IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 October 2024, accepted 22 October 2024, date of publication 25 October 2024, date of current version 13 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3486653

RESEARCH ARTICLE

Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability

ABID BADSHAH[1], BASEM YOUSEF ALKAZEMI[2], (Senior Member, IEEE), FAKHRUD DIN[1], KAMAL Z. ZAMLI[3,4], (Member, IEEE), AND MUHAMMAD HARIS[4]

[1]Faculty of Information Technology (IT), Department of Computer Science and IT, University of Malakand, Dir Lower, Chakdara, Khyber Pakhtunkhwa 18800, Pakistan
[2]Department of Software Engineering, College of Computing, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[3]Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), Pekan, Kuantan, Pahang 26600, Malaysia
[4]Faculty of Science and Technology, Universitas Airlangga, C Campus JI. Dr. H. Soekamo, Mulyorejo, Surabaya 60115, Indonesia

Corresponding author: Kamal Z. Zamli (kamalz@ump.edu.my)

Problem statement

ABSTRACT Agriculture is pivotal for the economy of a country as it is a major source of food, employment and raw materials. However, challenges such as diseases, soil degradation, and water scarcity persist. Technology adoption can address these issues, improving production and quality. Machine learning, a subset of Artificial Intelligence (AI), enables prediction, classification, and automation in agriculture. It optimizes irrigation, fertilization, and crop selection, aiding decision-making for food security and crop management. This study proposes two robust machine learning architectures for classification and regression based on distinct datasets. Firstly, we delve into a crop recommendation dataset obtained from Kaggle, consisting of various input attributes such as the pH of the soil, temperature, humidity, and nutrient levels. Leveraging machine learning classification techniques such as Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM), we suggest twenty-two different crops founded on these inputs. Through the use of K-fold cross-validation, Explainable AI (XAI) and feature engineering, we identify the best-performing model, with Random Forest coming out on top scoring an accuracy of 99.7% with precision, recall, F1 score, and confusion matrix. Secondly, we investigate wheat yield prediction data snagged from the World Bank and Food and Agriculture Organization (FAO), covering the years 1992-2013 for Pakistan. Using Multivariate Imputation by Chained Equations (MICE) to tackle data restrictions, we gauge wheat production for 2014-2024 and forecast the 2025 yield using machine learning regression models. Once again, using hyper parameter tuning with K-fold cross-validation, Support Vector Regressor (SVR) stands out as the top-performing model, achieving an accuracy of 99.9% with R^2 Score. Transparency and confidence in agricultural decision-making are increased when machine learning decisions are made comprehensible using Explainable AI (XAI) approaches. Two widely used XAI approaches, namely Feature Importance and Local Interpretable Model-Agnostic Explanations (LIME) are used to interpret and explain outcomes of the proposed models. The study can increase agricultural productivity, minimize risks, enhance food security, and promote more environmentally friendly farming approaches.

Research Objectives

Methods

key results

Conclusion

INDEX TERMS Agricultural planning, crop recommendation, crop yield forecasting, explainable AI, K-fold cross-validation, machine learning.

Present your topic and get the reader interested

I. INTRODUCTION

Many Asian countries rely heavily on agriculture as the foundation of their economy and the main source of income

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

Provide background or summarize existing research

for a sizable portion of the populace. But with companies that rely on agriculture, there is a conspicuous lack of quality control, which lowers output and has negative effects on farmers [1]. Due to their failure to repay bank loans taken out for farming activities, many farmers experience financial troubles and, in extreme situations, are pushed to extreme actions like suicide [2], [3]. Farmers are further driven into debt and hopelessness by these problems, which are made worse by the environment's constant change. By offering data-driven insights and facilitating improved decision-making about crop selection and agricultural practices, statistical and mathematical tools help reduce these risks [2]. Utilizing current data for categorization and prediction analytics, machine learning, a branch of artificial intelligence (AI), presents potential solutions. In addition to predicting crop growth and diagnosing illnesses, this technology can automate vital agricultural procedures like fertilization and irrigation [4]. While obtaining precise and reliable results is still a problem, precision agriculture focuses on site-specific advice. Still, a great deal of research is being done to create more accurate and effective crop forecast models [5]. It is feasible to select the best crops, increasing yield, quality, and profitability while reducing environmental effect, by evaluating soil conditions and using machine learning.

Detail your specific research problem

Numerous obstacles still remain in the way of machine learning's potential in agriculture, including those related to infrastructure, cost, social acceptability, data accessibility, quality, and environmental effect [4]. Because agricultural ecosystems are diverse and dynamic, traditional approaches for crop recommendation and yield prediction sometimes fall short of meeting these needs. Advanced, trustworthy models that can offer precise suggestions and production projections are desperately needed, particularly in areas like Pakistan where agriculture is a major industry.

Position your own approach

The goal of this work is to create a crop recommendation and yield forecast system using machine learning, specifically for Pakistan. A crop recommendation system is developed using a Kaggle dataset [6]. Based on weather and soil characteristics, it attempts to assist in choosing the optimal crops for cultivation. It comprises parameters such as soil pH, temperature, humidity, rainfall, phosphorus (P), potassium (K), and nitrogen (N) in the soil. There are 2200 entries in the collection, each of which describes a set of conditions and the ideal crop to produce under those circumstances. The recommendation method matches these parameters with 22 distinct crop varieties using classification techniques. Despite coming from India, a neighboring nation, the data was collected from many areas with comparable soil and weather conditions, allowing for precise crop recommendations and extending the dataset's applicability to comparable agricultural sites. Based on a number of input parameters, this approach is meant to suggest the best crops as shown in Fig. 1. The correlation matrix is depicted in Fig. 2.

Additionally, the study forecasts the accuracy of future output for twenty-two different crops using supervised machine

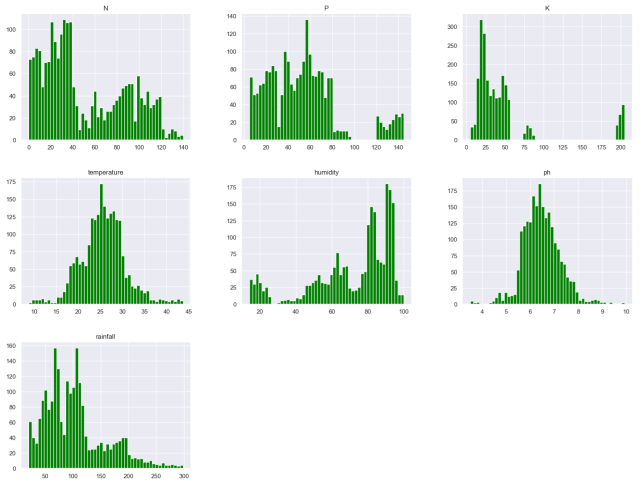

FIGURE 1. Block diagram outlining the histogram of different features.

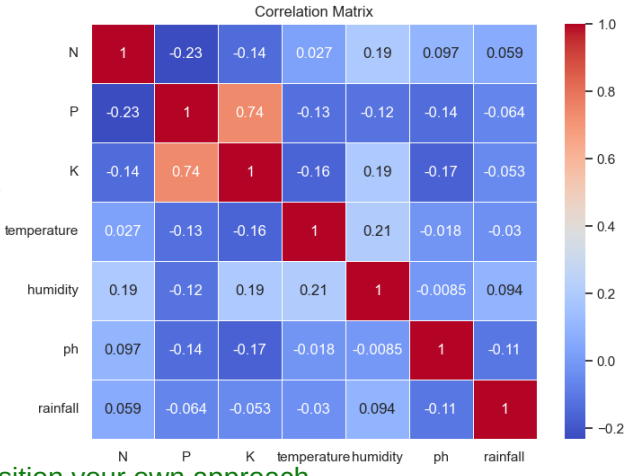

FIGURE 2. Block diagram outlining the correlation matrix.

learning methods. By extracting additional features from highly linked characteristics, the dataset will be improved. We used a variety of Explainable AI (XAI) approaches in our study to improve crop recommendation's interpretability and transparency. Local Interpretable Model-agnostic Explanations (LIME), and feature significance analysis are among the methods employed. The efficacy of distinct machine learning models will be evaluated through the application of measures, including recall, precision, F1 score, and confusion matrix.

Furthermore, our investigation entails yield prediction dataset procured from the World Bank and FAO [7], covering the years between 1992 and 2013, delving into wheat production in Pakistan. Fig. 3 and Fig. 4 show the features and labels pairplot and correlation matrix of the dataset, respectively.

