IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 19 September 2024, accepted 7 October 2024, date of publication 11 October 2024, date of current version 29 November 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3478739*

RESEARCH ARTICLE

# An Approach for Crop Prediction in Agriculture: Integrating Genetic Algorithms and Machine Learning

**TANJIM MAHMUD[1], (Member, IEEE), NIPPON DATTA[2], RISHITA CHAKMA[1], UTPOL KANTI DAS[1], MOHAMMAD TAREK AZIZ[2], MUSADDIKUL ISLAM[3], ABUL HASNAT MUHAMMED SALIMULLAH[3], MOHAMMAD SHAHADAT HOSSAIN[4,5], (Senior Member, IEEE), AND KARL ANDERSSON[5], (Senior Member, IEEE)**

[1]Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati 4500, Bangladesh
[2]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh
[3]Department of Economics, University of Chittagong, Chittagong 4331, Bangladesh
[4]Department of Computer Science and Engineering, University of Chittagong, Chittagong 4331, Bangladesh
[5]Cybersecurity Laboratory, Luleå University of Technology, 931 87 Skellefteå, Sweden

Corresponding authors: Tanjim Mahmud (tanjim_cse@yahoo.com), Nippon Datta (22mcse031@student.cuet.ac.bd), and Karl Andersson (karl.andersson@ltu.se)

**ABSTRACT** **Objectives:** The agricultural sector in many South Asian countries, including Bangladesh and India, plays a pivotal role in the economy, with a significant portion of the population dependent on it for livelihood. However, farmers often encounter challenges such as unpredictable weather conditions, soil variability, and natural disasters like floods and erosion, leading to substantial crop losses and financial strain. Despite government subsidies, many farmers struggle to sustain their livelihoods, resulting in a decline in interest in agriculture. Our focus lies on predicting the classification of various crops, including rice, jute, maize, and others, based on a combination of soil and weather features. Soil features, including Nitrogen, Phosphorus, Potassium, and pH levels, along with weather variables such as Temperature, Humidity, and Rainfall, are utilized as inputs for the predictive model. **Methods:** In this study, we address the critical issue of crop prediction by leveraging advanced machine-learning techniques and integrating genetic algorithms into the predictive model. Our proposed approach employs a hybrid methodology, where a Genetic Algorithm is utilized to optimize the hyperparameters of the model, enhancing its performance and robustness. Specifically, we employ a Random Forest classifier, a powerful ensemble learning technique, to classify the class labels associated with 22 different types of crops. **Findings:** The model's accuracy is evaluated extensively, demonstrating a remarkable accuracy rate of 99.3%. Additionally, we utilized Local Interpretable Model-agnostic Explaination(LIME) and SHapley Additive exPlanations(SHAP) Explainable AI (XAI) methods to interpret and validate the model's predictions. **Novelty:** The study presents a unique method for crop prediction that combines machine learning (ML) with genetic algorithms (GAs). The goal of this integration is to improve crop forecast models' interpretability and accuracy. Due to the nature of local approximation LIME may yield contradictory answers. On the other hand, for sophisticated models and extensive datasets, SHAP can be computationally costly. By improving feature selection and model parameters, the integration of GAs with ML models overcomes these drawbacks and produces predictions that are more reliable and accurate. The high accuracy achieved by our system underscores its potential to mitigate crop losses and enhance agricultural productivity, thereby contributing to the sustainability and prosperity of the agricultural sector in any country.

**INDEX TERMS** Agriculture, crop prediction, genetic algorithm, fitness function, random forest, explainable AI (XAI).

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

## I. INTRODUCTION

The agricultural sector holds paramount importance in the economies of South Asian countries such as Bangladesh and India, serving as a primary source of livelihood for a substantial portion of the population. In Bangladesh, agriculture constitutes the largest employment sector, contributing 14.2 percent to the Gross domestic product (GDP) in 2017 and employing approximately 42.7 percent of the workforce [1]. Despite facing challenges like unfavorable weather conditions, Bangladesh's agriculture sector has witnessed steady growth in food grain production, attributed to improved flood control, irrigation practices, and efficient use of fertilizers [2].

