Explainable AI Models for Blueberry Yield Prediction: A Step Towards Trustworthy Precision Agriculture

Niranjan Deokule and Geetanjali Kale

***Abstract*—Advancements in technology across the globe are leading to the adoption of Artificial Intelligence (AI) and Machine Learning (ML) as essential components in contemporary agricultural practices. However, the complex and opaque nature of many ML models poses challenges in interpreting predictions for end-users. To address this, Explainable AI (XAI) methods are leveraged to identify the key features influencing model predictions and their impact. This study evaluates eight ML models, with Gradient Boosting and CatBoost emerging as the top performers, achieving cross-validation scores of 133.20 and 134.78, respectively, and an R^2-score of 0.99. To understand the influence of individual features on yield predictions, Explainability methods such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) were utilized. The results emphasize the significant contribution of Explainable AI (XAI) in improving the clarity, understanding, and practical application of machine learning models within agriculture, thereby promoting increased confidence and wider acceptance of AI-based approaches in precision farming.**

***Index Terms*—Machine learning, Crop yield prediction, Artificial Intelligence, Explainable AI (XAI)**

I. Introduction

CROP yield forecasting plays a vital role in agricultural predictive analytics, providing essential insights to farmers, government bodies, and agricultural institutions. Accurate predictions aid stakeholders in anticipating crop yields for specific seasons, enabling informed decision-making regarding planting and harvesting schedules to optimize output. Furthermore, precise yield forecasts support improved pricing, profitability, and policy formulation. Accurately forecasting crop yields is a complex task, largely due to the intricate nature of agricultural systems and the impact of unpredictable variables like pest infestations, plant diseases, extreme weather events, soil properties, genetic traits of crops, and diverse farming practices. To overcome these issues, several machine learning (ML) methods such as regression analysis, decision tree algorithms, and artificial neural networks have been utilized. These methods utilize extensive crop yield datasets encompassing crop specifics, location, planting dates, soil characteristics, and weather conditions. By analyzing these datasets, ML algorithms establish relationships between inputs and outputs to generate accurate predictions. Despite their effectiveness, many ML models are inherently complex and opaque, lacking transparency in explaining their predictions, the underlying rationale, or the data patterns they identify. This lack of interpretability can undermine user confidence in their predictions, limiting the practical utility of ML models in agriculture. To overcome this challenge, Explainable Artificial Intelligence (XAI) has gained prominence as an effective approach to increase the clarity and interpretability of machine learning models. XAI methods enable the identification of key features influencing predictions, thereby enhancing user confidence and supporting well-informed decision-making. This study applied XAI methods to ML models for predicting wild blueberry yields, focusing on both model performance and interpretability. Specifically, we utilized SHAP and LIME to identify significant features and understand their influence on yield predictions.

Explainable Artificial Intelligence (XAI) is an emerging field dedicated to making machine learning models more transparent and understandable. It aims to improve users' trust and confidence in AI systems by offering clear insights into how decisions are made, all while preserving the models' accuracy and performance. In May 2017, the Defense Advanced Research Projects Agency (DARPA) launched a five-year initiative called the Explainable Artificial Intelligence (XAI) program, aiming to improve the transparency and interpretability of machine learning methods without compromising their accuracy [1]. This initiative aims to enhance transparency by offering clear interpretations of model predictions, ultimately building confidence in their dependability. XAI holds immense potential for agricultural applications by offering actionable insights into crop cultivation and growth. For instance, XAI enables early detection of issues, allowing farmers to take corrective measures promptly and mitigate risks. Additionally, it supports efficient resource management strategies for critical inputs such as water, fertilizers, and pesticides, ensuring their sustainable use.

Utilizing Explainable AI (XAI) techniques in predicting wild blueberry yields brings considerable advantages in terms of transparency and model understanding. This study evaluated eight different ML models to predict wild blueberry yields, leveraging a spatially explicit simulation computing model to construct the dataset [2]. The dataset comprises 17 features, including spatial plant characteristics, bee species composition, weather variations, and resulting yields. Pollination simulation models were employed to study pollen dispersal patterns, the scheduling of pollination events, and the effectiveness of different pollination strategies. These analyses provided valuable insights for enhancing crop yields

Manuscript received November 18, 2023; revised May 31, 2025.

This work was supported in part by the Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Maharashtra, India.

Niranjan Deokule is a Research Scholar at Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Maharashtra, India.(email: niranjan.deokule@gmail.com).

Dr. Geetanjali Kale is an Associate Professor and Head of Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Maharashtra, India. (email: gvkale@pict.edu)

and maximizing agricultural productivity. This research utilized both LIME and SHAP techniques to improve the predictive accuracy of machine learning models and identify the crucial factors influencing yield predictions. We incorporated Explainable AI (XAI) techniques to improve the clarity and transparency of machine learning model outputs, thereby fostering greater trust and understanding among end-users. This method enhanced insight into the elements affecting blueberry production while also aligning with the larger goal of promoting sustainable and efficient farming techniques.

The structure of this paper is as follows: Section II presents an in-depth overview of current approaches to agricultural yield prediction, emphasizing machine learning methods and the difficulties they face in modeling intricate, non-linear data relationships. Section III delves into the Dataset, Data Preprocessing, Feature Selection, and Model Training, offering a detailed explanation of the dataset utilized, preprocessing procedures, feature selection strategies, and the training workflow for different machine learning models.Section IV details the Results and Discussion, highlighting model performance evaluation using cross-validation metrics, scatter plot illustrations, and significant interpretations derived from the findings. Finally, Section V explores explainability methods, such as SHAP and LIME to assess feature importance, offering insights into the key factors that most influence yield prediction outcomes.

II. LITERATURE REVIEW

The growing importance of crop yield prediction has led to a comprehensive exploration of machine learning (ML) algorithms in agricultural science. Numerous studies have focused on various crops, each emphasizing different aspects of ML applications in agriculture, as shown in Table I.

M. Rashid et al. focused on predicting palm oil yield, providing a comprehensive review of the global status of palm oil yield prediction [3]. The study examined widely adopted features and prediction methods, offering a critical analysis of advanced machine learning approaches for crop yield forecasting. It emphasized their implementation in the palm oil sector and provided a comparative review with existing research in the field. Rashid et al. also examined the benefits and challenges of using ML for crop yield prediction, addressing both current and future challenges in agriculture. They proposed potential solutions to existing issues in crop yield prediction and explored future perspectives for ML-based palm oil yield prediction. They explored various domains including remote sensing, plant development and disease detection, mapping, tree enumeration, and the selection of suitable features and algorithms. Concluding their analysis of prior research, they proposed a potential architecture for machine learning-based prediction of palm oil yield.

Nigam et al. conducted a crop yield estimation study using data on rainfall, temperature, area, and season obtained from various Indian government departments [4]. They utilized various machine learning models, including Random Forest, XGBoost, K-Nearest Neighbors, and Logistic Regression, for their analysis. Additionally, they employed a hybrid model, Multiple Linear Regression - Artificial Neural Network (MLR-ANN), to predict paddy crop yield [5]. In this hybrid approach, Multiple Linear Regression was utilized to initialize the input weights and biases of the ANN's input layer, leading to enhanced prediction accuracy when compared to conventional machine learning techniques. In the East Godavari region of India, a study utilized Multiple Linear Regression and a density-based clustering technique to forecast rice yield [6]. The dataset consisted of eight features, including year, rainfall, sowing area, fertilizers, production, and yield. Multiple Linear Regression was employed to predict rice yield, highlighting the importance of incorporating various agricultural and environmental factors to enhance prediction accuracy.

Kamath et al. developed a comprehensive framework to support rapid agricultural production forecasting, consisting of four modules: Crop, Soil, Weather, and Predict [7]. This framework utilized the Random Forest technique and historical data from official government websites to identify significant crops. By integrating multiple data sources and leveraging advanced ML techniques, the framework demonstrated significant potential for accurate and timely agricultural forecasting. Abbas et al. focused on predicting future potato tuber yields using datasets from six fields in Atlantic Canada [8]. Four machine learning models—Linear Regression (LR), Elastic Net (EN), k-Nearest Neighbors (k-NN), and Support Vector Regression (SVR)—were utilized in the study. The input dataset comprised variables including horizontal and vertical soil electrical conductivity, moisture content of the soil, field slope, soil pH, soil organic matter (SOM), normalized difference vegetation index (NDVI), and the yield of potato tubers. Among these models, SVR exhibited superior performance across all datasets, underscoring the effectiveness of Support Vector Regression in agricultural yield prediction.