Multivariate Imputation by Chained Equations (MICE) was used to impute data for 2014–2024 even though the dataset is limited to 2013. Predictions are made with accuracy through synthetic data that faithfully replicates

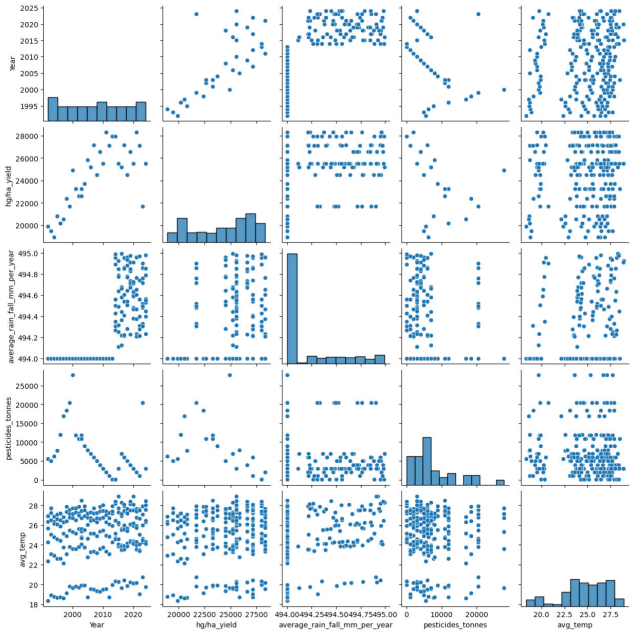

FIGURE 3. Block diagram outlining the pairplot of features and label.

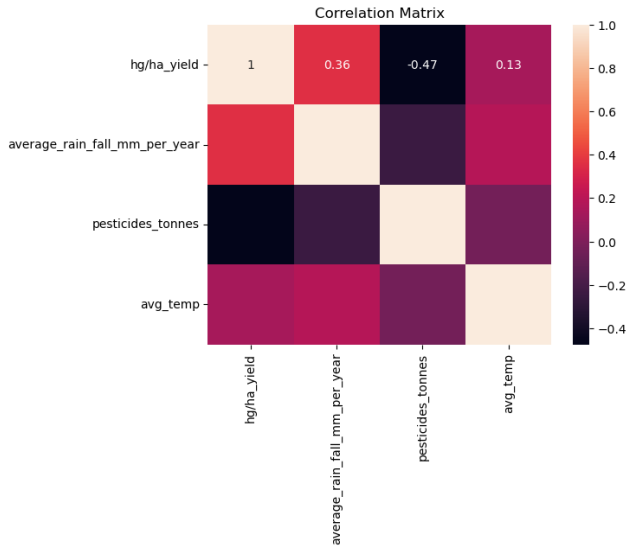

FIGURE 4. Block diagram outlining the correlation matrix of yield dataset.

real data patterns. Multiple complete datasets were created using MICE, allowing machine learning models to for reliably estimate wheat output for 2025. These models were then tested by K-fold cross-validation. This study's scope includes comprehensive data analysis, feature engineering, Explainable AI, MICE imputation technique, model training, and performance assessment to guarantee the accuracy and robustness of the yield projections and crop recommendations, which significantly advances the field of agricultural analytics.

Farmers may make more informed decisions by using the actionable insights that are provided by the identification of key factors impacting crop output. The study shows how machine learning algorithms might increase agricultural sustainability and production. Additionally, the study offers a workable paradigm for wise crop management and decision-making that can be modified and applied in a range of agricultural contexts. This work intends to advance sustainable agriculture by overcoming the shortcomings of current approaches and providing a reliable crop recommendation and yield prediction solution. By giving farmers in Pakistan and related areas specific, data-driven recommendations, it aims to enhance their standard of living. Because of its practicality and adaptability, the framework created in this study may be extensively used to assist intelligent agriculture practices. By employing this methodology, the research advances the robustness and effectiveness of farming systems, so bolstering food sovereignty and financial steadiness in rural areas.

In order to improve agricultural decision-making our study contributes by utilizing these two datasets. We maximize yields and optimize land usage by offering precise crop recommendations using robust machine learning models that take environmental aspects into account. We also provide farmers and officials with precise estimates through our yield prediction models which aid in resource planning. Farming becomes more sustainable and effective as a result of these efforts which also lower agricultural difficulties, improve the utilization of resources, and increase the supply of food.

Concretely, the study contributions are summarized as follows:

1) Thorough assessment of prediction models: Determine which robust machine learning classifiers (Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors Regressor (KNNR), Support Vector Regressor (SVR), and Extreme Gradient Boosting Regressor (XGBoostR)) are the most accurate models for crop recommendation and yield prediction by evaluating them using metrics like Precision, Recall, and F1 Score for classification and MSE, RMSE, Standard Deviation, and MAE for regression.
2) Integration of multiple data sources: Two distinct datasets [6], [7], for crop classification and yield prediction, are used to create robust and generic prediction models. For enhancing quality of the adopted datasets, efficient techniques such as feature scaling and multivariate imputation by chained equations (MICE) are employed. These techniques improved the predictive accuracy of the models.
3) Interpretation and applicability of the models: Through Feature Importance and Local Interpretable Model-Agnostic Explanations (LIME), outcomes of the models can easily be accessed and utilized by farmers and policy makers.

The aftermath of the paper goes as followed. Section II examines the various literature review related to this study; Section III presents the detailed description of the proposed

Give an overview of the paper's structure

framework called crops recommendation and yield prediction through Machine Learning approaches and then compression between them! Section IV introduces the experimental results and discussions, all verifying the model performance. Eventually, Section V wraps it up with the conclusion and future scope of the study.

II. RELATED WORKS

Many nations, especially in Asia, still rely heavily on agriculture for their economies and means of subsistence. Yet, obstacles like soil erosion, climate change, and financial strains have made it harder and harder for farmers to sustain profitability and production. By offering data-driven insights and suggestions, machine learning (ML) approaches present viable answers to these problems.

Crop recommendation systems, which examine a range of parameters such as soil properties, climate, and past crop data to suggest the best crops for certain areas, have made machine learning models increasingly important. Recent advancements in this field demonstrate the potential of many machine learning strategies. For instance, [8] showed the usefulness of more basic ML models in agricultural contexts by using linear regression models to forecast crop production based on climate and soil data. In order to forecast maize yield, [9] used Random Forest regression, highlighting the stability of ensemble approaches in handling intricate agricultural datasets. Support Vector Regression (SVR) was used by [10] to predict agricultural yields, demonstrating SVR's adaptability to non-linear correlations between variables. Reference [11] highlights the development of a deep learning model that integrates multi-level data to categorize plant diseases which highlighting the application of machine learning to farming to enhance crop health via precise classification. Gradient Boosting Regression (XGB) was also used by [12] to estimate rice yield, demonstrating how well it can capture complex patterns in agricultural data. Taken as a whole, these research demonstrate the variety of approaches and the increasing complexity of machine learning applications in agriculture, providing insightful information for improving crop suggestions and yield forecasts.

One important area where machine learning models have shown great promise is yield prediction. These models offer insightful crop production estimates that can help farmers and governments make well-informed decisions. Notable improvements in this field have been made possible by recent studies. For example, [13] used climate and remote sensing data to forecast agricultural yield using Decision Tree-based regression models. Their research emphasizes how crucial it is to combine different data sources in order to improve forecast accuracy. Reference [14] uses a neural network model to identify plant diseases. To increase prediction accuracy, the system augments meteorological data and uses a multi-level attention strategy. Reference [15] combines text and visual data to identify pests in wolfberry farming. By focusing on pest control that is essential for maximizing crop health and output which relates to our study. In their investigation of deep learning methods for predicting wheat production, [16] shown how well neural networks comprehend intricate relationships seen in agricultural data. Convolutional Neural Networks (CNNs) were used by [17] to estimate rice yield, highlighting the need of adding image-based data to yield prediction models. Moreover, [18] examined the use of Long Short-Term Memory (LSTM) networks for spatial yield forecasting. Their study advances the field of agricultural production forecasting by highlighting the advantages of recurrent neural networks in handling sequential data.

For prediction models to be effective, it is necessary to comprehend the elements that influence crop output. New studies have shed important light on these variables. For example, [19] carried out an extensive investigation into the ways in which crop output is influenced by soil conditions, which is essential for choosing pertinent features for machine learning models. In their investigation of how climatic factors affect agricultural production, [20] emphasized the need of weather information for precise yield forecasts. Precision farming, which optimizes agricultural inputs to maximum production and sustainability, has greatly improved as a result of the use of machine learning to agriculture. A evaluation of how machine learning approaches may improve crop management practices was conducted by [21]. The use of machine learning algorithms in automated irrigation systems was investigated by [22], who demonstrated how data-driven techniques may increase water usage efficiency. In their discussion of the application of machine learning models for disease and pest detection, [23] emphasized the importance of early diagnosis in crop health management.

Additionally, a number of research have suggested useful frameworks for machine learning-based intelligent crop management. Reference [24] developed a machine learning-based crop selection and optimization of yield system especially for small-scale farmers. A precision agriculture decision support system that combines crop, weather, and soil data to provide real-time suggestions was presented by [25]. A cloud-based platform for agricultural data analysis was provided by [26], which facilitates the use of machine learning models in remote and resource-constrained locations. There are still difficulties in using machine learning in agriculture, despite these developments. Reference [27] talked about the shortcomings of the machine learning models that are currently in use, especially with regard to data variety and model interpretability. Reference [28] discussed the infrastructural issues that farmers in poor countries confront and offered fixes to increase accessibility to machine learning technology. In their study of the moral issues surrounding the use of machine learning models in agriculture, [29] emphasized the significance of justice and openness.

