantage of SHAP

SHAP [32] is presented as a strategy for converting the blackbox of MLA into a form representing the effect of each feature on wild blueberry prediction. The ratings are based on significant contributions to a certain feature. SHAP was applied to the 13 features in our study.

The primary benefits of using SHAP are as follows:

(1) Relationship between specified characteristics and prediction.

(2) Investigation of the factors influencing predictions to obtain more specific information; and

(3) Interpretation of the blackbox of the MLA.

The technique used XGBoost in SHAP. In Figure 6, according to SHAP, red indicates the greater value for each selected feature on its unit scale, whereas blue indicates the lower value. An increase in RD, MaxUTR, CS, and HB values negatively affect the yield prediction, whereas an increase in OS, BB, and AD values have a positive impact on wild blueberry yield prediction, and the relationship between OS, BB, and AD and wild blueberry yield prediction is directly proportional.

According to Figure 6, reducing RD, CS, and HB values increases the wild blueberry yield prediction accuracy, indicating an inverse proportional relationship. The x-axis represents the SHAP value (impact on the model output), and the y-axis represents the feature value.

An increase or decrease in MaxLTR, MinLTR, and AvLT values did not affect wild blueberry yield prediction. The features are arranged according to the main features that have a significant impact on wild blueberry yield prediction. An increase in AvUTR, and MinUTR values decreases the target output.

Figure 6. SHAP represents the impact of features on the prediction.

V. MACHINE LEARNING MODELS

A. BASIC MODELS

In this study, LGBM, GBR, XGBoost, and Ridge were combined using the stacking and cascading techniques.

B. Light Gradient Boosting Machine

LGBM [33] is a boosting ensemble model that builds a potentially effective model by combining linked weak learners. LGBM enhances the capabilities of gradient-boosted decision tree models by increasing running time and decreasing memory usage while maintaining excellent accuracy. While other tree algorithms develop horizontally, the LGBM algorithm grows vertically, which means that it undergoes leaf-wise tree development to minimize overfitting. However, other tree algorithms grow level-wise. LGBM selects the leaf with the highest growth loss. When expanding the same leaf, the loss can be reduced by more than that of a level-wise method.

C. Gradient Boosted Regression

GBR [34] is one of the most popular algorithms among the MLA used in this study. It minimizes bias and variance. GBR
is a stage-wise additive model that constructs learners while learning. Existing trees in the model were not modified, and new trees were inserted sequentially. The new weak learner focuses on areas that influence performance. The contribution of each weak learner to the final prediction was based on a gradient optimization method designed to reduce the total error of the learner.

D. Extreme Gradient Boosting

Many agricultural researchers have used this ensemble boost method. XGBoost [35] uses decision trees as foundation learners and combines many weak learners to generate a strong learner. Consequently, it is called an ensemble learning approach because the final prediction integrates the outputs of numerous models. It works well on small and medium-sized datasets.

E. Ridge

Ridge regression [36] is extension of linear regression that modifies the loss function to reduce the model complexity. This is accomplished by introducing a penalty parameter that is equal to the square of the coefficient magnitudes. OLS + alpha × summation is the loss function (squared coefficient values)
The alpha parameter should be chosen in the loss function. A low alpha value can result in overfitting, whereas a high alpha value can result in underfitting.

VI. PERFORMANCE EVALUATION APPROACHES

A. BAYESIAN OPTIMIZATION OF HYPERPARAMETERS

Bayesian optimization [37] generates a posterior distribution of functions (Gaussian process) that best characterizes the to-be-improved function. The posterior distribution improves as the number of observations increases, and the algorithm becomes more convinced of which regions of the parameter space are worth exploring and which are not. The algorithm iterates to balance its exploration and exploitation requirements while considering what it learns about the target function. A Gaussian process is fitted to the known samples (previously examined points) at each step, and the posterior distribution, together with an exploration strategy (such as upper confidence bound or expected improvement), is used to decide the next point to be researched.

The key advantage of using Bayesian optimization is to achieve the best accuracy by selecting the appropriate hyperparameters for each MLA; most studies have concluded that Bayesian optimization is faster than grid search [38], and that grid search achieves higher accuracy than Bayesian optimization, but takes longer to find optimal hyperparameters.

We tuned the hyperparameters of each MLA. To reduce the hyperparameters of each MLA, as shown in Tables. 1–17, we used the split train-test function (training set 80% and test set 20%) for each MLA using Bayesian optimization.

| TABLE 1. REPRESENTS THE BEST HYPERPARAMETER FOR STOCHASTIC GRADIENT DESCENT (SGD). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| learning_rate                 | 'optimal'|
| loss                          | 'epsilon_insensitive'|
| penalty                       | 'l1'     |

| TABLE 2. REPRESENTS THE BEST HYPERPARAMETER FOR KERNEL RIDGE (KR). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| alpha                         | 1.0027362833183082 |

| TABLE 3. REPRESENTS THE BEST HYPERPARAMETER FOR SUPPORT VECTOR REGRESSION (SVR). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| C                             | 999.9330399497585 |
| gamma                         | 'auto'   |

| TABLE 4. REPRESENTS THE BEST HYPERPARAMETER FOR LINEAR REGRESSION (LR). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| fit_intercept                 | True     |
| normalize                     | True     |
| copy_X                        | True     |
| n_jobs                        | None     |

| TABLE 5. REPRESENTS THE BEST HYPERPARAMETER FOR ELASTIC NET (EN). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| alpha                         | 0.05317254699201772 |
| l1_ratio                      | 0.9997328488140344 |
| max_iter                      | 585364   |
| selection                     | 'random' |