In another significant study, Andrew Crane-Droesch investigated the impact of climate change on agriculture by applying a semiparametric variation of a deep neural network to a corn yield dataset [9]. The dataset used in this research was sourced from the QuickStats database provided by the USDA's National Agricultural Statistics Service (NASS). It incorporated historical climate information from the METDATA meteorological dataset and projected weather patterns derived from the MACA (Multivariate Adaptive Climate Analogues) dataset. The study demonstrated the effectiveness of deep learning models in capturing intricate relationships between climatic factors and crop yield outcomes, offering valuable perspectives on the potential consequences of climate change for agricultural output. Pant et al. compiled a comprehensive dataset from the Food and Agriculture Organization of the United Nations (FAOSTAT) and the World Data Bank repository to forecast crop yields [10]. This dataset included attributes such as country, year, yield values, average rainfall, pesticide usage, and average temperature. They employed four machine learning models—Support Vector Machine (SVM), Gradient Boosting Regressor, Random Forest Regressor, and Decision Tree—to predict the yields of maize, potatoes, paddy, and wheat crops. Notably, the Decision Tree model achieved the highest accuracy of 96% for crop yield prediction, demonstrating the potential of decision trees in agricultural forecastingThe dataset comprised various attributes, including the country, year, crop yield figures, average rainfall, pesticide consumption, and mean temperature. To estimate the yields of crops such as maize, potatoes, paddy, and wheat, four machine learning algorithms

TABLE I
SURVEY OF CROP YIELD PREDICTION METHODS

Author	Objective	Dataset	Feature	Prediction Algorithm	Performance
Nigam et al. [4]	Crop Yield Prediction	Indian Government Dataset	Temperature, Rainfall, Area, Season	Random Forest Classifier, XGBoost Classifier, KNN Classifier and Logistic Regressor	Random Forest Classifier predicted the crop yield with better accuracy 67.80% based on rainfall, temperature, area and season.
Maya et al. [5]	Paddy Crop Prediction	Dataset collected from Statistical, Meteorological and Agricultural Departments of Tamilnadu, India	Area,maximum temperature, number of openwells, canals length and number of tanks	Multiple Linear Regression and Artificial Neural Network (MLR-ANN)	MLR-ANN has achieved better crop yield prediction accuracy.
D. Ramesh et al. [6]	Region specific crop yield analysis	Dataset collected from East Godavari district of Andhra Pradesh	Year, Rainfal, Area of Sowing, Yield, Fertilizers (Nitrogen, Phosphorous and Potassium) and Production	Multiple Linear Regression and Density-based clustering technique	The Density-based clustering technique showed superior performance in predicting rice yield compared to Multiple Linear Regression.
Pallavi Kamath et al. [7]	Crop Yield Prediction	Dataset collected from Official Government Websites	District, crop, soil type and area	Random Forest Algorithm	The Random Forest model attained a 98% accuracy rate in forecasting the most suitable crop for future cultivation.
Farhat Abbas et al., [8]	Potato tuber yield prediction	The datasets PE-2018, NB-2017, and NB-2018 were collected from three agricultural fields in Prince Edward Island and three fields in New Brunswick, Canada	Volumetric moisture content, soil electrical conductivity, slope, NDVI, along with HCP and PRP parameters	Linear Regression (LR), Elastic Net (EN), K-Nearest Neighbor (K-NN), and Support Vector Regression (SVR)	The SVR models outperformed all others, achieving the lowest RMSE values of 5.97, 4.62, 6.60, and 6.17 t/ha, respectively.
Andrew Crane-Droesch [9]	Corn yield prediction	Dataset collected from NASS, METDATA and MACA	Precipitation, Relative humidity, Wind speed, Air temperature, Shortwave radiation,Total precipitation, Latitude/longitude, Growing degree-days, Time, Soil, Proportion irrigated, County	Parametric and semi-parametric deep neural network	The use of bagging significantly enhanced the accuracy of both the parametric model and the SNN. However, the bagged SNN demonstrated the best performance.
Pant et al. [10]	Crop Yield Prediction-Maize, Potatoes, Rice (Paddy) and wheat	FAO data and world data bank repository	Country, year, crop yield, mean precipitation, pesticide usage, and mean temperature	Support Vector Machine (SVM), Gradient Boosting Regressor, Random Forest Regressor, and Decision Tree	The decision tree achieved highest accuracy 96% to predict crop yield.
Khaki et al. [11]	Corn and soybean yield prediction	Yield performance, management, weather, and soil Dataset	Features describing details about Weather , Soil and Management Practices	A hybrid CNN-RNN model, Random Forest (RF), deep fully connected neural network (DFNN), and the least absolute shrinkage and selection operator (LASSO).	Hybrid CNN-RNN Model outperfomred other models with RMSE 9% and 8% for corn and soybean yield prediction.
Morales et al. [12]	Sunflower and wheat crop yield prediction	Synthetic dataset collected from DSSAT 4.8	Sowing date, N applied , irrigation applied, anthesis date, rainfall, mean maximum and minimum temperature, mean solar radiation , average soil depth , year and cultivar	Linear model, Random forest and ANN	The Random Forest algorithm outperformed the other models with a Root Mean Square Error of 35-38%.
S. Hazra et al.[13]	Crop yield prediction	Dataset collected from US field data collection	Crop type, soil type, pesticide type and usage, season and crop damage	XGBoost, MLP, LightGBM, SVM, ANN, RF and K-NN	XGBoost Classifier achieved 84.79% average accuracy better than other models.
S Iniyan et al. [14]	Crop yield prediction	District-wise data collected from the Indian agricultural website	Precipitation, humidity, temperature, area, soil type, crop type, season	Multiple LR, DT, Gradient Boosting, Elastic Net, Lasso, Ridge, LSTM	LSTM has achieved best accuracy 86.3% and outperformed other models.

were utilized: Support Vector Machine (SVM), Gradient Boosting Regressor, Random Forest Regressor, and Decision Tree. Among these, the Decision Tree model delivered the best performance, attaining an accuracy of 96% in crop yield prediction. This highlights the effectiveness of decision trees in agricultural yield forecasting.

A hybrid Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) model based on deep learning was utilized to predict corn and soybean yields [11]. The dataset comprised details related to crop yield outcomes, agricultural management techniques, climatic variables, and soil properties. The hybrid model's predictive capability was evaluated against a deep fully connected neural network (DFNN) and the Least Absolute Shrinkage and Selection Operator (LASSO). The hybrid CNN-RNN approach demonstrated superior accuracy, yielding a reduced Root Mean Square Error (RMSE) of 9% for predicting corn yield and 8% for soybean yield. This highlights the model's effectiveness

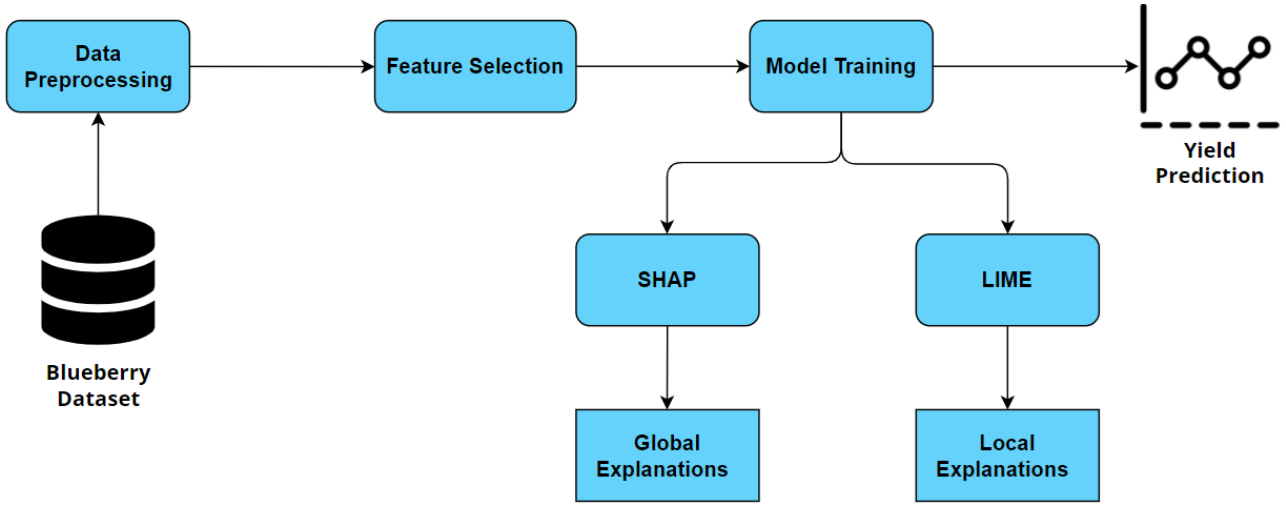

Fig. 1. System Architecture

in leveraging the strengths of CNN and RNN to interpret intricate patterns within agricultural datasets. Morales et al. employed a synthetic dataset from the Decision Support System for Agrotechnology Transfer (DSSAT) to predict sunflower and wheat crop yields [12]. This dataset comprised 11 features related to weather, soil, and management practices. Multiple modeling approaches such as linear regression, Random Forest, and artificial neural networks (ANN) have been utilized to forecast crop yields. The Random Forest algorithm demonstrated superior performance, with an RMSE ranging from 35% to 38%, highlighting its effectiveness in handling diverse agricultural datasets. S. Hazra et al. conducted a comprehensive study on crop yield prediction using seven machine learning models, including XGBoost, SVM, ANN, LightGBM, Multi-Layer Perceptron (MLP), k-nearest neighbor (KNN), and Random Forest (RF) [13]. They utilized field observation data from the United States, which included details on crop types, soil types, pesticide usage, seasons, and crop damage. The XGBoost Classifier achieved an average accuracy of 84.79%, outperforming other models. In addition, the study evaluated five Convolutional Neural Network (CNN) architectures—VGG-19, VGG-16, Inception v3, ResNet-50, and EfficientNet-B0—for predicting crop types. Among these, VGG-19 demonstrated the highest accuracy in recognizing crop images.