To sum up, the incorporation of machine learning techniques into crop selection and yield prediction has

demonstrated encouraging prospects, providing data-driven methods to enhance agricultural sustainability and productivity. Subsequent investigations have to concentrate on surmounting current obstacles and improving machine learning models in order to provide precise and practical insights for farmers worldwide.

Restate your thesis or research problem

III. METHODOLOGY

Initially the study employs a crop recommendation dataset to help choose the best crops to cultivate. This data is based on weather and soil parameters. It comprises input attributes like the pH, temperature, humidity, rainfall, and the amounts of nitrogen (N), potassium (K), and phosphorus (P) in the soil. There are 2,200 items in the collection and each of which describes a set of conditions and the best crop to produce under those settings. The model that makes recommendations uses a variety of seven different robust machine learning approaches to generate models and matches these factors with 22 different crop kinds. These models advise farmers on which crops are best to plant, and our objective is to forecast precise outcomes from a limited dataset without over-fitting. The addition of several classifiers, each of which has been refined and assessed to determine which is best for the input data, is a noteworthy improvement. Furthermore, our methodology incorporates feature engineering and k-fold cross-validation techniques, which are essential for enhancing the precision of outcomes from small datasets and helping farmers choose the most relevant characteristics. We suggested applying different machine learning classification methods for crop analysis that have different properties. Based on their distinct qualities, these algorithms—which include Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM)—were carefully chosen. We carried out a thorough examination of the data, emphasizing the crucial elements that contributed to the high accuracy. We emphasized the value of obtaining high-quality data and the significance of doing preprocessing operations including feature engineering, cleaning, and transformation (e.g., MinMax Scalar). An overview of our approach to crop analysis and classification, incorporating machine learning methods, is shown in Fig. 5.

Explain the approach you chose

Referring to Fig. 5, we have incorporated a multi-stage process into our framework. The following steps are included in these stages:

1) Compiling Dataset
2) Feature Engineering
3) Explainable AI (XAI) Techniques
4) Feature Selection
5) Using Different Machine Learning Algorithms
6) Offering Suggested Crops through Testing

Explain the methods you used to analyze the data you collected

A. COMPILING DATASET

The dataset contains information on temperature, precipitation, humidity, soil pH, nitrogen (N), phosphorus (P), and potassium (K). The Kaggle platform provided the dataset. Mango, papaya, apple, banana, orange, pomegranate, grapes, watermelon, muskmelon, coconut, mung beans, mung bean, chickpea, kidney beans, pigeon peas, black gram, cotton, coffee, jute, and moth beans are among the unique crops that can be found in the dataset.

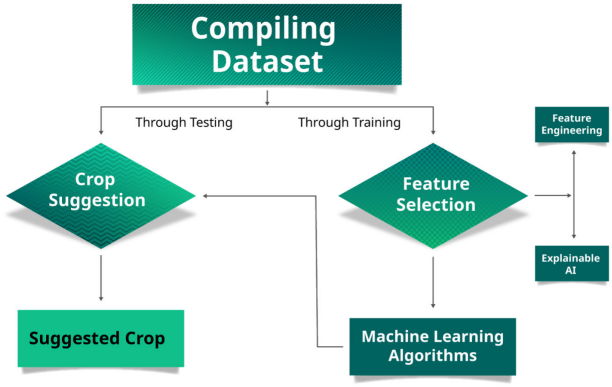

FIGURE 5. Block diagram outlining the whole proposed system methodology.

B. FEATURE ENGINEERING

In machine learning, feature engineering is the process of adding new features or changing preexisting ones in order to enhance model performance. It covers methods such as generating interaction terms, normalizing data, encoding categorical variables, and collecting new features from preexisting information. More accurate predictions are produced by models with improved underlying pattern recognition due to effective feature engineering. It is an essential phase in the pipeline of preprocessing data that improves the model's capacity to generalize from the training set.

Explain any uncommon methodology you use

C. EXPLAINABLE AI (XAI) TECHNIQUES

Explainable AI (XAI) in machine learning refers to methods that render model outputs comprehensible to humans. It increases openness and trust by offering insights into the decision-making processes of models. XAI ensures that consumers can comprehend and have confidence in AI-driven suggestions by streamlining the decision-making process and assisting users in properly interpreting outcomes. This is important in areas where decisions affect actual results, such as agriculture. Our research incorporates XAI to enhance the crop recommendation model's interpretability. We focused on using XAI techniques such as Local Interpretable Model-agnostic Explanations (LIME) and feature importance to understand the behavior of the models. These included XAI techniques are as follows:

1) LIME (Local Interpretable Model-agnostic Explanations):
LIME uses a more straightforward, interpretable model to locally approximate the model in order to explain

Evaluate and justify the methodological choices you made

specific predictions. It facilitates comprehending the relative contributions of each characteristic to a given forecast.

$$\hat{f}(x) \approx g(x) \tag{1}$$

2) Feature Importance Analysis:
Finding the characteristics that have the most impact on the model's predictions is made easier with the aid of feature importance analysis.

$$\text{Importance}(j) = \sum_{t \in \text{trees}} \text{Reduction}(t, j) \tag{2}$$

D. ML METHODS

Several machine learning techniques, including Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM), are utilized in this suggested system. We address many facets of data evaluation and categorization challenges using a range of these machine learning algorithms. Every algorithm has distinct qualities and benefits that make it appropriate for particular use cases. Here is a thorough explanation complete with benefits, use cases, and justification:

1) EXTRA TREE CLASSIFIER (ETC)

During training, the collaborative learning technique Extra Tree Classifier (EXT) builds several decision trees. By adding unpredictability to the tree-building process, it increases diversity and minimizes overfitting.

- Justification: Using randomized feature and data subsets, this ensemble approach creates many decision trees.
- Benefits: capable of handling both categorization and regression problems; less prone to overfitting than a single decision tree.
- Use Case: Good for datasets where obtaining excellent performance at a reasonable computing cost and decreasing variance are crucial.

2) LOGISTIC REGRESSION (LR)

One statistical technique for binary classification is Logistic Regression (LR).

- Justification: A linear framework that is beneficial in binary and multi-class classification scenarios. It models a binary dependent variable using the logistic function.
- Benefits: Simple to use, comprehensible, and offers categorization probabilities. functions best in cases when there is a roughly linear connection between the target variable and the characteristics.
- Use Case: Fits small- to medium-sized datasets with a decision boundary that is linear.

$$P(\text{Outcome} = 1|\text{Feature}) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 \text{Feature})}} \tag{3}$$

- $P(\text{Outcome} = 1|\text{Feature})$ denotes the probability that the target variable (Outcome) will be 1 given the input (Feature).
- α_0 and α_1 are the new coefficients.

3) DECISION TREE CLASSIFIER (DT)

For classification problems, the Decision Tree Classifier (DTC) constructs a decision tree model. By using majority voting, it divides the feature space into regions and gives each region a class label.

- Justification: To simulate decisions and their potential effects, a non-linear framework that divides data into branches is used.
- Benefits: Handling and interpreting numerical and categorical data is made simple. doesn't call for scaling of features.
- Use Case: Useful in situations where interpretability is essential, such as financial decision-making, for determining feature relevance and making judgments.

4) RANDOM FOREST (RF)

Random Forest is a popular collaborative learning method for applications involving regression and classification. It consists of many decision trees, each developed on a distinct subset of the data. The average of each tree's projections yields the final forecast.

- Justification: A group of decision trees that reduces overfitting and raises prediction accuracy.
- Benefits: Offers feature importance, is resistant to overfitting, and manages missing values effectively. functions effectively with data that is both categorized and numerical.
- Use Case: Suitable for intricate datasets that contain missing or noisy data.

5) K-NEAREST NEIGHBOUR (KNN)

A simple yet effective method for machine learning problems with classification is the K-Nearest Neighbors (KNN) classifier. A class label is assigned to a data point according to the adjacent neighbors' majority categorization.

- Justification: Data is categorized using a straightforward instance-based learning technique by comparing it to a large class of its closest neighbors.
- Benefits: Does well with tiny datasets, requires no training step, and is simple to comprehend and apply.
- Use Case: Fits well with datasets with extremely non-linear feature relationships.

$$\hat{y} = \text{mode}(y_{i_1}, y_{i_2}, \ldots, y_{i_k}) \tag{4}$$

6) GAUSSIAN NAIVE BAYES (GNB)

The Gaussian Naive Bayes (GNB) classifier is a probabilistic model that relates the Bayes theorem under the premise of feature independence. It calculates the likelihood of a class given a set of attributes using the Gaussian probability density function.

- Justification: A probabilistic classifier that assumes feature independence and is based on the Bayes theorem.
- Benefits: Easy to use, quick, and efficient with small datasets. handles continuous data with a Gaussian distribution assumption.

- Use Case: When the independence requirement is properly met, effective in text categorization and spam detection.

$$P(c|\mathbf{v}) = \frac{1}{Z}P(c)\prod_{j=1}^{m}P(v_j|c) \tag{5}$$

In this equation:

- $P(c|\mathbf{v})$ represents the probability of class c given the features $\mathbf{v}$.
- Z is a normalization factor.
- $P(c)$ is the prior probability of class c.
- $P(v_j|c)$ is the conditional probability of feature v_j given class c.
- m is the number of features.

7) SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) classifier is a useful method for solving classification difficulties. Its decision limit is defined by a hyperplane that optimizes the distance among units.

- Justification: A strong supervised learning method that determines the best hyperplane to split groups along for categorization.
- Benefits: Versatile, resilient against overfitting (particularly when using kernels), efficient in high-dimensional areas.
- Use Case: Good for text classification and picture categorization, and appropriate for datasets with a distinct margin of separation.

$$\text{minimize}\frac{1}{2}\|\mathbf{v}\|^2 + \lambda\sum_{i=1}^{m}\zeta_i \tag{6}$$

subject to:

$$z_i(\mathbf{v}\cdot\mathbf{x}_i + d) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$$

where:

- $\mathbf{v}$ is the weight vector,
- d is the bias term,
- λ is the regularization parameter,
- $\mathbf{x}_i$ is the feature vector for the ith data point,
- z_i is the class label for the ith data point,
- ζ_i are slack variables.

In our second study domain, we used data on wheat production yield in Pakistan, which includes data from 1992 to 2013 as our second research undertaking. The outputs of the dataset are the area, expressed in hectares, and the yield, expressed in hectograms per hectare (Hg/Ha). The average temperature, the amount of pesticide used in tons, and the average yearly rainfall in millimeters are the inputs or qualities, as seen in Fig. 6. The next year, 2025, crop production was predicted using five different machine learning regression techniques.

The diagrammatic representation the entire procedure includes a multi-stage process into our proposed architecture (refer to Fig. 7).

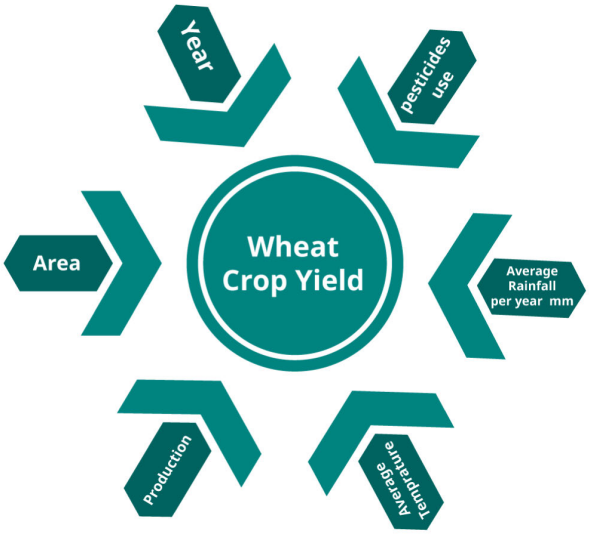

FIGURE 6. Block diagram showing yield input attributes and the final output.

These stages in the proposed architecture are:

- Data Collection
- Preprocessing
- Apply Robust ML algorithms
- Check predictions performance

a: DATA COLLECTION

The yield prediction dataset, which examined wheat output in Pakistan, was obtained from the World Bank and FAO, covered the years 1992–2013.

Describe how you collected the data you used

b: PREPROCESSING

Data cleaning is the first step after gathering data from publically accessible repositories. Shared columns can be used to merge the data frames after the data cleaning process is complete. To guarantee uniformity across all attributes, normalization is required. The final elements of the generated data frame are expected to be crop (Wheat), expressed in hectograms per hectare (Hg/Ha). We utilized the MICE (Multiple Imputation by Chained Equations) imputer technique to resolve missing data spanning from 2014 to 2024. The steps associated with MICE are given below:

- To begin, use first approximations such as mean or median to fill in the blank numbers.
- Based on other variables, use regression models to forecast each variable's missing values.
- Use the anticipated values to update any missing data.
- Imputation should be done several times for each variable (chaining).
- To handle variability, create many imputed datasets (usually five to ten) and aggregate the results using Rubin's guidelines.

Let Y denote the dataset with missing values, where $Y = (Y_{obs}, Y_{mis})$. The MICE algorithm involves the following steps:

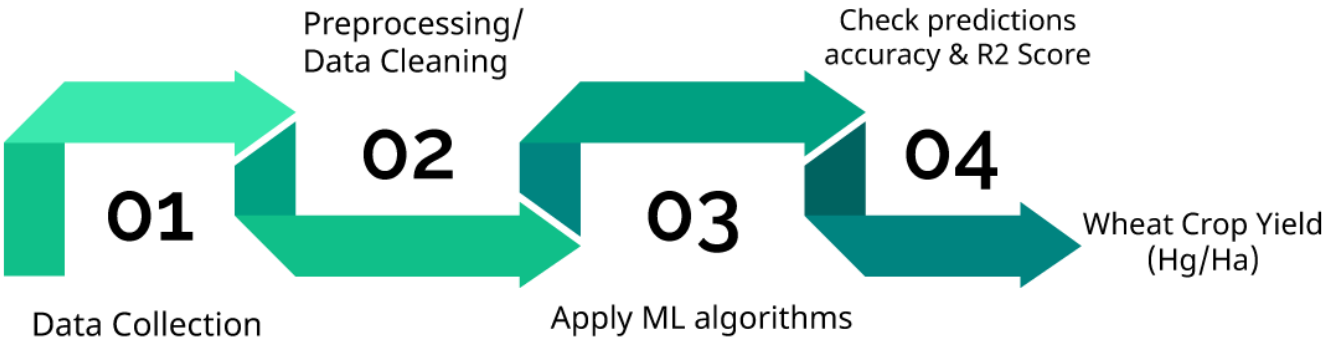

FIGURE 7. Block diagram illustrating the proposed methodology.

1) **Initialize:** Impute missing values Y_{mis} with initial estimates $\hat{Y}_{mis}^{(0)}$. For example:

$$\hat{Y}_{mis}^{(0)} = \bar{Y}_{obs} \tag{7}$$

where $\bar{Y}_{obs}$ is the mean of the observed values.

2) **Iterative Imputation:** For each variable Y_j with missing data, perform regression imputation. Predict Y_j based on other variables:

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{8}$$

where $X_1, X_2, \ldots, X_k$ are the other variables, and ϵ is the error term.

3) **Update:** Update the imputed values with the predictions:

$$\hat{Y}_{mis}^{(t+1)} = \hat{Y}_j \tag{9}$$

where $\hat{Y}_j$ are the predicted values from the regression model.

4) **Iterate:** Repeat the imputation for all variables with missing data multiple times, updating values each iteration:

$$\hat{Y}_{mis}^{(t+1)} \text{ for } t = 1, 2, \ldots, T \tag{10}$$

where T is the number of iterations.

5) **Create Multiple Datasets**: Generate k imputed datasets:

$$D^{(1)}, D^{(2)}, \ldots, D^{(k)} \tag{11}$$

Each dataset represents a distinct imputation outcome.

6) **Results**:Analyze each imputed dataset separately and combine the results using Rubin's rules:

$$\bar{\theta} = \frac{1}{n}\sum_{j=1}^{n} \theta^{(j)}$$
$$\text{Var}(\bar{\theta}) = \frac{1}{n}\sum_{j=1}^{n} \text{Var}(\theta^{(j)}) + \frac{1+\frac{1}{n}}{n}\text{Var}(\theta^{(j)})_{\text{between}} \tag{12}$$

where $\theta^{(j)}$ represents the estimates from each imputation dataset, and $\text{Var}(\theta^{(j)})_{\text{between}}$ denotes the variance between imputations.

c: APPLY ROBUST ML ALGORITHMS

To ensure optimal efficiency, we have predicted outcomes employing many machine learning techniques and then evaluated them. To anticipate agricultural yield, we have used a number of machine learning models in our research. To improve model performance, we used optimizers for machine learning algorithms to apply hyperparameter adjustment in our research study. We've also done a comparison study between these models, which consist of:

- Decision Tree Regressor (DTR)
- Random Forest Regressor (RFR)
- K-Nearest Neighbors Regressor (KNNR)
- Support Vector Regressor (SVR)
- Extreme Gradient Boosting Regressor (XGBoostR)

Here is a thorough explanation of each algorithm's benefits, use cases, and justification. It also explains how optimizers may improve performance by adjusting hyperparameters:

1. Decision Tree Regressor (DTR):

- Justification: The Decision Tree Regressor was selected because of its capacity to represent intricate, non-linear associations between the target variable and features.
- Benefits:
  Simplicity and Interpretability: The judgments made by the model are easily interpreted since they are simple to comprehend and depict. Non-Linear Relationships: Do not need a linear assumption to describe complex patterns.
  Importance of Features: Indicates which features are most crucial for forecasts.
- Use Case: Perfect for datasets that benefit from interpretability and basic decision rules. It is useful to know which factors have the greatest impact on agricultural output in order to anticipate yield.
- Optimization:

Hyperparameter Tuning: The model's hyperparameters include the tree's depth, the minimum number of samples needed to split a node, the minimum number of samples needed at a leaf node, and the standard used to gauge the quality of a split (such as Mean Squared Error or Mean Absolute Error). Grid Search and Random Search techniques were used to investigate different combinations of these hyperparameters. Furthermore, 5-fold cross-validation was applied to evaluate the resilience and performance of the model. Finding the ideal settings that strike a compromise between underfitting and overfitting was the goal of the tuning procedure. Shallower trees with more samples may underfit the data, whereas deeper trees with fewer minimum samples may capture more complexity but may also result in overfitting.