| TABLE 6. REPRESENTS THE BEST HYPERPARAMETER FOR BAYESIAN RIDGE (BR). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| n_iter                        | 7288     |

| TABLE 7. REPRESENTS THE BEST HYPERPARAMETER FOR THE LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO). |
|-------------------------------|----------|-----------------|
| HYPERPARAMETER                | VALUES   |
| selection                     | cyclic   |
| alpha                         | 1.002533096859825 |
| max_iter                      | 29988    |

| TABLE 8. REPRESENTS THE BEST HYPERPARAMETER FOR RIDGE. |
|-------------------------------|----------|-----------------|
| Hyperparameter                | VALUES   |
| alpha                         | 1.002502854 |
| max_iter                      | 5546     |
| solver                        | 'lsqr'   |
TABLE 9. REPRESENTS THE BEST HYPERPARAMETER FOR MULTI-LAYER PERCEPTRON REGRESSOR (MLPR).

| HYPERPARAMETER  | VALUES                     |
|-----------------|----------------------------|
| alpha           | 0.795078277961322          |
| learning_rate   | ‘invscaling’               |
| max_iter        | 77563                     |
| shuffle         | False                     |
| solver          | ‘lbfgs’                   |

TABLE 10. REPRESENTS THE BEST HYPERPARAMETER FOR ADAPTIVE BOOSTING (AdaBoost).

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| learning_rate   | 1.5             |
| loss            | ‘square’        |
| n_estimators    | 1285            |

TABLE 11. REPRESENTS THE BEST HYPERPARAMETER FOR DECISION TREE REGRESSOR (DTR).

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| criterion       | ‘mae’           |
| max_depth       | 3050.0          |
| max_features    | ‘auto’          |
| min_samples_leaf| 3               |
| splitter        | ‘random’        |

TABLE 12. REPRESENTS THE BEST HYPERPARAMETER FOR K NEAREST NEIGHBOR (KNN).

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| algorithm       | ‘ball_tree’     |
| leaf_size       | 64              |
| n_neighbors     | 10              |
| p               | 1               |
| weights         | ‘distance’      |

TABLE 13. REPRESENTS THE BEST HYPERPARAMETER FOR LGBM.

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| bagging_fraction| 0.667417297    |
| bagging_freq    | 6               |
| boosting        | ‘gbdt’          |
| max_depth       | 5               |
| num_leaves      | 65443           |

TABLE 14. REPRESENTS THE BEST HYPERPARAMETER FOR RANDOM FOREST (RF).

| HYPERPARAMETER  | VALUES         |
|-----------------|----------------|
| max_depth       | 940.0          |
| min_samples_leaf| 2              |
| min_samples_split| 4           |
| n_estimators    | 402            |

TABLE 15. REPRESENTS THE BEST HYPERPARAMETER FOR EXTRA TREES REGRESSOR (ETR).

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| criterion       | ‘mae’           |
| max_depth       | 100.0           |
| min_samples_leaf| 2              |
| min_samples_split| 5           |
| n_estimators    | 9271           |

TABLE 16. REPRESENTS THE BEST HYPERPARAMETER FOR GBR.

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| criterion       | mse             |
| max_depth       | 3000            |
| max_features    | auto            |
| min_samples_leaf| 5              |
| min_samples_split| 45            |
| n_estimators    | 906             |

TABLE 17. REPRESENTS THE BEST HYPERPARAMETER FOR XGBOOST.

| HYPERPARAMETER  | VALUES          |
|-----------------|-----------------|
| base_score      | 0.5             |
| booster         | ‘gbtree’        |
| colsample_bylevel| 1            |
| colsample_bynode| 1              |
| colsample_bytree| 0.729795278     |
| gamma           | 0.002979511     |
| gpu_id          | -1              |
| importance_type | ‘gain’          |
| interaction_constraints | ‘()’          |
| n_estimators    | 100             |
| n_jobs          | 8               |
| num_parallel_tree| 1             |
| reg_alpha       | 0               |
| reg_lambda      | 1               |
| scale_pos_weight| 1              |
| subsample       | 1               |
| tree_method     | ‘exact’         |
| validate_parameters| 1            |
| verbosity       | None            |

B. Performance Metrics

We examined 17 regression MLA and the stacking and cascading approaches using four measurable formula equations.

1. \( R^2 \) [39] is the coefficient of determination, which is used to determine how much variance in the response variable can be explained by the independent variables.

2. The difference between the expected and actual values was calculated using RMSE.

3. The absolute mean of the real and predicted yield values is the mean absolute error (MAE).

4. The relative root mean squared error (RRMSE) was obtained by dividing the RMSE by the real-value average.
The formula equations were obtained for RMSE and MAE [40], while RRMSE was determined according to

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \]

\[ \text{MAE} = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n} \]

where \( n \) is the count of the real yield, \( Y_i \) is the real yield value, and \( \hat{Y}_i \) is the MLA yield prediction for wild blueberries. \( R^2 \) is calculated as follows:

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \]

\[ \text{RRMSE} = \frac{\text{RMSE}}{\bar{Y}} \times 100 \]

\( \bar{Y} \) is the mean real yield when the value of \( R^2 \) is close to 1.0, and the relationship between the MLA forecasted and real yield is more linear. An RMSE of zero shows that the MLA output yield corresponds perfectly to the real yield.