S. Iniyan conducted a crop yield prediction study using a dataset from an Indian agricultural website, organized by district [14]. The dataset comprised variables like precipitation, humidity, temperature, geographical area, soil classification, crop variety, and seasonal data. The research investigated eight machine learning techniques, such as Multiple Linear Regression, Decision Trees, Gradient Boosting, Elastic Net, Lasso, Ridge, and Long Short-Term Memory (LSTM) networks. Among these, the LSTM model delivered the best performance with an accuracy of 86.3%, highlighting the effectiveness of recurrent neural networks in modeling temporal patterns in agricultural datasets.

Elavarasan et al. introduced a deep reinforcement learning model, Deep Recurrent Q-Network (DRQN), to predict the yield of paddy crops in the Vellore district of southern India, achieving an accuracy of 93.7% [15]. The dataset used in this study included detailed information on climate, soil, groundwater properties, and fertilizer consumption specific to the study area. Important factors including evapotranspiration, frequency of ground frost, nutrient levels in groundwater, occurrence of wet days, and properties of aquifers were taken into account. The research highlights the capability of reinforcement learning methods in managing intricate agricultural datasets and enhancing the precision of crop yield predictions. H. S. Midtiby presented an innovative method for estimating pumpkin yield using drone-acquired images [16]. The methodology included creating an orthomosaic, extracting color models from a randomly selected portion of the dataset, applying color-based segmentation, and identifying pumpkin clusters to estimate the count of pumpkins within each cluster. This method achieved high precision and recall, with the lowest score being 0.959. The use of drone technology combined with image processing techniques illustrates the potential of remote sensing and computer vision in agricultural yield estimation.

Collectively, these studies highlight the promising role of machine learning in improving the accuracy of crop yield predictions. They demonstrate the successful application of a range of ML models across different agricultural settings, reflecting the ongoing advancement of techniques and the adoption of novel strategies like hybrid models. The insights gained from this body of work form a strong basis for continued exploration and innovation in ML-driven agricultural forecasting, including the use of Explainable AI (XAI) to boost model interpretability and stakeholder confidence.

III. Methodology

The system architecture in the figure 1 depicts a process for predicting blueberry yield using machine learning techniques. The wild blueberry dataset underwent a preprocessing stage to address noisy and missing values. This step ensured the dataset's quality and prepared it for effective modeling, thereby mitigating potential adverse impacts on predictive accuracy. A feature selection process based on the Pearson correlation coefficient was employed to identify features most relevant to the target variable, wild blueberry yield. The approach emphasized features that exhibited the highest correlation with the yield variable, thereby refining the dataset for machine learning applications. Subsequently,

TABLE II
DESCRIPTION OF WILD BLUEBERRY DATASET

Feature	Unit	Description	Min	Max	Median	Std Deviation
clonesize	m^2	Average blueberry clone size	10.00	40.00	18.76	6.99
honeybee	$bees/m^2/min$	Density of Honeybee in the field	0.00	18.43	0.41	0.97
bumbles	$bees/m^2/min$	Density of Bumblebee in the field	0.00	0.59	0.28	0.06
andrena	$bees/m^2/min$	Density of Andrena bee in the field	0.00	0.75	0.46	0.16
osmia	$bees/m^2/min$	Density of Osmia bee density in the field	0.00	0.75	0.56	0.16
MaxOfUpperTRange	Fahrenheit	Highest record of the upper band daily air temperature during the bloom season	69.70	94.60	82.27	9.19
MinOfUpperTRange	Fahrenheit	Lowest record of the upper band daily air temperature	39.00	57.20	49.70	5.59
AverageOfUpperTRange	Fahrenheit	Average of the upper band daily air temperature	58.20	79.00	68.72	7.67
MaxOfLowerTRange	Fahrenheit	Highest record of the lower band daily air temperature	50.20	68.20	59.30	6.64
MinOfLowerTRange	Fahrenheit	Lowest record of the lower band daily air temperature	24.30	33.00	28.69	3.20
AverageOfLowerTRange	Fahrenheit	Average of the lower band daily air temperature	41.20	55.90	48.61	5.41
RainingDays	Inches	Total number of days during the bloom season	1.00	34.00	18.30	12.12
AverageRainingDays	Inches	Day The average of raining days of the entire bloom season	0.06	0.56	0.32	0.17
fruitset	Time	Transitioning time of fruit set	0.19	0.65	0.50	0.07
fruitmass	Weight	Mass of the fruit set	0.31	0.54	0.44	0.04
seeds	Number	Number of seeds in fruitset	22.08	46.59	36.12	4.37
yield	kg/ha	Yield of blue berry	1945.53	8969.40	6012.84	1356.95

the refined dataset was divided into 80% for training and 20% for testing purposes. This partitioning was essential for model development and evaluation, ensuring the trained models were adequately validated on unseen data. To forecast wild blueberry yield, eight distinct machine learning models were implemented. This ensemble approach improved both the precision and resilience of the predictive system. To interpret the outcomes and gain insight into the decision-making process, two explainable AI (XAI) techniques were applied. LIME (Local Interpretable Model-Agnostic Explanations) provided localized interpretations for individual predictions, offering a clearer view of specific model decisions. In contrast, SHAP (Shapley Additive Explanations) offered a global perspective by highlighting the most influential features across the entire dataset. The final output of the system included the predicted yield along with comprehensive insights that merged both local and global explanation frameworks.

A. Dataset

Blueberries, part of the Vaccinium genus in the Cyanococcus section, are perennial plants known for producing blue or purple berries. This genus also encompasses cranberries, bilberries, huckleberries, and Madeira blueberries. North America is home to both wild (lowbush) and cultivated (highbush) blueberries, which range in height from 10 cm (4 inches) to 4 m (13 feet). The dataset used for predictive modeling, known as the Wild Blueberry Pollination Simulation Model, was created using an open-source, spatially explicit computer simulation program [2]. The dataset captures a range of influential variables, including plant layout, pollination patterns, diversity of bee species, and environmental conditions such as weather. These factors play a significant role—both independently and in combination—in shaping pollination effectiveness and the productivity of wild blueberry agroecosystems. It includes six categories for clone size, seven for honeybee density, ten for bumblebee density, and twelve categories each for Andrena and Osmia bee densities. Moreover, air temperature and precipitation are represented across three distinct levels. These variables were systematically varied to generate unique conditions for simulations. The dataset comprises 777 records, each containing 17 attributes associated with plant spatial patterns, pollination mechanisms (including outcrossing and self-pollination), bee species diversity, climatic variables, and wild blueberry yield, as detailed in Table II. The simulation model has been substantiated through three decades of field studies and experimental data from Maine, USA, and the Canadian Maritimes, establishing its utility for testing hypotheses and advancing theoretical research in wild blueberry pollination.

- Pollinator Density: The density of pollinators such as honeybees, bumbles, and other bees varies significantly across the field, with honeybee density ranging from 0 to 18.43 bees/m²/min, while bumblebee density shows a narrower range (0 to 0.59 bees/m²/min).
- Temperature Variations: During the bloom season, the daily upper band of air temperature varies between 69.70°F and 94.60°F, with a median temperature of 82.27°F. In contrast, the lower band temperatures fall within the range of 24.30°F to 33.00°F.
- Rainfall Patterns: The count of rainy days throughout the bloom season varies significantly, ranging between 1 and 34 days. On average, there are approximately 13.80 rainy days, with a standard deviation of 12.12 days.
- Fruitset and Yield: Fruitset, indicating the transition time for fruit set, ranges from 0.19 to 0.65, while the mass of the fruit set (fruitmass) has a median value of 0.54. The blueberry yield ranges significantly, from