2. Random Forest Regressor (RFR):

- justification: Random Forest Regressor combines many trees to boost prediction accuracy and resilience, which is an improvement over Decision Tree Regressor.
- Benefits: Ensemble learning reduces overfitting and increases accuracy by combining predictions from several decision trees.
  Robustness: Because several trees are averaged, it can handle noisy data and outliers more effectively.
  Feature significance: Offers cumulative measurements of feature significance.
- Use Case: Fits well with intricate datasets that have a lot of characteristics and interactions. By reducing the influence of noise and outliers, it offers a solid and trustworthy estimate for yield prediction.
- Optimization:
  Hyperparameter Tuning: During the tuning process, certain hyperparameters were taken into account, such as the number of trees, their maximum depth, the amount of features to be taken into account at each split, the minimum sample split, and the minimum sample leaf. In order to determine the best values for these parameters, several combinations were tested using Grid Search and Random Search approaches. Additionally, the performance and stability of the model were evaluated using 5-fold cross-validation. The goal of the tuning procedure was to control computational complexity while improving model performance. Notably, adding more trees increased computation time while typically improving accuracy.

3. K-Nearest Neighbors Regressor (KNNR):

- Justification: The K-Nearest Neighbors Regressor is used since it's easy to use and good at identifying small-scale patterns and correlations in the data.
- Benefits:
  Instance-Based Learning: Enables adaptable and intuitive modeling by generating predictions based on the similarity of examples.
  Non-Parametric: Does not presuppose that the connection between the target and characteristics has a certain shape.
- Use Case: Ideal for datasets with substantial local patterns and correlations. By taking comparable cases into account, yield prediction may adjust to changing circumstances.
- Optimization:
  Hyperparameter Tuning: The number of neighbors, the weighting strategy (uniform or distance), and the closest neighbors search algorithm (auto, ball_tree, kd_tree, brute) are the hyperparameters for the nearest neighbors algorithm. Grid Search and Random Search techniques were both used to find the optimal set of these hyperparameters. Furthermore, 5-fold cross-validation was used to assess the model's effectiveness and capacity for generalizing to new data. The trade-off between variance and bias must be balanced in order to determine the impact of adjusting these factors. While fewer neighbors capture more local trends but run risk of overfitting, more neighbors tend to smooth forecasts but can also cause underfitting. In order to accomplish this balance and enhance the overall performance of the model, optimal parameters were chosen.

4. Support Vector Regressor (SVR):

- Justification: Support Vector Regressor (SVR) is utilized because of its versatility in modeling complicated connections and its capacity to handle high-dimensional data.
- Benefits:
  Effective in High Dimensions: Functions effectively when there are a lot of features compared to samples.
  Robustness: Able to manage non-linearity and produce forecasts highly accurate forecasts.
- Use Case: Fit for high-feature datasets or datasets with non-linear feature-to-target relationships. Support Vector Regressor aids in capturing the intricate correlations that exist between agricultural production and meteorological factors when predicting yield.
- Optimization:
  Hyperparameter Tuning: The Support Vector Regression (SVR) model underwent hyperparameter tuning through the examination of many critical parameters, including kernel type, the regularization parameter C, epsilon (ϵ), and gamma (γ). The flexibility and performance of the model are greatly influenced by the kind of kernel, which might be linear, polynomial, or RBF. A greater value of the regularization parameter C usually leads to a narrower margin and fewer support vectors, which may cause overfitting. This parameter manages the trade-off between obtaining a low training error and a low testing error. On the other hand, a smaller C value results in a bigger margin but might lead to underfitting. The epsilon-SVR model's epsilon (ϵ) defines the tube that accepts predictions, while gamma (γ) controls how

much a single training sample influences the decision boundary, particularly in the sigmoid, polynomial, and RBF kernels.

To investigate different combinations of these parameters, both Grid Search and Random Search approaches were used throughout the tuning process. The model's performance was assessed using 5-fold cross-validation to guarantee successful generalization. This method aids in choosing the ideal parameters that strike a compromise between model fit and complexity, guaranteeing reliable model performance on a variety of datasets.

5. *Extreme Gradient Boosting Regressor (XGBoostR):*

- Justification: The boosting approach of Extreme Gradient Boosting Regressor (XGBoostR) allows for state-of-the-art performance in prediction tasks, which is why it was selected.
- Benefits: Boosting Technique: Increases accuracy by building models sequentially to fix earlier iterations' mistakes. Regularization: Consists of regularization to enhance model generalization and avoid overfitting. Efficiency: Well-known for its quickness and potency while working with big datasets.
- Use Case: Optimal for intricate datasets where superior precision and efficiency are essential. Extreme Gradient Boosting Regressor can manage the interplay between several characteristics in yield prediction and generate precise forecasts.
- Optimization:

Hyperparameter Tuning: The Learning Rate (eta), Max Depth, Number of Estimators, Subsample, Colsample Bytree, and Gamma are the main hyperparameters that need to be adjusted for the model. While Max Depth establishes the maximum depth of each tree, Learning Rate regulates each tree's contribution to the final prediction. The overall number of trees in the model is shown by the Number of Estimators. The proportion of samples used to fit specific trees is defined by the Subsample parameter, while the fraction of features utilized for each tree is indicated by the Colsample Bytree. Gamma denotes the minimal decrease in loss necessary to create a further division, hence managing the intricacy of the trees.

Grid Search and Random Search were used in the tuning process to investigate different combinations of these hyperparameters. The model was evaluated using 5-fold cross-validation to make sure it was generalizable. During this procedure, the best hyperparameters were chosen in order to balance model complexity and performance. More specifically, the Max Depth and Number of Estimators impact the model's capability, whilst the Learning Rate determines how quickly the model learns. Regularization is influenced by the Subsample and Colsample Bytree parameters, and Gamma is essential for managing the trees' complexity.

Discuss any obstacles and their solutions

IV. RESULTS AND DISCUSSION

This section discusses and presents the outcomes of all the included classification and regression machine learning models. We applied the suggested recommendation method to classify 22 crops, including cotton, rice, bananas, apples, and coffee, among others. Seven machine learning techniques were used for the categorization, including Random Forest, Extra Tree Classifier, Logistic Regression, Decision Tree, K-Nearest Neighbour, Gaussian Naive Bayes, and Support Vector Machine.

We have created a wheat yield forecast model for Pakistan. Five machine learning techniques, including the Decision Tree Regressor, were used to forecast wheat yields, which are expressed in hectograms per hectare. Anaconda and Jupyter Notebook, two Python tools, were used in the study for data preparation, visualization, prediction, and categorization. Preprocessed data is divided into training (70 %) and testing (30 %) sets. Training set helps machine learning algorithm to learn from available data whereas testing set evaluates its performance. Finding the optimal models for crop categorization and yield data prediction was the aim of this approach.

A. DATA COLLECTION

In the first study, we used seven machine learning algorithms to evaluate farming data and suggest the best crops for farmers. Features like nitrogen concentration, temperature, soil pH, rainfall, humidity, phosphorus content, and potassium content are included in the dataset, which was obtained via Kaggle. It has 2200 records with 22 crop labels, including beans, rice, black gram, jute chickpeas, among other crops.

The outcomes showed that the most accurate algorithms were Random Forest, Extra Tree Classifier, and Gaussian NB attaining 99 while others going beyond 96. Subsequently, we employed k-fold cross-validation techniques, which enhanced performance and other metrics such as confusion matrix, precision (13), recall (14), and F1 Score (15) in comparison to baseline models as shown in Table 1. The precision of positive predictions is represented by the ratio of true-positive (TP) to the total of TP and false-positive (FP). The ratio of TP to the total of TP and false-negative (FN) is known as recall, or sensitivity, and it indicates how well the model can detect positive occurrences. The F1 score, which is particularly helpful for unbalanced datasets, is the harmonic mean of accuracy and recall. It offers a single statistic that balances both features.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{13}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{14}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

We use additional methods to improve our model and identify the critical components required to get maximum

TABLE 1. Comparing various models' Precision, Recall, and F1 Score.