The number of independent factors used to predict the target variable is taken into consideration by the Adjusted R-squared (Adj \( R^2 \)) [41]. This allows us to see if adding new variables to the model improves model fit.

\[ \text{Adj R}^2 = \left( 1 - \frac{\left( 1 - R^2 \right) (n-1)}{(n-k-1)} \right) \]

The number of data points in our dataset is \( n \), the number of independent variables is \( k \).

As a result, if the Adj \( R^2 \) value does not increase sufficiently because of the inclusion of a new independent variable, the value of \( R^2 \) will decline.

### VII. RESULTS

We conducted intensive experiments to determine the best MLA combination using the stacking and cascading techniques. We also evaluated the performance of these two techniques with other traditional and modern MLA. The best hyperparameters were selected using Bayesian optimization to improve accuracy.

#### A. COMPARISON OF THE PERFORMANCE OF 17 MLA, CR, AND SR

Tables 19.1-19.3 summarize the performance of the 17 MLA, CR, and SR for different input features selected using feature engineering selection techniques. The specifications of the computer used in these experiments are as follows: AMD Phenom™ || X6 1100T Processor 3.30 GHz, 8.00 Gigabyte RAM, and Windows 10.

The wild blueberry dataset was split into 80% training and 20% test sets using the split–train–test function in the scikit-learn package. The with the training dataset consisted of 621 randomly chosen records; the test set comprised 156 records. We developed a novel MLA combination using the stacking and cascading techniques to predict wild blueberry yield using the dataset generated through simulations. Each MLA and the stacking, and cascading techniques were compared in terms of \( R^2 \), RMSE, and RRMSE for all features and each different subset of features, as we examined the M1, M2, M3, M4, and M5 datasets.

The findings revealed that CR, GBR, XGBoost, and SR had the lowest RMSE values of 227.88, 210.339, 208.542, and 173.068 kg/ha on M1 features, respectively. This shows that SR can predict more accurately than all traditional models and modern MLA. The MAE of CR, GBR, and SR values varied between 133.854 and 169.674 kg/ha. SR had the highest \( R^2 \) value, 0.985 on M1 dataset.

Concerning the RRMSRE, SR outperformed the 17 MLA and CR. Concerning RRMSE, MLA accuracy is excellent when the RMSE is less than 10%, it is good if the RMSE is greater than or equal to 10 percent and less than 20%, it is fair if the RMSE is greater than or equal to 20% and less than 30%, and it is poor if the RRMSE is greater than or equal to 30 percent. Therefore, in terms of RRMSE, SR, CR, XGBoost, GBR, ETR, RF, LGBM, KNN, DTR, AdaBoost, MLPR, Ridge, Lasso, BR, EN, LR, and SVR are classified as excellent for the M1, M2, M3, M4, and M5 datasets, whereas KR and SGD are classified as good for the M1, M2, M4, and M5 datasets; however, KR is classified as good and SGD is classified as poor for the M3 dataset. The findings revealed that XGBoost, CR, GBR, and SR reduced RMSE values from 147.47 to 227.88 kg/ha for the M1, M2, M3, M4, and M5 datasets.

For the M1, M2, M3, M4, and M5 datasets, the MAE scores for CR, GBR, XGBoost, and SR ranged from 132.38 to 169.674 kg/ha. When we assessed all MLA, CR, and SR, we discovered that SR had the best \( R^2 \) values, ranging from 0.984 to 0.985.

The second best-performing MLA, in terms of \( R^2 \), was for a different input dataset; GBR had \( R^2 \) of 0.978 for the M1, M2, and M3 datasets, but CR had 0.979 on the M4 dataset, and XGBoost achieved 0.978 to 0.981 for the M1 and M5 datasets.

The third best performing model, CR had \( R^2 \) ranging from 0.974 to 0.979 on the M1, M2, M3, and M5 datasets, whereas GBR had 0.978 on the M4 dataset.
The fourth best performing model, XGBoost had $R^2$ ranging from 0.975 to 0.976 for the M2, M3, and M4 datasets, whereas GBR had 0.978 for the M5 dataset. However, SGD was inaccurate, with the highest RMSE, ranging from 1265.228 to 2330.476 kg/ha and the highest MAE score, ranging from 1041.074 to 1881.774 kg/ha for the M1, M2, M3, M4, and M5 datasets. This illustrates that SGD was not accurate compared with the other 17 MLA, CR, and SR.

In comparison to the 17 MLA and cascade approaches, the stacking technique had the highest $R^2$ and the best RMSE and MAE values.

This study extensively investigated more MLA to improve prediction performance. The results in Figure 7 show that SBEFS was the best feature selection technique. Among all datasets, the M3 dataset had the highest adj $R^2$. Adj $R^2$ decreases when more irrelevant variables are included in a model. Adj $R^2$ increases when more relevant variables are included. $R^2$ after adjustment will always be smaller than or equal to $R^2$.

From Table 18.1 to 18.3, CR achieved the highest adj $R^2$ for M4 dataset compared to the other four ML that cooperate in the combination. In summary, the SR approach outperformed the 17 MLA and CR in this study on the M1, M2, M3, M4, and M5 datasets. SR was more accurate than CR.