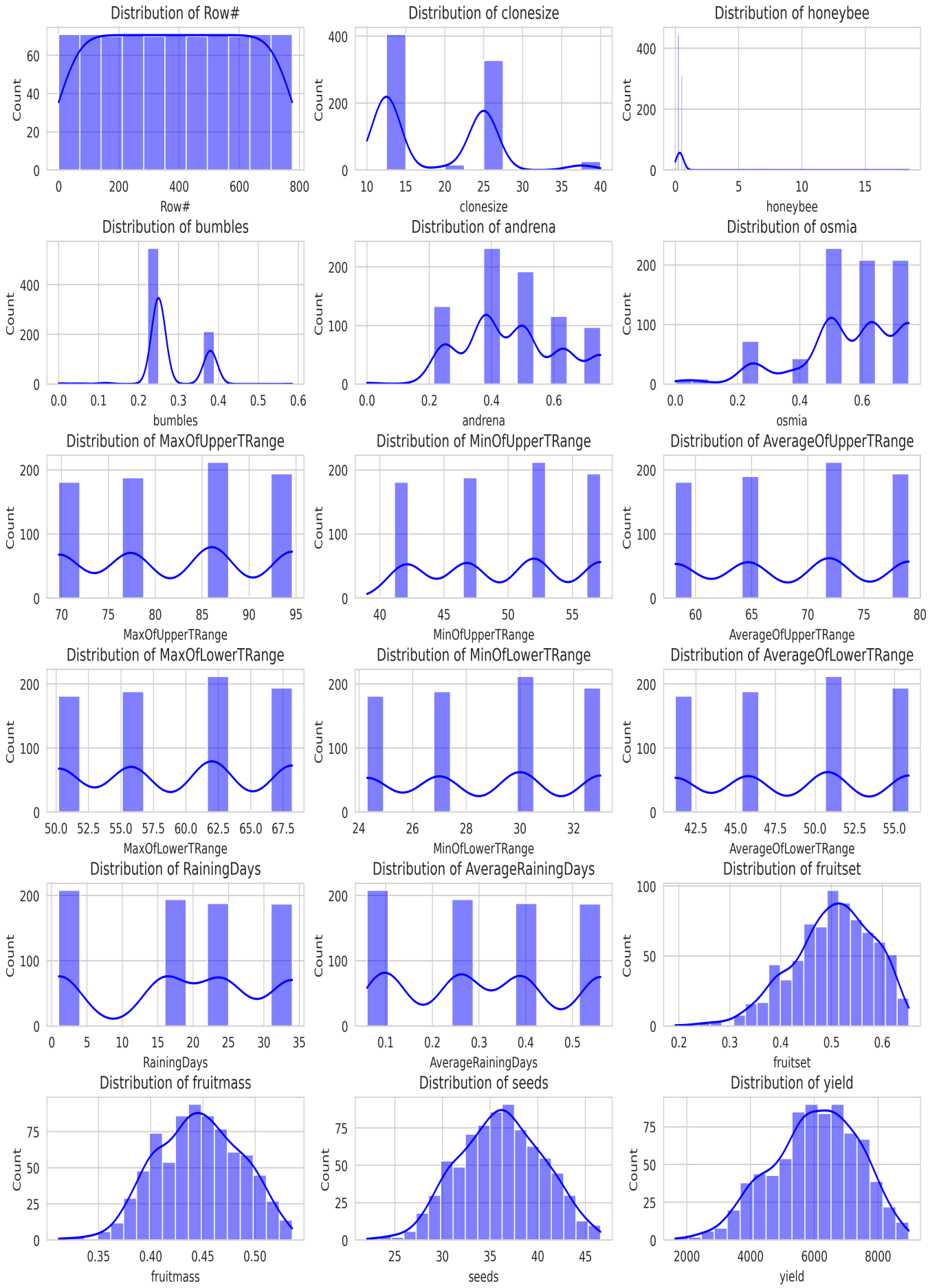

Fig. 2. Feature Distribution of wild blueberry dataset

1945.53 kg/ha to 8969.40 kg/ha, with a median yield of 6012.84 kg/ha.
- Clone Size: The size of individual blueberry clones typically varies between 10.00 m² and 40.00 m². The median clone size is approximately 18.76 m², with a standard deviation of 6.99 m².

B. Data Prepossessing

Real-world datasets often exhibit issues such as incompleteness, inconsistency, and inaccuracy, which can significantly impact the effectiveness of machine learning models. These problems can lead to models learning faulty patterns, ultimately affecting their accuracy and performance. Training machine learning models with such noisy data is likely to yield suboptimal results, as the models may struggle to identify meaningful patterns. The distribution of 16 features, along with the target variable 'yield,' is shown in Figure 2. The temperature and precipitation columns exhibit four distinct levels, corresponding to the conditions: Warm and Dry, Warm and Wet, Cool and Dry, and Cool and Wet. The use of these discrete numerical variables in various combinations within the simulation model is illustrated in the plots.

Data preprocessing constitutes a critical preliminary phase in the machine learning workflow, aimed at enhancing data quality, consistency, and suitability for subsequent analytical and modeling tasks. For instance, the wild blueberry dataset contained null values and outliers, causing an imbalance in its features. To mitigate this, a series of preprocessing steps was implemented:

1) Handling Missing Values: Missing entries within the dataset were detected and addressed using suitable imputation strategies. For columns containing numerical data, the median of each respective column was used to replace the missing values.

2) Outlier Detection and Removal: To prevent distortion in model performance caused by outliers, statistical techniques such as Z-score analysis and the Interquartile Range (IQR) method were employed to identify and eliminate anomalous data points from the dataset.

3) Min-Max Scaling: A Min-Max scaling technique was applied to specific columns within the dataset. This scaling method is particularly useful when the features have varying ranges, as it helps to standardize their impact on the model. By normalizing the values, this scaling process contributes to a more balanced and reliable dataset.

C. Feature Selection

Feature selection was crucial in pinpointing the most significant attributes influencing the target variable, yield. By narrowing down the dataset to only the most relevant features, this step effectively reduced dimensionality and improved both the performance and clarity of the predictive model. The steps undertaken for feature selection are detailed below:

1) Correlation Analysis: A color-coded correlation heatmap (Figure 3) was created to visually illustrate the strength and direction of relationships among the variables in the dataset. This analysis helped identify significant correlations among the attributes and the target variable, yield.

2) Feature Selection Based on Correlation:
- Fruitset, Fruitmass, and Seeds: The heatmap revealed significant correlations between the attributes fruitset, fruitmass, seeds, and yield. To address multicollinearity, only the variable most strongly correlated with yield—fruitset—was included in the final set of features (Figure 4).
- RainingDays and AverageRainingDays: Similarly, due to a strong correlation between RainingDays and AverageRainingDays, only RainingDays was included in the final feature set.

3) Refinement of Feature Set:
- Temperature-Related Attributes: Attributes such as AverageOfUpperTRange, MinOfLowerTRange, AverageOfLowerTRange, and AverageRainingDays exhibited high correlations with MaxOfUpperTRange, MinOfUpperTRange, and MaxOfLowerTRange. Therefore, the redundant attributes were removed.
- Final Feature Set: After refining the feature set based on correlation analysis, the final subset comprised 10 attributes: clonesize, honeybee, bumbles, andrena, osmia, MaxOfUpperTRange, MinOfUpperTRange, MaxOfLowerTRange, RainingDays, and fruitset.

D. Model Training

1) Data Partitioning: Before initiating the training process for the machine learning models, the wild blueberry dataset was split into two distinct subsets: 80% was allocated for training purposes, while the remaining 20% was set aside for testing. This approach allowed the models to learn from a major share of the data, ensuring a fair evaluation using the unseen portion.

2) Machine Learning Models: A total of eight machine learning models were employed to predict wild blueberry yields. The selected models included a combination of linear and non-linear algorithms to capture diverse patterns within the data.
- Linear Regression (LR): A linear regression model was applied to identify and quantify the relationship between the yield (target variable) and the ten chosen features from the wild blueberry dataset.
- Decision Tree: A decision tree model was utilized to effectively capture and manage non-linear interactions between input features and the target variable.
- Random Forest: To enhance predictive accuracy and ensure greater robustness, the Random Forest technique—an ensemble-based learning approach—was employed by aggregating the outputs of several decision trees.
- AdaBoost: AdaBoost was utilized as an ensemble technique that integrates several weak classifiers, often decision trees, to form a more accurate and robust predictive model. It iteratively adjusts weights for misclassified samples to improve overall accuracy.
- Gradient Boosting Regressor: The model was employed to construct an ensemble of individually weak predictive models, commonly decision trees, in order to enhance overall accuracy.
- LightGBM (LGBM): LightGBM, a gradient boosting framework known for its speed and scalability, was