Classifier	Precision	Recall	F1 Score
Extra Trees	0.9929	0.9927	0.9927
Logistic Regression	0.9589	0.9586	0.9585
Decision Tree	0.9884	0.9882	0.9882
Random Forest	0.9948	0.9945	0.9945
K-Nearest Neighbours	0.9817	0.9814	0.9814
GaussianNB	0.9919	0.9914	0.9913
SVM	0.9868	0.9868	0.9868

accuracy. We chose the most pertinent characteristics by employing strong machine learning models and feature engineering approaches. We also made use of XAI approaches, such as feature significance analysis to comprehend and display the influence of features on model predictions and LIME for model interpretability. According to the *Feature Importance Comparison*, The Extra Trees Classifier, Random Forest Classifier, and Decision Tree Classifier feature significance ratings are displayed in Fig. 10. The Random Forest Classifier prioritizes 'humidity' (0.199) and 'rainfall' (0.167), suggesting that these attributes are crucial for its forecasts. Decision Tree Classifier gives priority to rainfall' (0.263) and Phosphorus (P) (0.227), but Extra Trees Classifier values 'humidity (0.178) and 'Potassium (K) (0.169) higher. Because it effectively distributes significance among important characteristics, Random Forest Classifier performs best overall, improving prediction accuracy. Support Vector Machine, Gaussian Naive Bayes, and K-Nearest Neighbors Classifier are the models that are unsuitable for feature significance comparison. The Gaussian Naive Bayes algorithm concentrates on probabilities without offering feature significance ratings, while the K-Nearest Neighbors Classifier employs distances rather than feature weights. Support Vector Machine lacks a clear relevance metric, particularly when using non-linear kernels like RBF. Although normalization may be required for comparability, coefficients from Logistic Regression can be seen as measures of feature relevance. The model emphasizes humidity a lot which shows how important it is for crop categorization. Moreover, the relevance of rainfall and potassium emphasizes how crucial it is to incorporate water resources and nutrient availability into agricultural operations. In addition to making accurate predictions the Extra Trees Classifier makes clear how different traits affect crop production. The feature models correlation matrix and bar plot to compare each model, are shown in Figures 8 and 9 respectively.

Because of its functionality the agricultural planners who are interested in efficiently maximizing crop output will find it to be a valuable resource. In the end the model capacity to incorporate many elements and provide insightful analysis of their applicability improves agricultural decisions.

The *LIME Explanations Comparison Graph* in Fig. 13, shows how characteristics affect the model's predictions. Negative values point to a disadvantage, whilst positive ones show a feature improves forecasts. The model with

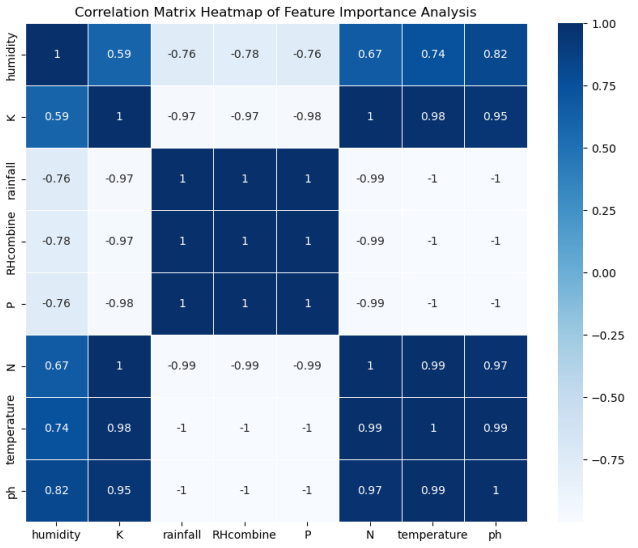

FIGURE 8. The graph displays the correlation of feature importance for all models.

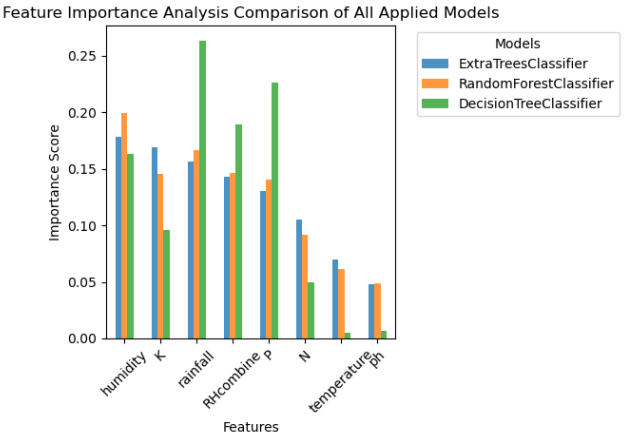

FIGURE 9. The bar graph displays the comparison of feature importance for all models.

the highest performance is Gaussian Naive Bayes (GNB), which has significant positive contributions from characteristics such as 0.87 < RHcombine <= 1.24 and 60.16 < humidity <= 80.55. This implies that it makes good use of these attributes to enhance predictions. The general resilience of the model is partly a result of its balanced negative consequences across different characteristics. The majority of the input parameters show consistently modest negative correlations which indicating a steady performance devoid of a clear preference for any one scenario. It is noteworthy that there is a positive link between it and relative humidity which is suggesting that it is sensitive to favorable growing circumstances. The residual plot and correlation matrix for each of the models in the LIME are shown in Figures 11 and 12 respectively.

This ensemble model improves generality and accuracy which makes it a great option for agricultural production

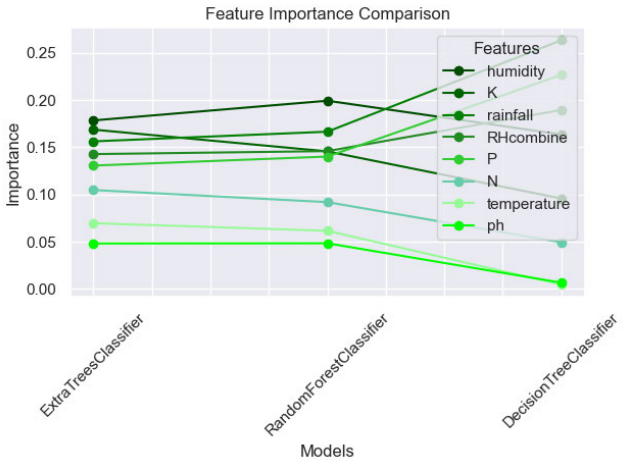

FIGURE 10. The graph displays the comparison of feature importance for each classifier.

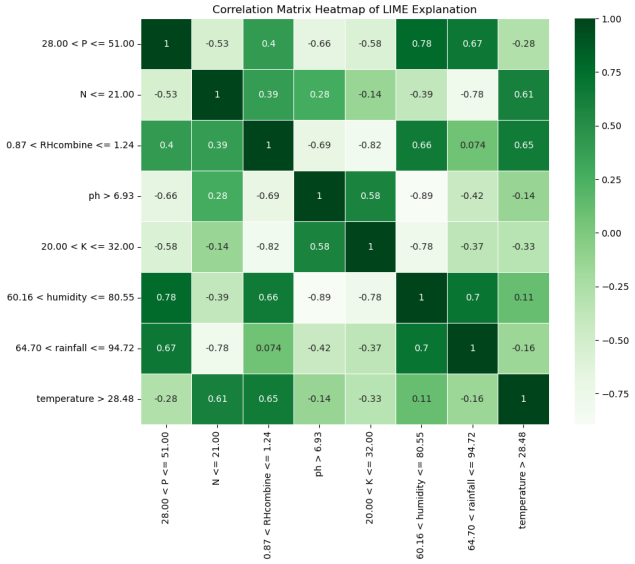

FIGURE 11. The correlation matrix the LIME explanations comparison of all models.

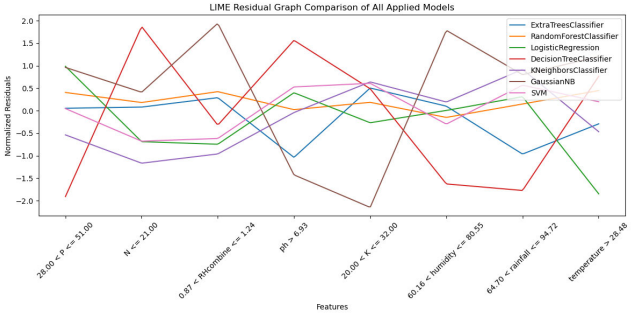

FIGURE 12. The residual graph displays the LIME explanations comparison of all models.

estimates and crop recommendations. Our comprehension of its actual usefulness will be strengthened by additional assessment of its correctness and precision.

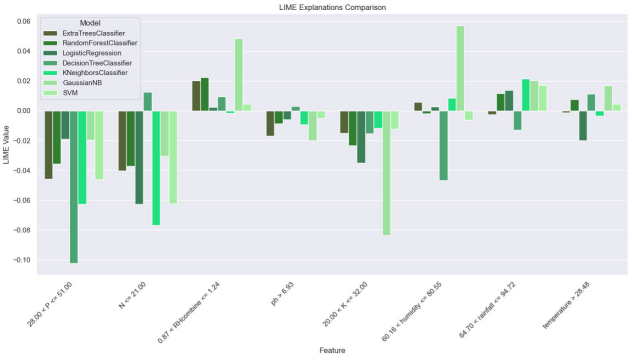

FIGURE 13. The graph displays the LIME explanations comparison for each classifier.

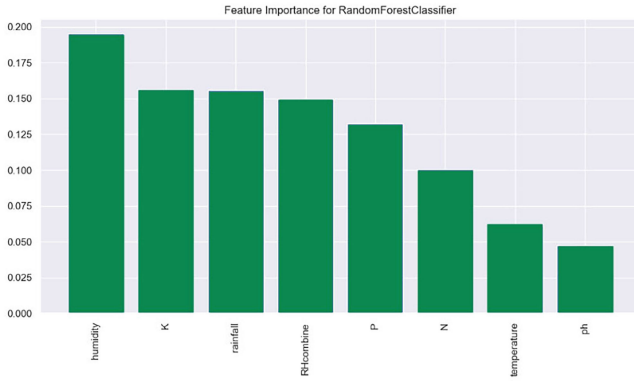

FIGURE 14. The feature significance of Random Forest Classifier.