![Figure 7](image-url)  
*Figure 7. Represent the Adj $R^2$ and $R^2$ were used to assess the performance of the SR on the dataset: M1, M2, M3, M4, and M5.*

| Model | M1 | M2 | M3 | M4 | M5 |
|-------|----|----|----|----|----|
| LGBM  | 0.955315 | 0.955315 | 0.954999 | 0.954999 | 0.953201 |
| GBR   | 0.974387 | 0.973916 | 0.975089 | 0.974731 | 0.973510 |
| XGB   | 0.974276 | 0.973808 | 0.97759 | 0.977276 | 0.979510 |
| Ridge | 0.974274 | 0.973801 | 0.977592 | 0.977271 | 0.979510 |
| CR    | 0.974274 | 0.973767 | 0.977593 | 0.97724 | 0.979510 |

Table 18.1. Represent the Adj $R^2$ and $R^2$ were used to assess the performance of the CR and the four ML in the combination applied on the dataset: M1, M2.

| Model | M3 | M4 |
|-------|----|----|
| LGBM  | 0.955315 | 0.954999 |
| GBR   | 0.97508 | 0.974731 |
| XGB   | 0.977963 | 0.977646 |
| Ridge | 0.977957 | 0.97764 |
| CR    | 0.977957 | 0.97761 |

Table 18.2. Represent the Adj $R^2$ and $R^2$ were used to assess the performance of the CR and the four ML in the combination applied on the dataset: M3, M4.

| Model | M5 |
|-------|----|
| LGBM  | 0.933201 |
| GBR   | 0.97351 | 0.97324 |
| XGB   | 0.97951 | 0.97929 |
| Ridge | 0.9795 | 0.97929 |
| CR    | 0.9795 | 0.97926 |

Table 18.3. Represent the Adj $R^2$ and $R^2$ were used to assess the performance of the CR and the four ML in the combination applied on the dataset: M5.
Figure 8. Scatter diagrams were used to assess the performance, actual, and prediction yield for SGD and SR on the dataset: (a) M1, (b) M2, (c) M3, (d) M4, and (e) M5.

Figure 8 displays a scatter diagram for locating the estimated and actual wild blueberry by SGD and SR for the M1, M2, M3, M4, and M5 datasets. The x-axis represents the actual yield (kg/ha), and the y-axis represents the prediction yield (kg/ha), as shown in Figure 8. The SR and SGD techniques were applied to the wild blueberry datasets. In addition, each colored point represents the MLA prediction. The two-colored points blue, and orange, represent the predictions of SGD and SR, respectively, as shown in Figure 8.

The highest $R^2$ was achieved by SR for M1, M2, M3, M4, and M5 datasets, indicating that the points of the output of SR are closer to linear, which is almost perfect, based on the scatter diagram. The highest error from the SGD model is also the worst for the M1, M2, M3, M4, and M5 datasets.
### Table 19.1: Performance Evaluation for 17 MLA, CR and SR.