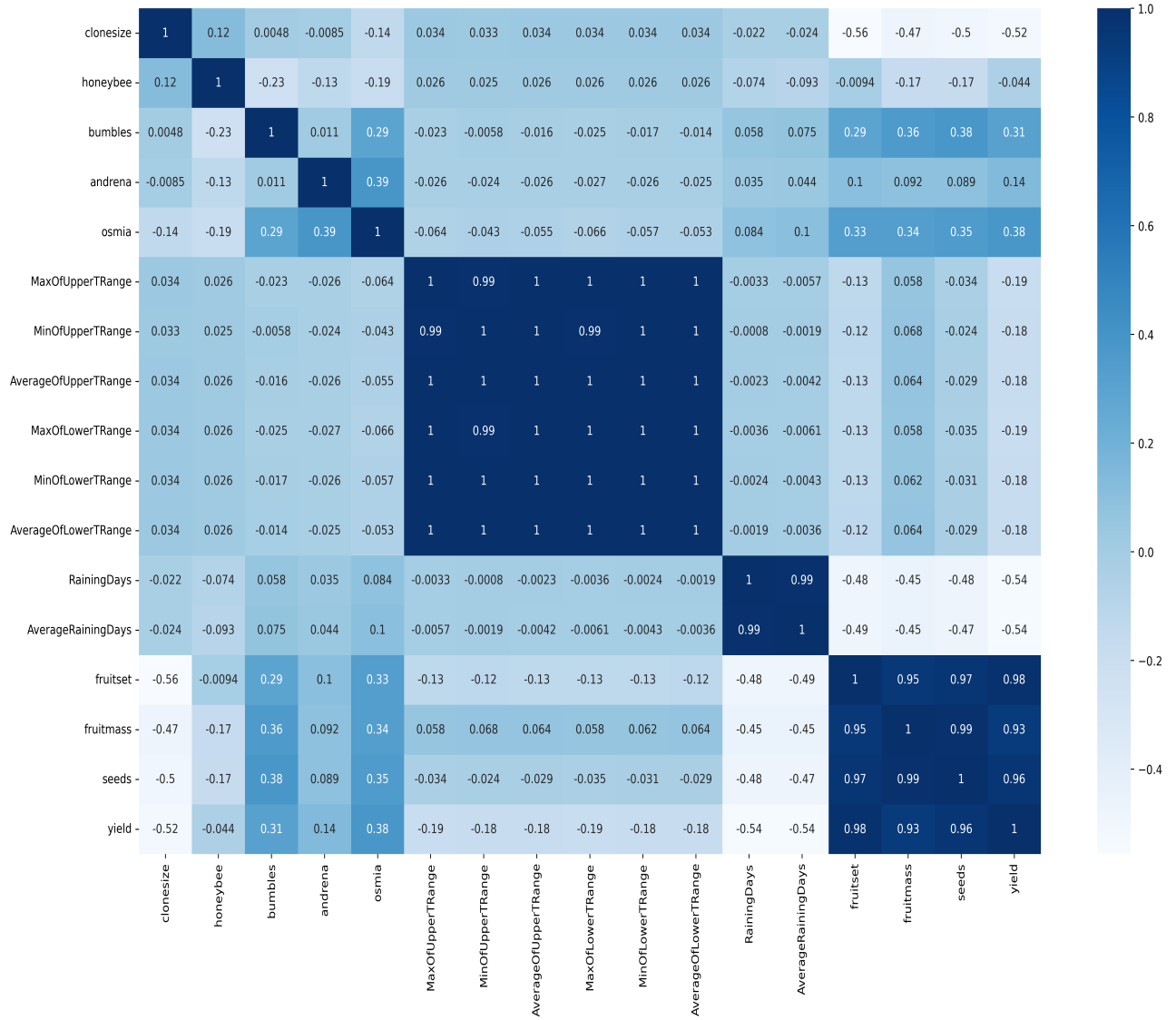

Fig. 3. Correlation heatmap

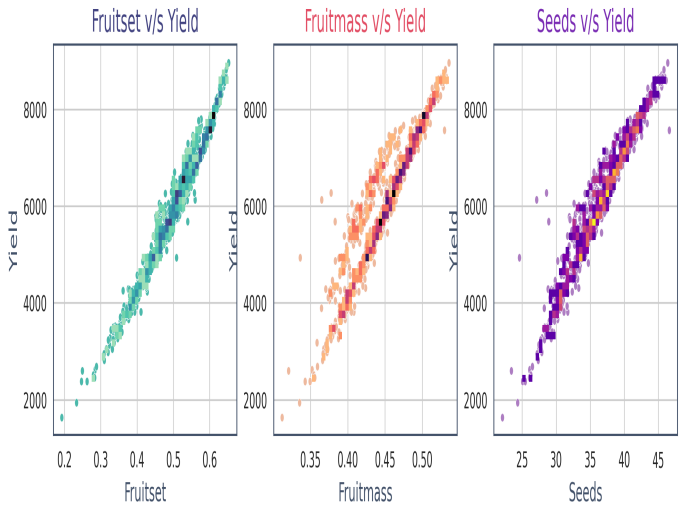

Fig. 4. Feature Selection Based on Correlation for Fruitset, Fruitmass, and Seeds

employed to efficiently process large datasets and deliver strong predictive performance. Its capability to handle categorical variables and model complex data relationships made it an ideal option for the task.

- XGBoost: XGBoost, a powerful gradient boosting algorithm optimized for speed and performance, was used for its regularization techniques and ability to prevent overfitting. It effectively leverages parallel processing to enhance predictive accuracy.
- CatBoost: CatBoost, a gradient boosting technique known for effectively processing categorical variables, was selected due to its computational efficiency and strong predictive capabilities.

E. Results and Discussion

Several machine learning regression models were analyzed to predict wild blueberry yield. The models assessed in this study include Linear Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, LightGBM, XGBoost, and CatBoost. Among these, Linear Regression was applied to identify a linear association between the yield (target

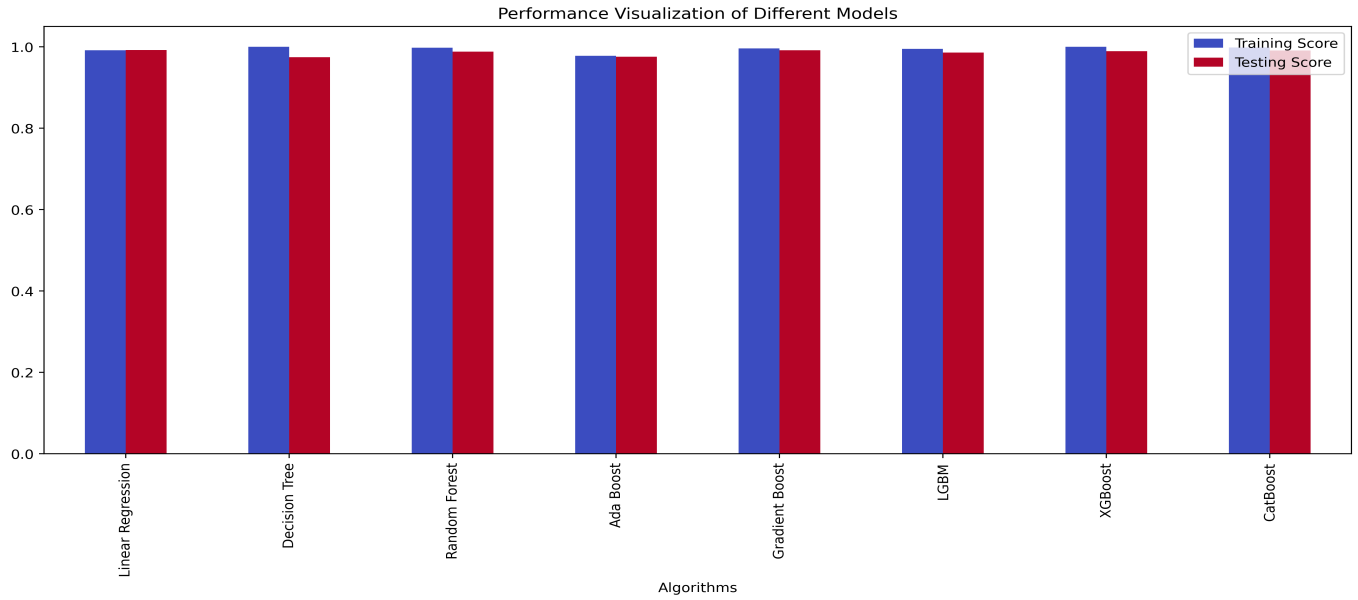

Fig. 5. Performance visualization of different machine learning models

variable) and the ten chosen features from the wild blueberry dataset.

TABLE III
CROSS VALIDATION SCORES

Model	Cross Validation
Linear Regression	159.93
Decision Tree	188.35
Random Forest	155.76
Ada Boost	216.74
Gradient Boost	133.20
LGBM	156.81
XGBoost	145.60
CatBoost	134.78

Additionally, tree-based machine learning algorithms like decision trees and random forests were utilized to predict wild blueberry yield. Boosting algorithms, known for their exceptional performance and higher accuracy compared to other machine learning techniques, played a crucial role in this analysis. Boosting algorithms are highly effective because they learn from the mistakes of earlier models, enabling them to reduce the likelihood of repeating the same errors made by previous weak learners.

1) Model Performance: The cross-validation scores (Table III) indicated varying levels of prediction accuracy across the models, with CatBoost achieving the lowest cross-validation error (134.78), closely followed by Gradient Boost (133.20). Decision Tree and AdaBoost exhibited higher error rates, highlighting their limitations with this dataset. The Gradient Boosting model demonstrated the best cross-validation performance, achieving a score of 133.20, which indicates its strong generalization capability compared to the other models. CatBoost closely followed Gradient Boost, with a cross-validation score of 134.78, making it one of the top-performing models in terms of predictive consistency. Random Forest and LGBM displayed similar performance in cross-validation, with scores of 155.76 and 156.81, respectively, which reflects their moderate accuracy across different folds. Linear Regression, while computationally efficient, had a cross-validation score of 159.93, indicating that it may not be as robust as more complex models like Gradient Boosting and CatBoost. AdaBoost recorded the highest cross-validation score of 216.74, reflecting weaker generalization performance compared to the other models.

TABLE IV
PERFORMANCE EVALUATION OF MACHINE LEARNING METHODS

Algorithm	MAE	MSE	RMSE	R2	Time (s)
Linear Regression	124.95	25578.09	159.93	0.98	0.0065
Decision Tree	140.67	35896.92	189.46	0.98	0.0114
Random Forest	118.61	24030.65	155.01	0.98	0.4152
Ada Boost	176.59	48816.98	220.94	0.97	0.1361
Gradient Boost	102.55	17990.46	134.12	0.99	0.1807
LGBM	114.66	24592.36	156.81	0.98	0.1229
XGBoost	110.37	21199.42	145.60	0.98	0.5732
CatBoost	98.02	18167.93	134.78	0.99	1.2029

The performance of eight distinct machine learning models was assessed based on various evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), and the time taken for computation in seconds, as presented in Table IV. Among these, Gradient Boosting—a robust ensemble learning technique—demonstrates improved predictive accuracy by sequentially training models. Each subsequent learner focuses on correcting the errors of its predecessor by minimizing a specific loss function, often defined using metrics like mean squared error or cross-entropy, through gradient descent optimization. In contrast, CatBoost, another gradient boosting method, demonstrated versatility in handling both numerical and categorical features. What distinguishes CatBoost is its ability to convert categorical features into numerical ones without requiring feature encoding techniques. It also utilizes the Symmetric Weighted Quantile Sketch (SWQS) technique to efficiently manage missing values in the dataset automatically. This dual strategy not only reduces the problem of overfitting but also improves the accuracy of wild blueberry yield predictions.