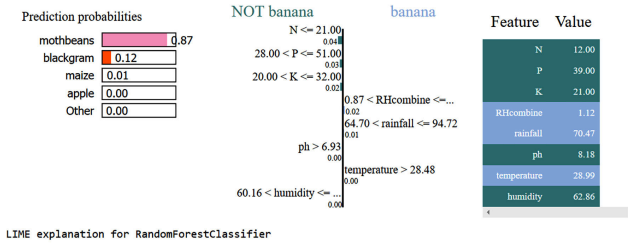

FIGURE 15. The LIME interpretations of Random Forest Classifier.

According to our findings, the Extra Tree Classifier and Gaussian Naive Bayes both had accuracy rates of 99.4% and 99.2%, respectively. However, Random Forest Classifier, with 99.7% accuracy, was the most accurate. Fig. 14 and Fig. 15 depict the feature significance and LIME interpretations for Random Forest model. The accuracies of Logistic Regression, K-Nearest Neighbour, and Support Vector Machine were 95.9%, 98.1%, and 98.9%, respectively. Fig. 16 illustrates how we evaluated each model by comparing it based on these parameters after obtaining the accuracy and cross-validation scores for each model.

These findings emphasize how crucial it is to choose the right characteristics for crop recognition, and the features that are found can direct further agricultural data research.

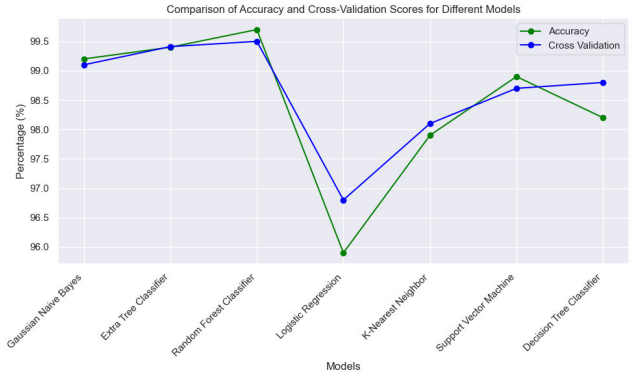

FIGURE 16. The comparison of the cross-validation and accuracy scores for each classification technique is shown on the line graph.

We also looked into how crop labels may affect the accuracy of the system. Based on factors such as growth characteristics, utilization, water needs, and harvest techniques, we divided crops into broad groups. The purpose of this categorization was to find out if the algorithm could correctly anticipate these more general groups. Overall, our research highlights the value of feature selection in raising model accuracy as well as the efficiency of machine learning in crop recommendation.

Predicting the wheat crop production for the next year was the main goal of the second research. Since our original dataset only included data from 2013 to 2024, we initially used the MICE imputation approach to produce data values from 2014 to 2024. Preprocessed data is divided into training (70%) and testing (30%) sets. Training data teaches machine learning algorithms for accurate predictions; testing data evaluates their performance. Five regression models were then trained for this prediction job.

Mean Absolute Error	Mean Squared Error
$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
Root Mean Squared Error	**Standard Deviation**
$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	$\text{STD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Table 2 assesses the effectiveness of several regression approaches using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Standard Deviation (STD). Support Vector Regressor performs better than the other models, having the lowest RMSE (0.3236), MAE (0.1614), and MSE (0.1047). This suggests that Support Vector Regressor makes the most reliable and accurate predictions. On the other hand, the error rates are greater for the K-Nearest Neighbors, Extreme Gradient Boosting, Random Forest, and Decision Tree regressions; the MAE and RMSE values are highest for the K-Nearest Neighbors and Extreme Gradient Boosting regressions. As a result, Support Vector Regressor performs the best in this evaluation.

We compared and assessed each model according to performance criteria in order to determine which was the greatest match and to make sure it was neither overfitting nor underfitting. Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbor Regressor, Support Vector Regressor andExtreme Gradient Boosting Regressor were among the models employed in this investigation. Referring to Fig. 17, we computed their R^2 (coefficient of determination) score by comparing their results using the Root Mean Square Error (RMSE).

The Support Vector Regressor has the highest R^2 value of (99.99%), according to the data. Fig. 18 shows the anticipated and actual values for the algorithms. In order to improve the predictability of wheat yields, this study proposes a machine learning method that uses ensemble models to improve crop yield prediction at the county level. Furthermore, when compared to real yield data, the predictions made by these machine learning models showed less bias.

TABLE 2. Comparison of regressors.

Regressor	MAE	MSE	RMSE	Standard deviation
SVR	0.1614	0.1047	0.3236	26.97
KNN Regressor	0.4540	0.3538	0.5948	26.92
XGB Regressor	0.4407	0.4346	0.6592	27.13
RF Regressor	0.3931	0.3754	0.6127	27.05
DT Regressor	0.4282	0.4289	0.6549	27.14

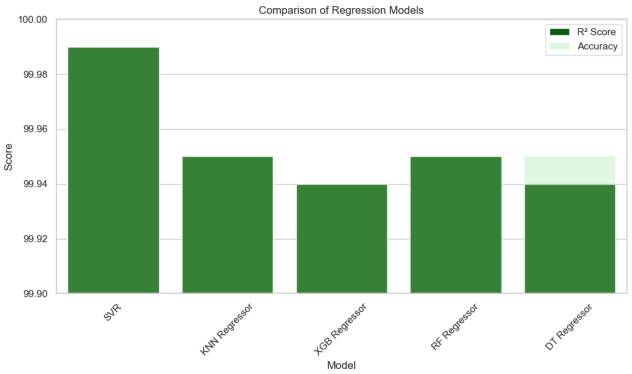

FIGURE 17. The graph shows accuracy and R^2 score of each regression algorithms.

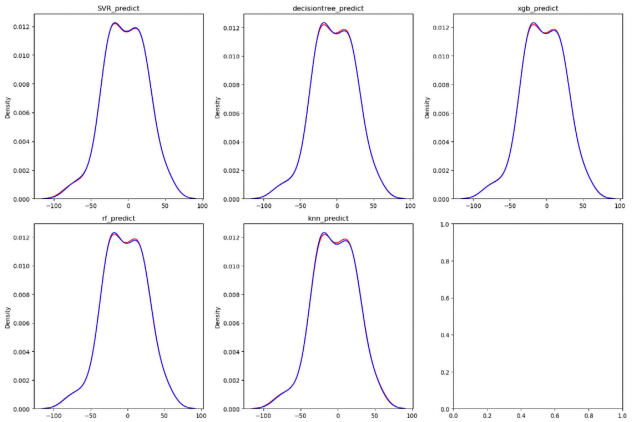

FIGURE 18. Actual vs predicted values of the algorithms.

V. CONCLUSION

In conclusion, our first study developed crop recommendation models that use cutting-edge machine learning algorithms to determine which crops are most suited for planting. This approach is adaptable and scalable, making it simple to apply to new data sets and geographical locations. Also transparency and trust are improved by explainable artificial intelligence (XAI)-based crop advice systems. XAI offers lucid insights into model selections, demonstrating how factors like as rainfall and soil pH affect suggestions. This openness encourages user trust and helps farmers make better decisions. Furthermore, XAI permits a more efficient use of AI insights by highlighting potential biases, guaranteeing accountability, and supporting user education by emphasizing the significance of different crop selection factors. The results have a number of advantageous ramifications for the agriculture industry. First off, farmers may use this method to choose crops with more knowledge and understanding. Secondly, these techniques may be employed by governments to formulate policies that promote agriculture. Thirdly, companies may use these insights to create innovative goods and services that boost the agriculture sector and contribute to price stability. All things considered, this study represents a major breakthrough in agriculture by providing governments, and companies with a scalable and accurate model.

Summarize the main points

State the significance or results

The research domain can be expanded in a variety of ways, such as:

1) Surveying farmers to evaluate the financial savings made possible by these methods would demonstrate the machine learning models' financial impact.
2) Creating a mobile application that serves as an end-to-end system for agribusiness owners and farmers, making the models that have been presented easily accessible.
3) Compiling information from diverse areas to assess the models' accuracy in various circumstances.
4) Attempting larger datasets, which would improve the models' comprehension of the connections between different variables and agricultural yields.
5) Evaluating the technique's advantages for farmers and the larger ecosystem in terms of the economy and ecology.
6) Putting sensors on farms to gather data in real time, enhance crop suggestions, reduce crop loss risks, and improve sustainability and decision-making.

Different machine learning regression algorithms were used in the second research task in this study to predict Pakistan's wheat crop productivity. Using input parameters such the production year, average annual rainfall (mm), pesticide usage (tonnes), and average temperature for that year, we were able to forecast wheat yield in hectogram per hectare (Hg/Ha) using a dataset. When estimating agricultural yield, the Support Vector Regressor showed the best accuracy shown in Fig. 13. Future work in this research will entail adding more pertinent data to the model to improve its predictions.