| MLA  | M1        | M2 (Forward) | M2 (Forward) |
|------|-----------|--------------|--------------|
|      | R²        | RMSE         | MAE          | R²        | RMSE         | MAE          | RRMSE       | RRMSE       |
| SGD  | 0.054     | 1381.73      | 1058.873     | 0.203     | 1267.884     | 1076.379     | 17.59%      | 17.59%      |
| KR   | 0.698     | 779.593      | 619.408      | 0.558     | 944.015      | 743.566      | 13.09%      | 13.09%      |
| SVR  | 0.744     | 718.652      | 515.084      | 0.748     | 712.723      | 510.975      | 9.88%       | 9.88%       |
| LR   | 0.876     | 499.051      | 405.415      | 0.875     | 501.002      | 410.311      | 6.95%       | 6.95%       |
| EN   | 0.876     | 499.46       | 405.17       | 0.867     | 517.893      | 420.26       | 7.18%       | 7.18%       |
| BR   | 0.876     | 498.972      | 405.281      | 0.875     | 501.282      | 410.197      | 6.95%       | 6.95%       |
| Lasso| 0.835     | 576.034      | 470.995      | 0.815     | 610.866      | 499.685      | 8.47%       | 8.47%       |
| Ridge| 0.819     | 603.685      | 503.779      | 0.804     | 628.987      | 525.865      | 8.72%       | 8.72%       |
| MLPR | 0.882     | 487.628      | 403.705      | 0.587     | 912.737      | 726.68       | 12.66%      | 12.66%      |
| AdaBoost | 0.897 | 455.521      | 365.235      | 0.898     | 453.378      | 365.303      | 6.29%       | 6.29%       |
| DTR  | 0.89      | 489.972      | 405.281      | 0.875     | 501.282      | 410.197      | 6.95%       | 6.95%       |
| KNN  | 0.905     | 435.967      | 336.22       | 0.906     | 435.452      | 335.788      | 6.04%       | 6.04%       |
| LGBM | 0.95      | 317.555      | 172.63       | 0.95      | 317.555      | 172.63       | 4.40%       | 4.40%       |
| RF   | 0.953     | 305.482      | 203.327      | 0.953     | 305.102      | 202.917      | 4.23%       | 4.23%       |
| ETR  | 0.955     | 298.937      | 205.714      | 0.958     | 289.258      | 198.403      | 4.01%       | 4.01%       |
| GBR  | 0.978     | 210.339      | 154.657      | 0.978     | 208.59       | 153.839      | 2.89%       | 2.89%       |
| XGBoost | 0.978 | 208.542      | 162.296      | 0.975     | 147.47       | 147.47       | 3.05%       | 3.05%       |
| CR   | 0.974     | 227.88       | 169.674      | 0.977     | 212.675      | 163.458      | 2.95%       | 2.95%       |
| SR   | **0.985** | **173.068**  | **133.854**  | **0.984** | **177.746**  | **134.594**  | **2.46%**   | **2.46%**   |
| MLA      | M3 (BACKWARD) | M4 (VIF) |          |          |          |          |          |          |
|----------|---------------|----------|----------|----------|----------|----------|----------|----------|
|          | R²  | RMSE     | MAE      | RRMSE    | R²  | RMSE     | MAE      | RRMSE    |
| SGD      | -1.69 | 2330.476 | 1881.774 | 32.33%   | 0.206 | 1265.23  | 1041.074 | 17.55%   |
| KR       | 0.55  | 952.953  | 751.321  | 13.22%   | 0.55  | 952.505  | 750.877  | 13.21%   |
| SVR      | 0.748 | 712.707  | 510.96   | 9.88%    | 0.75  | 709.932  | 508.956  | 9.85%    |
| LR       | 0.875 | 501.002  | 410.311  | 6.95%    | 0.799 | 635.693  | 517.593  | 8.81%    |
| EN       | 0.808 | 622.504  | 507.562  | 8.63%    | 0.8   | 634.767  | 517.476  | 8.80%    |
| BR       | 0.875 | 501.394  | 410.273  | 6.95%    | 0.8   | 634.026  | 517.433  | 8.79%    |
| Lasso    | 0.804 | 628.963  | 516.545  | 8.72%    | 0.802 | 630.846  | 517.504  | 8.75%    |
| Ridge    | 0.799 | 636.581  | 532.365  | 8.83%    | 0.797 | 638.030  | 533.798  | 8.86%    |
| MLPR     | 0.868 | 515.436  | 426.553  | 7.15%    | 0.891 | 466.941  | 363.849  | 6.47%    |
| AdaBoost | 0.895 | 459.953  | 367.036  | 6.38%    | 0.899 | 450.668  | 364.386  | 6.25%    |
| DTR      | 0.905 | 436.162  | 303.259  | 6.05%    | 0.878 | 496.051  | 308.02   | 6.88%    |
| KNN      | 0.905 | 436.015  | 336.845  | 6.04%    | 0.912 | 420.653  | 325.378  | 5.83%    |
| LGBM     | 0.95  | 317.555  | 172.63   | 4.40%    | 0.947 | 324.994  | 181.609  | 4.50%    |
| RF       | 0.954 | 304.418  | 202.631  | 4.22%    | 0.953 | 306.057  | 203.653  | 4.24%    |
| ETR      | 0.958 | 288.778  | 198.163  | 4.00%    | 0.959 | 284.739  | 194.094  | 3.95%    |
| GBR      | 0.978 | 208.016  | 153.734  | 2.88%    | 0.978 | 209.665  | 153.957  | 2.90%    |
| XGB      | 0.976 | 217.12   | 145.519  | 3.01     | 0.975 | 223.75   | 147.683  | 3.10%    |
| CR       | 0.979 | 210.939  | 161.878  | 2.92%    | 0.979 | 205.361  | 162.543  | 2.84%    |
| SR       | **0.985** | **174.551** | **133.372** | **2.42%** | **0.984** | **179.854** | **134.055** | **2.49%** |
### Table 19.3: Performance Evaluation for 17 MLA, CR and SR on M5.

| MLA | M5 (7Features) | R² | RMSE   | MAE  | RRMSE |
|-----|----------------|----|--------|------|-------|
| SGD |                | 0.159 | 1302.543 | 1134.828 | 18.07% |
| KR  |                | 0.553 | 949.132 | 744.917 | 13.16% |
| SVR |                | 0.756 | 700.408 | 501.942 | 9.71% |
| LR  |                | 0.794 | 643.336 | 537.542 | 8.92% |
| EN  |                | 0.795 | 643.034 | 537.432 | 8.92% |
| BR  |                | 0.795 | 642.068 | 537.028 | 8.90% |
| Lasso |            | 0.796 | 641.49 | 537.078 | 8.90% |
| Ridge |           | 0.797 | 639.237 | 536.745 | 8.86% |
| MLPR |            | 0.799 | 636.319 | 535.936 | 8.82% |
| AdaBoost |       | 0.895 | 458.906 | 369.454 | 6.36% |
| DTR  |                | 0.909 | 428.357 | 311.824 | 5.94% |
| KNN  |                | 0.912 | 420.74 | 325.69 | 5.83% |
| LGBM |                | 0.953 | 307.354 | 168.769 | 4.26% |
| RF   |                | 0.953 | 305.564 | 202.978 | 4.23% |
| ETR  |                | 0.964 | 269.283 | 189.285 | 3.73% |
| GBR  |                | 0.978 | 209.057 | 153.255 | 2.90% |
| XGBoost |           | 0.981 | 190.788 | 145.686 | 2.64% |
| CR   |                | 0.979 | 203.389 | 161.291 | 2.82% |
| SR   | **0.984**      | **179.898** | **132.38** | **2.49%** |
VIII. DISCUSSION

This study examines how the stacking technique with the combination of different MLA and Bayesian optimization can overcome the limitations of modern MLA for agricultural yield prediction.

The accuracy is indicated in measures such as $R^2$, RMSE, and MAE, and we fulfilled our goal by attaining an $R^2$ close to 1. We also compared the stacking technique with the cascade approach and with 17 popular MLA. The stacking technique outperformed all other MLA and cascade approaches. We also compared our results to those of another study [17], and we discovered that the stacking strategy performed with maximum accuracy compared to that of another study.