Among the models tested, CatBoost achieved the lowest

MAE of 98.02 and one of the highest R^2 scores of 0.99, demonstrating strong predictive accuracy, although it had the longest computation time of 1.2029 seconds. The Gradient Boosting model displayed a strong balance between accuracy and efficiency, with an MAE of 102.55, an R^2 score of 0.99, and a moderate computation time of 0.1807 seconds. While Linear Regression offered the fastest computation time (0.0065 seconds), its performance was less accurate, with an MAE of 124.95 and an R^2 score of 0.98. Random Forest provided relatively accurate results, with an MAE of 118.61 and an R^2 score of 0.98, but its computation time (0.4152 seconds) was longer than some of the other models. AdaBoost had the weakest performance among the models, with the highest MAE (176.59) and RMSE (220.94), despite having a relatively short computation time of 0.1361 seconds.

Figure 5 illustrates a comparative analysis of the performance of various machine learning models based on their training and testing scores. Each model's predictive performance was represented by blue bars (training scores) and red bars (testing scores), allowing a visual assessment of overfitting or underfitting. Most models, including Gradient Boost, LGBM, XGBoost, and CatBoost, displayed minimal differences between training and testing scores, indicating strong generalization capabilities. These models achieved the highest testing scores, reflecting their superior predictive accuracy. Linear Regression exhibited slightly lower scores compared to other advanced models, suggesting it might not have captured the dataset's complexity as effectively. While competitive, AdaBoost's testing score lagged slightly behind ensemble methods like Gradient Boost and Random Forest. This visualization emphasized the comparative effectiveness of boosting techniques, particularly Gradient Boost, LGBM, XGBoost, and CatBoost, in achieving both robust training and testing performances.

2) Scatter Plot Analysis: Figure 6 illustrated the correlation between the predicted outcomes and the actual values for each model. The near-linear alignment of the data points for Gradient Boost, XGBoost, and CatBoost signified their strong predictive performance. In contrast, Linear Regression and Decision Tree exhibited relatively higher dispersion, suggesting limited accuracy in capturing the complexity of the data. All eight machine learning models showed a strong correlation between actual and predicted yield values, as evidenced by the close clustering of data points along the diagonal line—indicating high predictive accuracy. Among them, the Gradient Boosting Regressor and CatBoost Regressor stood out, displaying minimal variance from the diagonal, which highlights their exceptional effectiveness in accurately predicting yield outcomes. The Linear Regression model exhibited a reasonably strong performance, with its predictions generally aligning well with the actual yield values. However, it showed slightly lower accuracy compared to ensemble models such as Gradient Boosting and Random Forest. In contrast, the AdaBoost Regressor produced more dispersed prediction points, suggesting a less accurate fit and higher variability in yield estimates. On the other hand, the Random Forest Regressor and XGBoost Regressor delivered superior predictive accuracy, with their outputs closely following the ideal diagonal line—indicating their effectiveness in modeling yield outcomes. The LGBM Regressor model also displayed strong predictive performance, though it showed slightly more spread in the middle range compared to Gradient Boosting and CatBoost models. The differences observed between the actual and predicted yields, as determined by the Gradient Boosting model, served as an indicator of the model's predictive performance. Smaller differences indicated better model performance, as shown in Figure 7. The findings demonstrate that advanced ensemble methods (e.g., CatBoost, Gradient Boost) are well-suited for yield prediction in agricultural datasets.

F. Explainable AI Methods

1) SHapley Additive exPlanation (SHAP): Shapley Additive Explanation (SHAP), a technique based on game theory, was initially introduced by Lundberg and Lee [17]. It was used to attribute SHAP values to the features that significantly contributed to a model's predictions. Seireg et al. employed SHAP to analyze the influence of features on the predictions made by XGBoost [18]. They utilized stacking and cascading methods to integrate the predictions of four distinct machine learning models—LGBM, GBR, XGBoost, and Ridge—following the fine-tuning of their respective hyperparameters. Additionally, unique feature selection methods, including SFFS, SBEFS, VIF, and XFI, were employed to reduce model complexity. The predictive accuracy of the models was further enhanced through Bayesian optimization, which involved 17 different machine learning algorithms, all focused on predicting wild blueberry yield. A hybrid deep learning method, LSTM-CNN, was also proposed for predicting corn and soybean yields [19]. The dataset included information on average yield, environmental variables, management practices, soil data, and MODIS data. The SHAP XAI tool was used to identify the most influential features contributing to yield predictions.

The figure 8 illustrated how features affected Gradient Boost predictions using SHAP explanations. On the Y-axis, features were ordered by their average absolute SHAP values, while the X-axis displayed the SHAP values. High feature values were depicted in red, while low values were shown in blue. The analysis clearly highlighted that the Gradient Boost model's predictions were most significantly influenced by fruitset, RainingDays, osmia, MaxOfLowerTRange, and MaxOfUpperTRange. The results showed that greater fruitset values led to higher blueberry yields. Conversely, decreased RainingDays values corresponded to lower blueberry yields, while increased osmia values resulted in higher yields of blueberries.

Figure 9 highlights the key features that significantly influenced the predictions of the CatBoost model, as identified through SHAP analysis. Among these, fruitset, RainingDays, osmia, bumbles, and clonesize emerged as the most impactful variables contributing to the model's performance. It is noteworthy that fruitset, RainingDays, and osmia demonstrated consistent predictive influence across both the Gradient Boost and CatBoost models. This insight was derived from the SHAP feature importance, as depicted in Figure 10.This analysis of feature importance offered valuable insights into the features with high absolute Shapley values, highlighting their critical contribution to the CatBoost model's predictive performance.

By aligning these findings with our overall understanding of the issue, we can have confidence that the model is