Conclude your thoughts

REFERENCES

[1] A. Zakaria, A. Y. M. Shakaff, M. J. Masnan, F. S. A. Saad, A. H. Adom, M. N. Ahmad, M. N. Jaafar, A. H. Abdullah, and L. M. Kamarudin, "Improved maturity and ripeness classifications of *Magnifera indica* cv. Harumanis mangoes through sensor fusion of an electronic nose and acoustic sensor," *Sensors*, vol. 12, no. 5, pp. 6023–6048, May 2012.

[2] M. Garanayak, G. Sahu, S. N. Mohanty, and A. K. Jagadev, "Agricultural recommendation system for crops using different machine learning regression methods," *Int. J. Agricult. Environ. Inf. Syst.*, vol. 12, no. 1, pp. 1–20, Jan. 2021.

[3] D. Dighe, A. Pawar, R. Nagpure, and S. Kharat, "Survey of crop recommendation systems," *Int. Res. J. Eng. Technol.*, vol. 5, pp. 476–481, May 2018.

[4] R. V. Saraswathi, J. Sridharani, P. S. Chowdary, K. Nikhil, M. S. Harshitha, and K. M. Sai, "Smart farming: The IoT based future agriculture," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Jan. 2022, pp. 150–155.

[5] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, "Crop recommendation system for precision agriculture," in *Proc. 8th Int. Conf. Adv. Comput. (ICoAC)*, Jan. 2017, pp. 32–36.

[6] A. Ingle, "Crop recommendation dataset," *Kaggle*. Accessed: Jul. 24, 2024. [Online]. Available: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[7] Food Agricult. Org. (FAO), World Bank, Kaggle. *Crop Yield Prediction Dataset, Data Sources, and Agricultural Information*. Accessed: Jul. 3, 2024. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

[8] X. Gao, Y. Liu, L. Wang, and M. Sun, "Predicting crop yield using a linear regression model based on climate and soil data," *Comput. Electron. Agricult.*, vol. 193, Aug. 2022, Art. no. 106642.

[9] Y. Chen, X. Zhang, H. Li, and M. Hu, "Random forest regression for predicting maize yield with climate and soil data," *Field Crops Res.*, vol. 300, Apr. 2023, Art. no. 108938.

[10] L. Zhang, Y. Liu, Q. Chen, and Z. Wu, "Support vector regression for crop yield prediction using meteorological and soil variables," *Comput. Electron. Agricult.*, vol. 195, Dec. 2022, Art. no. 106788.

[11] G. Dai, Z. Tian, J. Fan, C. K. Sunil, and C. Dewi, "DFN-PSAN: Multi-level deep information feature fusion extraction network for interpretable plant disease classification," *Comput. Electron. Agricult.*, vol. 216, Jan. 2024, Art. no. 108481.

[12] M. I. Khan, T. A. Shah, S. Ahmed, and R. Singh, "Gradient boosting regression for predicting rice yield based on climate and soil features," *Environ. Model. Softw.*, vol. 164, Jun. 2024, Art. no. 105548.

[13] X. Jiang, Y. Sun, and F. Zhao, "Decision tree-based regression models for predicting crop yield using remote sensing and climate data," *Agricult. Syst.*, vol. 211, Jul. 2023, Art. no. 103277.

[14] G. Dai, J. Fan, Z. Tian, and C. Wang, "PPLC-Net: Neural network-based plant disease identification model supported by weather data augmentation and multi-level attention mechanism," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 5, May 2023, Art. no. 101555.

[15] G. Dai, J. Fan, and C. Dewi, "ITF-WPI: Image and text based cross-modal feature fusion model for wolfberry pest recognition," *Comput. Electron. Agricult.*, vol. 212, Sep. 2023, Art. no. 108129.

[16] Y. Liu, M. Gao, Z. Huang, and S. Wang, "Deep learning models for wheat yield prediction using climate and soil data," *Comput. Electron. Agricult.*, vol. 192, Oct. 2022, Art. no. 106576.

[17] T. T. Nguyen, H. T. Pham, D. P. Hoang, and N. H. Vu, "Predicting rice yield using convolutional neural networks with satellite imagery," *Remote Sens. Environ.*, vol. 265, Sep. 2023, Art. no. 112633.

[18] R. Sharma, P. K. Tiwari, and K. Gupta, "Temporal crop yield prediction using LSTM networks," *Agricult. Forest Meteorol.*, vol. 318, Mar. 2024, Art. no. 108924.

[19] H. Patel, K. Shah, and A. Mehta, "Impact of soil properties on crop yield: A comprehensive analysis," *Soil Sci. Soc. Amer. J.*, vol. 87, pp. 45–58, Jul. 2023.

[20] P. Singh, V. Kumar, and S. Nair, "Role of climatic variables in crop productivity: An empirical study," *Agric. Forest Meteorol.*, vol. 308, Jan. 2022, Art. no. 108992.

[21] A. Banerjee, S. Ray, and M. Sarkar, "Machine learning in precision agriculture: A review of techniques and applications," *Comput. Electron. Agricult.*, vol. 199, Oct. 2023, Art. no. 107175.

[22] M. Al-Yaari, F. Kogan, and M. A. Wahid, "Automated irrigation systems using machine learning algorithms: A review," *Irrigation Sci.*, vol. 40, pp. 245–261, Feb. 2022.
[23] D. Gomez, J. E. Garcia, and S. Lopez, "ML models for pest and disease detection in crops: Current status and future directions," *Comput. Electron. Agricult.*, vol. 202, Jan. 2024, Art. no. 107340.
[24] M. U. Hassan, S. Ahmed, and A. Khan, "A machine learning-based framework for crop selection and yield optimization for smallholder farmers," *Agricult. Syst.*, vol. 209, Nov. 2023, Art. no. 103204.
[25] S. Kim, H. Park, and M. Lee, "Decision support systems for precision agriculture: Integrating soil, weather, and crop data," *Comput. Electron. Agricult.*, vol. 197, Mar. 2022, Art. no. 106929.
[26] R. Silva, J. Rodrigues, and A. C. Antunes, "Cloud-based platforms for agricultural data analysis: A review," *Comput. Electron. Agricult.*, vol. 200, p. 107287, Sep. 2023.
[27] H. Zhang, M. Zhao, and Y. Li, "Challenges and limitations of current ML models in agriculture," *Agricult. Forest Meteorol.*, vol. 310, Jul. 2022, Art. no. 108934.
[28] S. Ahmed, M. U. Hassan, and N. Aslam, "Addressing infrastructure challenges for implementing ML in agriculture in developing regions," *J. Agricult. Informat.*, vol. 14, no. 1, pp. 22–36, Mar. 2023.
[29] V. Yadav, A. Sinha, M. R. Gupta, and S. Patel, "Ethical considerations in deploying ML models in agriculture," *AI & Soc.*, vol. 38, pp. 85–101, May 2023.

ABID BADSHAH is currently pursuing the master's degree in computer science with the Department of Computer Science and IT, University of Malakand. He has contributed to the field through a conference publication, significantly advancing the development of intelligent agricultural systems and enhancing methods for image classification and recognition. He focuses on bridging theory with practice, and fostering collaboration to solve complex challenges. His research interests include machine learning and AI with a strong emphasis on crafting innovative algorithms and practical applications across different domains.

BASEM YOUSEF ALKAZEMI (Senior Member, IEEE) received the Ph.D. degree from Newcastle University, U.K. He is currently a Professor with the Department of Software Engineering, College of Computing, Umm Al-Qura University, Saudi Arabia. He held several administrative positions at the Deanship of Information Technology, Umm Al-Qura University, as the Vice Dean, and was later appointed as the Dean of the Scientific Research Deanship. He supervised several postgraduate students who conducted their research in software continuous delivery (CD/CI), BPM, the IoT, big data for retailers, machine translation, and machine learning. His main research interests include software engineering, data mining, big data, and machine learning. He served as a reviewer for a number of international conferences and reputable journals in addition to a number of national universities in Saudi Arabia.

FAKHRUD DIN received the Ph.D. degree in computer science from Universiti Malaysia Pahang, Malaysia. He is currently an Assistant Professor with the Department of Computer Science and IT, University of Malakand, Pakistan. His research interests include soft computing, machine learning, search-based software engineering, and combinatorial testing.

KAMAL Z. ZAMLI (Member, IEEE) received the degree in electrical engineering from the Worcester Polytechnic Institute, Worcester, MA, USA, in 1992, the M.Sc. degree in real-time software engineering from Universiti Teknologi Malaysia, in 2000, and the Ph.D. degree in software engineering from Newcastle University, U.K., in 2003. His research interests include (combinatorial t-way) software testing and search-based software engineering.

MUHAMMAD HARIS is currently pursuing the Ph.D. degree in computer science with the University of Technology Malaysia. He is also a Lecturer with the Department of Computer Science and Bioinformatics, Khushal Khan Khattak University. He has published extensively in high-impact journals and conferences, contributing significantly to the advancement of intelligent systems and computational biology. His work aims to bridge the gap between theoretical research and practical implementation, fostering interdisciplinary collaborations to address complex challenges in the field. His research interests include machine learning and deep learning, with a focus on developing innovative algorithms and applications in various domains.