The stacking technique exhibits a higher performance for different subsets of features. Several feature selection strategies have been employed to reduce the complexity of the datasets as well as the time required for computation. The main advantage of using Bayesian optimization is that it increases the accuracy and obtains the best hyperparameter for each MLA.

The stacking technique employs a combination of MLA (LGBM, GBR, and XGBoost) to represent level-0, obtains the output prediction, and feeds it as an input with the dataset to Ridge to represent level-1, yielding the final ridge result. The stacking technique has been applied to a variety of feature selections, with M1 representing all input feature sets. M2 used SFFS, which resulted in 10 features being selected. M3 used SBEFS, which resulted in selecting 10 features, and M4 used VIF, which resulted in 9 features being selected. M5 used the XFI to create seven essential features. For the M1, M2, M3, M4, and M5 datasets, the stacking strategy attained the highest accuracy.

The cascading approach used LGBM, obtaining the output prediction to the GBR. The output prediction was obtained, which was fed to XGBoost, the old prediction of LGBM was removed, the output prediction of XGBoost feeds to the ridge, and the output prediction of GBR was removed. For the M4 dataset, the cascading technique yielded the second-best model.

Most research has employed the current MLA to forecast agricultural yield [7] [11]; however, only a few studies have used ensemble ML techniques in different applications [42] to gain flexibility and improve accuracy. Researchers have focused on simulating experimental implementation, feature selection approaches, employing sensitivity analysis to determine the importance of a feature influencing crop output, and measuring the predictive model’s performance [17].

According to prior research, various factors influence wild blueberry productivity, including bee populations and species, plant diseases, weeds, soil type, clone genetics, climate change, and pesticides [2]. A statistical study revealed that yield is directly related to an increase in bee density in the field [17]. The challenge in the agriculture industry is insufficient data, and features needed for analysis is inaccessible. Small datasets are not suitable for deep learning methods. In our study, we focused on improving prediction accuracy by Bayesian optimization and then combining different MLA using the stacking and cascading techniques.

Compared with earlier research that used the same dataset and had the same goal, our proposed stacking technique using a combination of LGBM, GBR, XGBoost, and Ridge provided the highest accuracy, as shown in Table 20. A previous study [17] applied XGBoost to the seven features with an $R^2$ of 0.938. The SR achieved the highest $R^2$ of 0.984 for the same seven features. The improvement increased $R^2$ by 0.046 and decreased RMSE by 163.128 kg/ha, as the SR has more accurate predictions than XGBoost. Our proposed SR improved $R^2$ by 0.187 and decreased the RMSE by 459.339 kg/ha compared to Ridge.

![Table 20](https://example.com/table20.png)

| STUDY   | FST  | TECHNIQUES                  | NF | $R^2$ | RMSE  |
|---------|------|-----------------------------|----|-------|-------|
| Obie et al. [17] | XFI  | XGBoost (80%-20%)          | 7  | 0.938 | 343.026 |
| PM      | XFI  | Stacking technique (LGBM, GBR, XGBoost and Ridge) using Bayesian optimization (80%-20%) | 13 | 0.985 | 173.068 |
| PM      | VIF  |                            | 7  | 0.984 | 179.898 |
| PM      | SFFS |                            | 10 | 0.984 | 177.746 |
| PM      | SBEFS|                            | 10 | 0.985 | 174.551 |

SHAP was used to analyze the 13 features of the prediction; SHAP translated the inputs and output prediction to determine the most important features for prediction, and SHAP’s graphical representation aided in translating the contribution of each feature to the output target in scientific information about wild blueberry yield. SHAP is used in the modeling process to interpret significant features, achieving good performance with few features.

We investigated the stacking and cascading techniques using a novel combination of four MLA (LGBM, GBR, XGBoost, and Ridge) and 17 MLA for wild blueberry yield prediction. A simulation model of wild blueberry pollination spanning 30 years in Maine, USA, and the Canadian Maritimes yielded an input dataset of 777 records and 13 characteristics involving meteorological data, average blueberry CS, and four distinct kinds of bees and their densities [17]. The best hyperparameters for each MLA were determined using a Bayesian optimization technique. The stacking strategy outperformed the cascading and other techniques for all unique subsets of the dataset. Regarding the maximum RMSE value achieved on the M1, M2, M3, M4, and M5 datasets, the SGD was the worst. We investigated four unique feature selections playing a vital role in lowering the complexity of the dataset.
Following the theorem of Bayesian optimization of the dataset subjected to feature selection approaches, we concluded that using the SBEFS technique, the stacking technique achieved similar accuracy of 0.985 for all features. SHAP provides more information about features significantly influencing wild blueberry yield. An increase in OS, BB, and AD densities boosted wild blueberry yield. In general, our analysis found that the stacking strategy achieved the best performance in terms of the examined metrics, namely, $R^2$, RMSE, and MAE for the varied subset features.