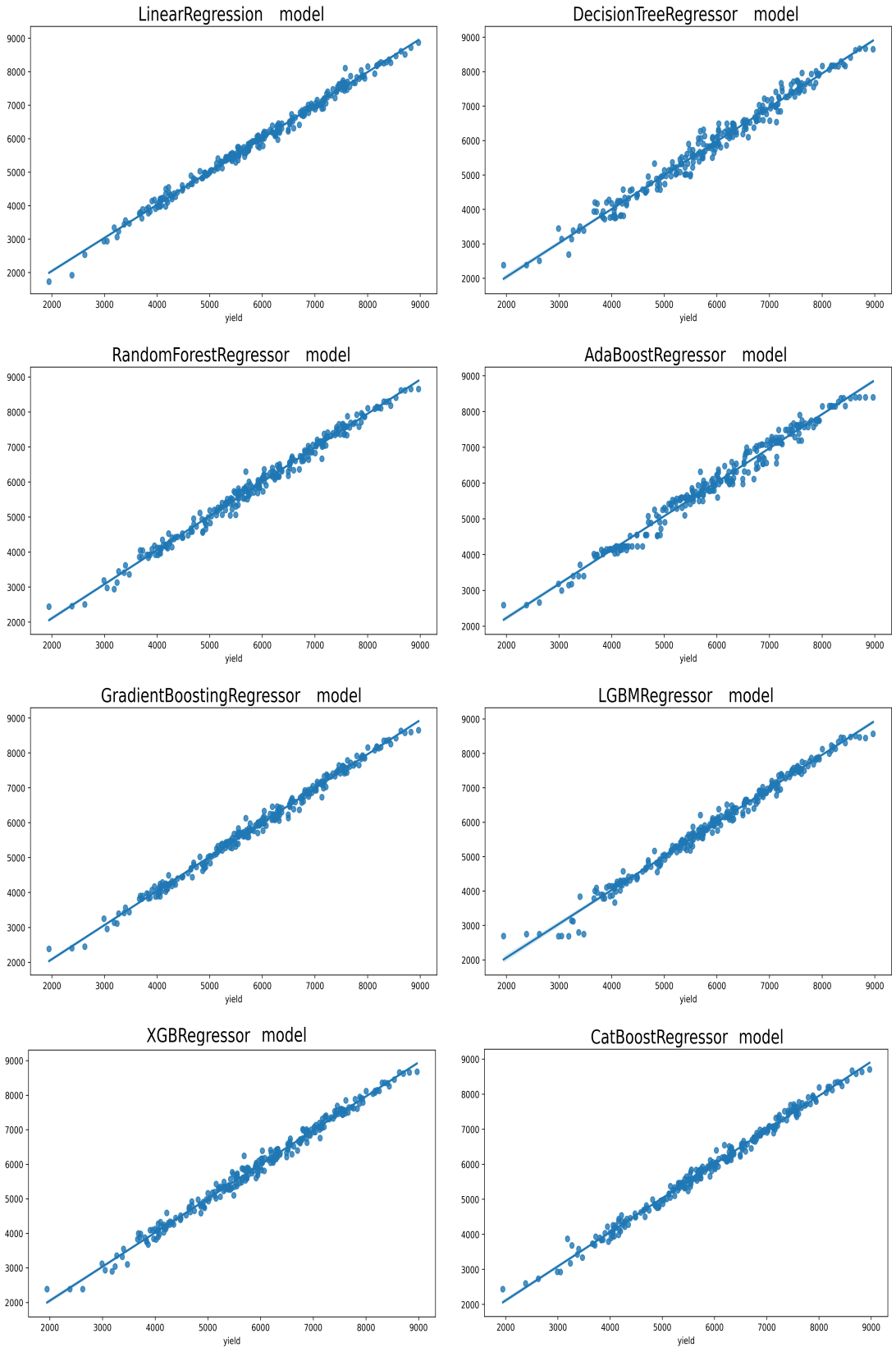

Fig. 6. Yield prediction of wild blueberry achieved by machine learning models

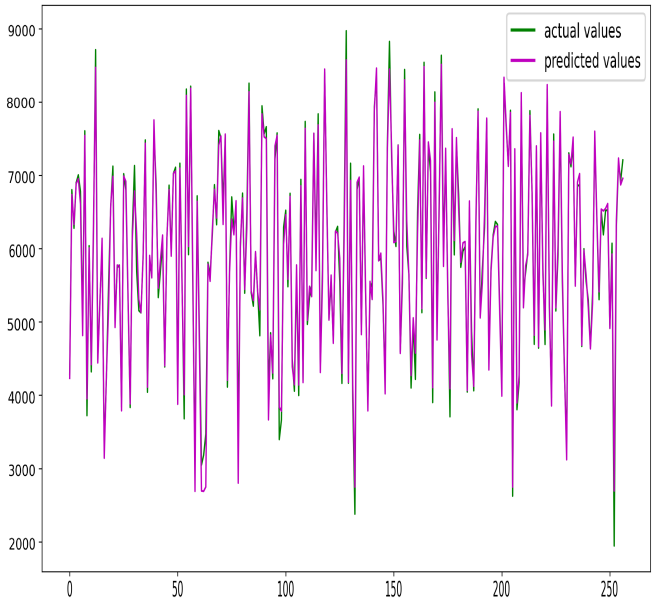

Fig. 7. Actual and predicted yield by Gradient Boost Model

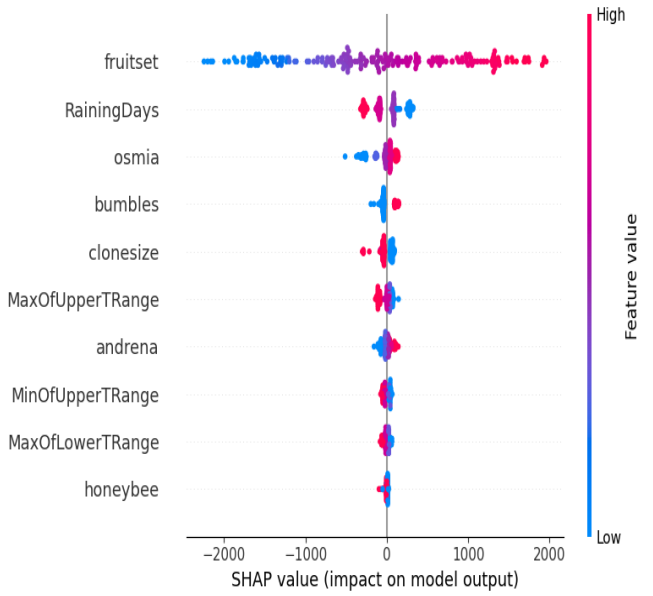

Fig. 9. Impact of features on CatBoost Model's Output by SHAP

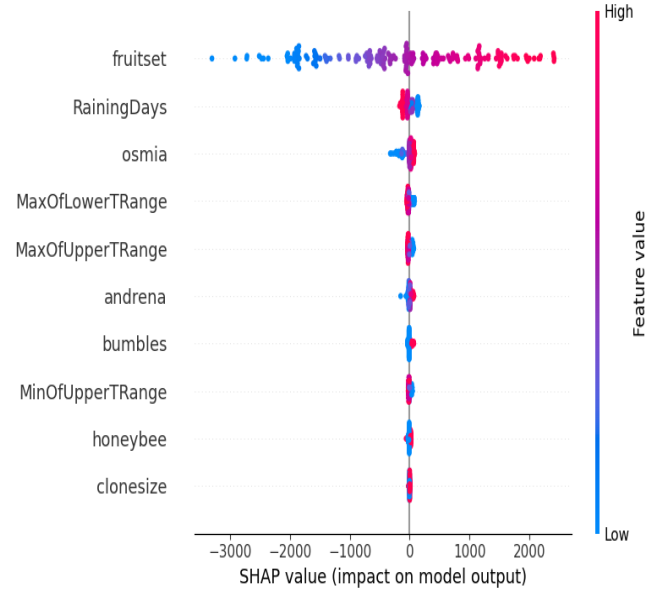

Fig. 8. Impact of features on Gradient Boost Model's Output by SHAP

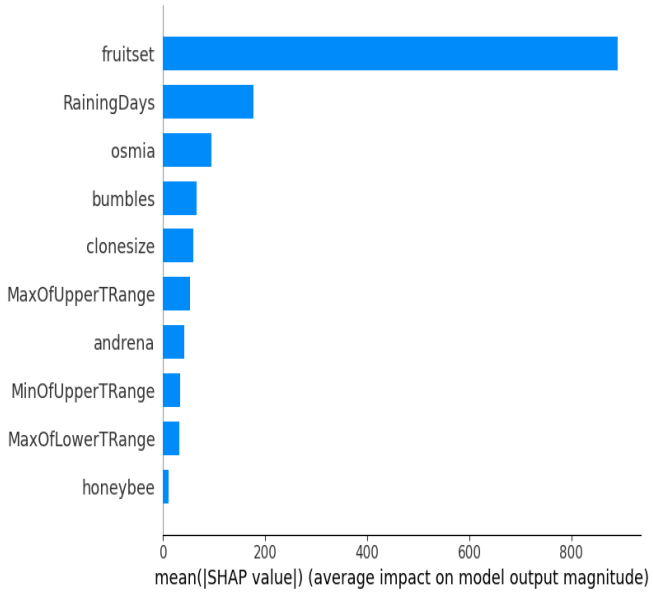

Fig. 10. Feature importance on CatBoost Model's Output by SHAP

intuitive and accurately predicting blueberry yields. It is probable that the model will predict higher blueberry yields for greater values of fruitset, RainingDays, and osmia.

2) Local Interpretable Model-Agnostic Explanations (LIME): LIME is a well-known approach within Explainable Artificial Intelligence (XAI) that gained recognition for its ability to interpret and explain the outputs of complex machine learning models. It was originally introduced by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin in 2016 [20]. This technique allowed for the assessment of the individual contributions of each variable within a dataset to each prediction generated by the model. LIME proved to be an effective approach for assessing the impact of individual variables on a machine learning model's predictions, as well as for comparing the relative importance of different variables in influencing the model's outcomes. Masahiro Ryo leveraged explainable artificial intelligence, incorporating techniques such as partial dependence plots (PDPs), pairwise interaction importance, permutation-based variable importance, and LIME, to investigate maize crop data [21]. The study employed a dataset containing 17 variables and applied predictive models such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting to estimate maize crop yield. LIME proved valuable in identifying the factors that wield the most significant or minimal influence on each prediction generated by a machine learning model. Additionally, it helped discern which variables hold the most or least sway over the outcome of each prediction in comparison to other variables. In practice, LIME generated an explanation for a particular instance within the test dataset, employing a Gradient Boost classifier. This explanation is then presented in a tabular format as shown in figure 11.In this table, the leftmost column illustrates the predicted probabilities for

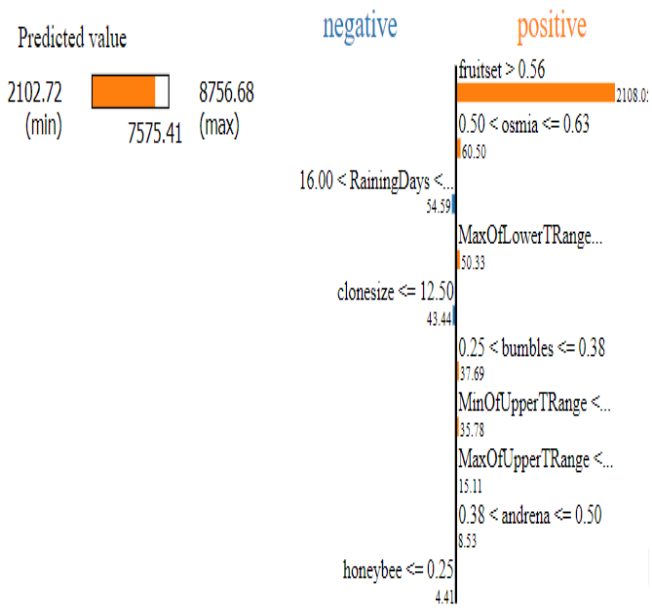

Fig. 11. Local Explanation by LIME for Gradient Boost

Feature	Value
fruitset	0.60
osmia	0.63
RainingDays	24.00
MaxOfLowerTRange	55.80
clonesize	12.50
bumbles	0.38
MinOfUpperTRange	46.80
MaxOfUpperTRange	77.40
andrena	0.50
honeybee	0.25