IX. CONCLUSION
In this empirical study, we used the cascading and stacking strategies to create prediction models for wild blueberry yield, utilizing two novel MLA combinations. The characteristics features used in our empirical investigation were derived from a simulated model of wild blueberry pollination. The data were generated using a computer model and integrated with weather data collected 30 years ago in Maine, USA. To obtain more specific information, SHAP explores the aspects of impact prediction by XGBoost. SHAP concludes that the features MaxLTR, MinLTR, and AvLTD do not influence the wild blueberry yield. The stacking and cascading approaches have been used for various subsets of features. To reduce complexity, feature selection approaches such as SFFS, SBEFS, VIF, and XFI were used to select unique features. Bayesian optimization was used to improve accuracy. After tuning the hyperparameters of the four MLA (LGBM, GBR, XGBoost, and Ridge), the stacking and cascading techniques were used to combine them. The performance evaluation was performed to calculate $R^2$, RMSE, MAE, and RRMSSE for SR and CR, and many techniques were used for all features and different subsets of features. The comparison was performed using SR and another study. Results show that SR outperforms CR, 17 MLA, and the method in another study. SR increases the weak Ridge model up to 0.187 $R^2$ which is a significant progress on the seven features and increases it by 0.046 $R^2$ more than that in the previous study. In the future, we will apply EMLA with deep learning techniques to a large wild blueberry yield dataset. A large dataset could be collected from various sources and have new features absent in our study, such as soil type and weeds.

REFERENCES
[1] E. Asare, A. K. Hoshide, F. A. Drummond, G. K. Criner, and X. Chen, “Economic Risk of Bee Pollination in Maine Wild Blueberry, Vaccinium angustifolium,” J. Econ. Entomol., vol. 110, no. 5, pp. 1980–1992, 2017, doi: 10.1093/jee/tox191.

[2] J. Popp, K. Petö, and J. Nagy, “Pesticide productivity and food security. A review,” Agron. Sustain. Dev., vol. 33, no. 1, pp. 243–255, 2013, doi: 10.1007/s13593-012-0105-x.

[3] F. Drummond, “Reproductive biology of wild blueberry (Vaccinium angustifolium aiton),” Agric., vol. 9, no. 4, 2019, doi: 10.3390/agriculture9040069.

[4] R. Tasnim, F. Drummond, and Y. J. Zhang, “Climate change patterns of wild blueberry fields in downeast, maine over the past 40 years,” Water (Switzerland), vol. 13, no. 5, 2021, doi: 10.3390/w13050594.

[5] J. K. Tuell and R. Isaacs, “Weather during bloom affects pollination and yield of highbush blueberry,” J. Econ. Entomol., vol. 103, no. 3, pp. 557–562, 2010, doi: 10.1603/EC09387.

[6] F. A. Drummond, J. Lund, and B. Eitzer, “Honey bee health in maine wild blueberry production,” Insects, vol. 12, no. 6, 2021, doi: 10.3390/insects12060523.

[7] A. Crane-Droesch, “Machine learning methods for crop yield prediction and climate change impact assessment in agriculture,” Environ. Res. Lett., vol. 13, no. 11, 2018, doi: 10.1088/1748-9326/aae159.

[8] B. Ji, Y. Sun, S. Yang, and J. Wan, “Artificial neural networks for rice yield prediction in mountainous regions,” J. Agric. Sci., vol. 145, no. 3, pp. 249–261, 2007, doi: 10.1017/S002185960600691.

[9] K. Matsumura, C. F. Gaitan, K. Sugimoto, A. J. Cannon, and W. W. Hsieh, “Maize yield forecasting by linear regression and artificial neural networks in Illin, China,” J. Agric. Sci., vol. 153, no. 3, pp. 399–410, 2015, doi: 10.1017/S0021859614000392.

[10] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, “Wheat yield prediction using machine learning and advanced sensing techniques,” Comput. Electron. Agric., vol. 121, pp. 57–65, 2016, doi: 10.1016/j.compag.2015.11.018.

[11] S. Contjui and A. Groza, “Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning,” Expert Syst. Appl., vol. 64, pp. 269–286, 2016, doi: 10.1016/j.eswa.2016.07.037.

[12] M. Shahhosseini, G. Hu, and S. V. Archontoulis, “Forecasting Corn Yield With Machine Learning Ensembles,” Front. Plant Sci., vol. 11, no. July, pp. 1–16, 2020, doi: 10.3389/fpls.2020.01120.

[13] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamarudin, 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, doi: 10.1109/ACCESS.2021.3075159.

[14] M. K. Uşar, M. Nour, H. Sindi, and K. Polat, “The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets,” Math. Probil. Eng., vol. 2020, 2020, doi: 10.1155/2020/2836236.

[15] M. Meroni, F. Waldner, L. Seguini, H. Kerdivles, and F. Rembald, “Yield forecasting with machine learning and small data: What gains for grains?,” Agric. For. Meteorol., vol. 308–309, no. April, 2021, doi: 10.1016/j.agrformet.2021.108555.

[16] H. Qu and F. Drummond, “Simulation-based modeling of wild blueberry pollination,” Comput. Electron. Agric., vol. 144, no. November 2016, pp. 94–101, 2018, doi: 10.1016/j.compag.2017.11.003.

[17] E. Y. Obsie, H. Qu, and F. Drummond, “Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms,” Comput. Electron. Agric., vol. 178, no. September, p. 105778, 2020, doi: 10.1016/j.compag.2020.105778.

[18] A. Tolk, “The next generation of modeling & simulation: Integrating big data and deep learning,” Simul. Ser., vol. 47, no. 10, pp. 368–375, 2015.

[19] D. Chicco and G. Jurman, “An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis,” IEEE Access, vol. 9, pp. 24485–24498, 2021, doi: 10.1109/ACCESS.2021.3057196.

[20] M. Graczysz, T. Lasota, B. Trawiński, and K. Trawiński, “Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell.
Lect. Notes Bioinformatics), vol. 5991 LNAI, no. PART 2, pp. 340–350, 2010, doi: 10.1007/978-3-642-12101-2_35.

[21] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D. Q. Wei, “PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method,” Front. Microbiol., vol. 9, no. OCT, pp. 1–9, 2018, doi: 10.3389/fmicb.2018.02571.

[22] M. Jiang et al., “NeuroPred-Fuse: an interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods,” Brief. Bioinform., vol. 22, no. 6, 2021, doi: 10.1093/bib/bbab310.

[23] D. Wang and X. Yue, “The weighted multiple meta-models stacking method for regression problem,” Chinese Control Conf. CCC, vol. 2019-July, no. 1, pp. 7511–7516, 2019, doi: 10.23919/ChCCC.2019.8665869.

[24] T. E. Franz et al., “The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield,” F. Crop. Res., vol. 252, 2020, doi: 10.1016/j.fcr.2020.107788.

[25] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh, “Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India,” Int. J. Biotemelorol., vol. 62, no. 10, pp. 1809–1822, 2018, doi: 10.1007/s00484-018-1583-6.

[26] P. S. Maya Gopal and R. Bhargavi, “A novel approach for efficient crop yield prediction,” Comput. Electron. Agric., vol. 165, no. June, p. 104968, 2019, doi: 10.1016/j.compag.2019.104968.

[27] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, “Corn Yield Prediction With Ensemble CNN-DNN,” Front. Plant Sci., vol. 12, no. August, pp. 1–13, 2021, doi: 10.3389/fpls.2021.709008.

[28] E. H. Bouras et al., “Cereal yield forecasting with satellite drought-based indices, weather data and regional climate indices using machine learning in morocco,” Remote Sens., vol. 13, no. 16, 2021, doi: 10.3390/rs13163101.

[29] S. Sayago and M. Bocco, “Crop yield estimation using satelllite images: Comparison of linear and non-linear models,” AgricScientia, vol. 35, no. 1, pp. 1–9, 2018, doi: 10.31047/1668-298x.v1.n5.S20447.

[30] S. Khaki, H. Pham, and L. Wang, “Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning,” Sci. Rep., vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-89779-x.

[31] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, and S. V. Archontoulis, “Maize yield and nitrate loss prediction with machine learning algorithms,” Environ. Res. Lett., vol. 14, no. 12, 2019, doi: 10.1088/1748-9326/ab5268.

[32] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. December, pp. 4766–4775, 2017.

[33] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, “Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data,” Agric. Water Manag., vol. 225, no. February, p. 105758, 2019, doi: 10.1016/j.agwat.2019.105758.

[34] A. Haithm, A. Y. Saleh, and A. Odašt, “Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance,” J. Institute Sci. Technol., vol. 37, no. 1, pp. 157–168, 2021.

[35] T. Hu and T. Song, “Research on XGboost academic forecasting and analysis modelling,” J. Phys. Conf. Ser., vol. 1324, no. 1, 2019, doi: 10.1088/1742-6596/1324/1/012091.

[36] A. V. Dorugade and D. N. Kashid, “Alternative method for choosing ridge parameter for regression,” Appl. Math. Sci., vol. 4, no. 9–12, pp. 447–456, 2010.

[37] J. Wu, Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, “Hyperparameter optimization for machine learning models based on Bayesian optimization,” J. Electron. Sci. Technol., vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11899/JEST.1674-862X.89040120.

[38] M. Wistuba, N. Schilling, and L. Schmidt-Thieme, “Hyperparameter optimization machines,” Proc. – 3rd IEEE Int. Conf. Data Sci. Adv. Anal. DSA2016, pp. 41–50, 2016, doi: 10.1109/DSAA.2016.12.

[39] S. Deb, “A novel robust r-squared measure and its applications in linear regression,” Adv. Intell. Syst. Comput., vol. 532, no. October 2016, pp. 131–142, 2017, doi: 10.1007/978-3-319-48517-1_12.

[40] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Comput. Electron. Agric., vol. 177, no. August, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.

[41] H. Pham, “A new criterion for model selection,” Mathematics, vol. 7, no. 12, pp. 1–12, 2019, doi: 10.3390/MATH7121215.

[42] P. Pintelas and I. E. Livieris, “Special issue on ensemble learning and applications,” Algorithms, vol. 13, no. 6, 2020, doi: 10.3390/A13060140.

HAYAM R. SEIREG had worked as a teaching assistant at El-Shorouk Academy, has a master's degree from the Arab Academy for Science, Technology, and Maritime Transport, and is now pursuing a PhD. Her research is primarily concerned with artificial intelligence, namely multi-agent systems and machine learning.

Yasser M.K. Omar is an Associate Professor at the Department of Computer Science and Engineering's Computer Science College of Computing and Information Technology at AASTMT. His research focuses on artificial intelligence, namely agent technology, multi-agent systems, and machine learning. He has published several publications on a variety of subjects.

Fathi E. Abd El-Samie is a Professor at the Department of Electronics and Communication at Menoufia University's Faculty of Electronic Engineering. His current research interests include medical image processing and digital communication. He has written several articles in these areas.

Adel S. El-Fishawy is a tenured Associate Professor at Menoufia University's Faculty of Electronics Engineering, Department of Electronics and Communication Engineering, with research interests in image processing and remote sensing. He had several articles in a variety of subjects.
Ahmed Elmhalawy is an Associate Professor in the Department of Computer Science and Engineering at Menoufia University’s Faculty of Electronics Engineering. His research interests are in artificial intelligence, namely agent technology, multiagent systems, and machine learning. He had a lot of papers on various subjects.