Fig. 12. Feature and their respective values for Gradient Boost

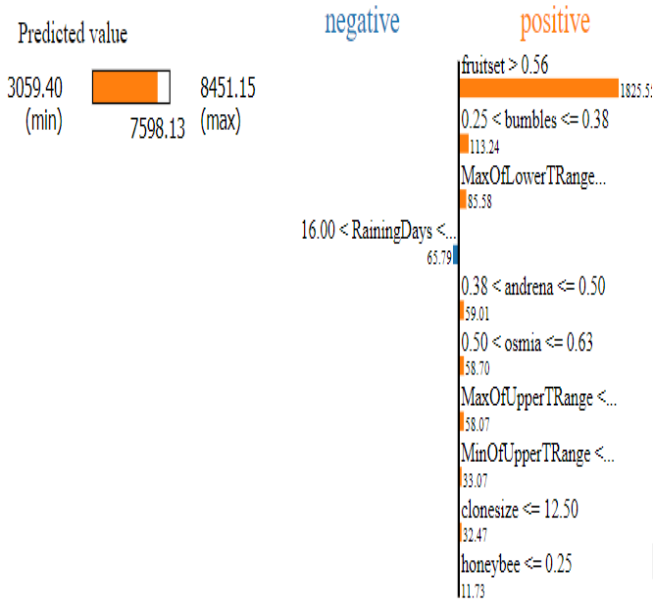

Fig. 13. Local Explanation by LIME for Cat Boost

wild blueberry yield. Features that had a positive impact on the model's predictions are highlighted in orange, whereas those that negatively influenced the outcome are shown in blue. For a particular instance from Blueberry dataset, the Gradient Boost has predicted the 7575.41 yield for which fruitset, osmia has positively contributed and RainingDays has negatively contributed towards the predicted yield. According to explanation generated by LIME as shown in figure 12, fruitset, osmia, RainingDays, MaxofLowerTRange and clonesize have contributed most in the Gradient Boost model's prediction of blueberry yield. The fruitset, bumbles, MaxofLowerTRange, RainingDays,

Feature	Value
fruitset	0.60
bumbles	0.38
MaxOfLowerTRange	55.80
RainingDays	24.00
andrena	0.50
osmia	0.63
MaxOfUpperTRange	77.40
MinOfUpperTRange	46.80
clonesize	12.50
honeybee	0.25

Fig. 14. Feature and their respective values for Cat Boost

and andrena have contributed most in the Cat Boost model's prediction as per the explanations generated for local instance by LIME as shown in figure 13. One of the case where Cat Boost has predicted the blueberry yield as 7598.13, fruitset, bumbles and MaxOfLowerTRange has positively contributed and RainingDays has negatively contributed towards the prediction. The feature and their respective values for Cat Boost are shown in figure 14. Higher the value of fruitset, higher the yield of blueberry and lower the value of RainingDays affected the blueberry yield. The patterns predicted for top features from wild

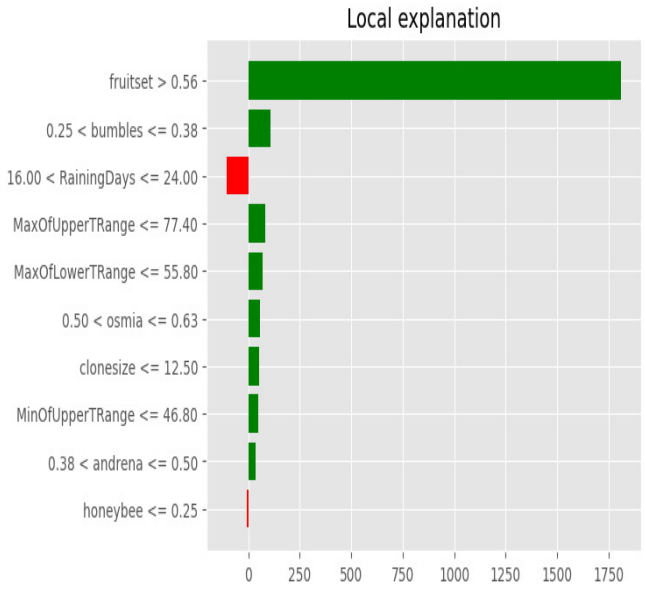

Fig. 15. Patterns detected by LIME for Cat Boost

blueberry dataset learned by Cat Boost model identified by LIME as shown in figure 15. The fruitset feature has the higher impact on model's prediction. The bumbles and MaxofLowerTRange have positive contribution in model output whereas the decrease in the feature value of RainingDays affected the yield of blueberry.

IV. Conclusion

The wild blueberry dataset, created through a spatially explicit simulation computing model, was examined to forecast blueberry yield using eight different machine learning models. Among these models, Gradient Boost and Catboost stood out as top performers, achieving cross-validation scores of 133.20 and 134.78, along with an impressive R^2-score of 0.99. To gain insights into model behavior, the study employed the Shapley Additive exPlanation (SHAP) method, which provides a global perspective, elucidating how the model behaves across all instances. To gain insight into individual predictions LIME was employed, providing localized explanations that help interpret the machine learning model's behavior for particular cases.

References

[1] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program", *AIMag* , vol. 40, no. 2, pp. 44-58, Jun. 2019.

[2] Obsie, Efrem, Qu, Hongchun, Drummond and Francis, "Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms", *Computers and Electronics in Agriculture*, 2020.

[3] Rashid, Md. Mamunur et al. "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction." IEEE Access 9 (2021): 63406-63439.

[4] Nigam, Aruvansh, Saksham Garg, Archit Agrawal and Parul Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms." *Fifth International Conference on Image Information Processing (ICIIP)*, 2019, 125-130.

[5] MayaGopalP., S. and R. Bhargavi, "A novel approach for efficient crop yield prediction." *Comput. Electron. Agric*, 2019.

[6] Ramesh, Dharavath and Vishnu Vardhan, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES", *International Journal of Research in Engineering and Technology*, 2015, 470-473.

[7] Pallavi Kamath, Pallavi Patil, Shrilatha S, Sushma, Sowmya S, "Crop yield forecasting using data mining", *Global Transitions Proceedings*, Volume 2, Issue 2, 2021, Pages 402-407, ISSN 2666-285X.

[8] F. Abbas, H. Afzaal, A. A. Farooque, and S. Tang, "Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms", *Agronomy,* vol. 10, no. 7, p. 1046, Jul. 2020, doi: 10.3390/agronomy10071046.

[9] Crane-Droesch, Andrew, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture", *Environmental Research Letters 13*, no. 11, 2018: 114003.

[10] Janmejay Pant, R.P. Pant, Manoj Kumar Singh, Devesh Pratap Singh, Himanshu Pant, "Analysis of agricultural crop yield prediction using statistical techniques of machine learning", *Materials Today: Proceedings* , Volume 46, Part 20,2021,Pages 10922-10926,ISSN 2214-7853.

[11] Khaki S, Wang L, Archontoulis SV, "A CNN-RNN Framework for Crop Yield Prediction", *Front Plant Sci.* , 2020.

[12] Morales A and Villalobos FJ, "Using machine learning for crop yield prediction in the past or the future", *Front Plant Sci*, 2023;14:1128388.

[13] Hazra Simanta, Karforma Sunil, Bandyopadhyay Abhishek, Chakraborty Sayantani and Chakraborty Debasis, "Prediction of Crop Yield Using Machine Learning Approaches for Agricultural Data", *techrxiv*, 2023.

[14] S Iniyan, V Akhil Varma, Ch Teja Naidu, "Crop yield prediction using machine learning techniques", *Advances in Engineering Software* ,Volume 175,2023,103326,ISSN 0965-9978.

[15] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications", in IEEE Access, vol. 8, pp. 86886-86901, 2020, doi: 10.1109/ACCESS.2020.2992480.

[16] Midtiby, Henrik Skov, and Elżbieta Pastucha. 2022. "Pumpkin Yield Estimation Using Images from a UAV" Agronomy 12, no. 4: 964. https://doi.org/10.3390/agronomy12040964

[17] Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions", *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, USA, 4768–4777, 2017.

[18] H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy and A. Elmahalawy, "Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," *IEEE Access*, vol. 10, pp. 64671-64687, 2022.

[19] Alexandros Oikonomidis, Cagatay Catal and Ayalew Kassahun (2022) "Hybrid Deep Learning-based Models for Crop Yield Prediction", Applied Artificial Intelligence, 36:1, DOI: 10.1080/08839514.2022.2031823

[20] Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin, "Model-Agnostic Interpretability of Machine Learning." *ArXiv,* abs/1606.05386, 2016.

[21] Masahiro Ryo, "Explainable artificial intelligence and interpretable machine learning for agricultural data analysis", *Artificial Intelligence in Agriculture*, Volume 6,2022,Pages 257-265,ISSN 2589-7217.