Similarly, India boasts a rich agricultural history dating back to the Neolithic period and ranks second worldwide in farm outputs. With agriculture employing over 50% of the Indian workforce and contributing 20.2% to the country's GDP, it remains a pivotal sector in India's socio-economic fabric [3]. However, challenges such as erratic weather patterns, soil variability, and natural disasters pose significant threats to crop yields, leading to substantial losses and financial strain among farmers. Despite government subsidies, many farmers struggle to sustain their livelihoods, resulting in a declining interest in agriculture.

To address these challenges, our study focuses on the critical issue of crop prediction by leveraging advanced machine learning techniques and integrating genetic algorithms into the predictive model. Machine learning is recognized as a versatile approach for addressing various challenges across diverse domains [4], [5], [6], [7]. For instance, it has been employed in human disease prediction [8], [9], [10], sports analytics [11], and natural language processing tasks [12], among other applications.

Our research aims to predict the classification of various crops, including rice, jute, maize, and others, based on a combination of soil and weather features. Soil attributes such as Nitrogen, Phosphorus, Potassium, and pH levels, along with weather variables like Temperature, Humidity, and Rainfall, serve as inputs for our predictive model.

We propose a hybrid methodology where a Genetic Algorithm [13] optimizes the model's hyperparameters, enhancing its performance and robustness. Specifically, we employ a Random Forest classifier [14], a powerful ensemble learning technique, to classify the class labels associated with 22 different types of crops. Additionally, we employ LIME and SHAP methods to interpret and validate the model's predictions, providing insights into the feature importance and decision-making process of the model.

By integrating genetic algorithms and machine learning techniques, our study presents an effective approach to crop prediction, offering valuable insights for farmers and policymakers alike. The high accuracy achieved by our system underscores its potential to mitigate crop losses and enhance agricultural productivity, thereby contributing to the sustainability and prosperity of the agricultural sector in any country.

The major objectives of this study are as follows:

1) To develop a hybrid methodology integrating genetic algorithms to optimize model hyperparameters, aiming to enhance prediction accuracy and reliability in crop forecasting.
2) To investigate the effectiveness of various data mining techniques in enhancing the performance of the crop prediction system.
3) To achieve high prediction accuracy across 22 different crop types, validating the efficacy of the proposed methodology.
4) To utilize LIME and SHAP Explainable AI (XAI) methods to interpret and validate the model's predictions, providing transparency and understanding of the model's decision-making process.
5) To conduct a comparative analysis to evaluate the performance of the developed crop prediction system against existing methods.

The rest of the paper is arranged as follows. Crop recommendations and the fundamental studies conducted by agricultural researchers are covered in section II. Section III presents the proposed crop prediction methodology. Results and analysis of the prediction system are covered in section IV, and the paper's conclusion and future directions are covered in section V.

## II. RELATED WORK

As crop prediction is a critical issue in the present world so many researchers are still working on it. In this section, some of the relevant recent studies are described briefly.

Jain et al. [15] proposed a method to identify the most suitable crop(s) to maximize yield by considering the analysis of all influencing parameters. These parameters include economic, environmental, and yield-related factors. Economic factors, such as market prices and demand, significantly impact crop selection, as do environmental factors like rainfall, temperature, soil type, chemical composition, and total produce.

Rani et al. [16] proposed a machine learning-based crop selection model using weather conditions and soil parameters. Weather analysis with Long short-term memory(LSTM) Recurrent neural networks(RNN) showed root mean square error(RMSE) of 5.023% for minimum temperature, 7.28% for maximum temperature, and 8.24% for rainfall, outperforming artificial neural networks(ANN). The Random Forest Classifier achieved 97.235% accuracy in crop selection, 96.437% in predicting resource dependency, and 97.647% in determining the appropriate sowing time.

Jain and Ramesh [17] proposed a crop selection method to maximize yield based on weather and soil parameters. They also suggested the proper sowing time for suitable crops using seasonal weather forecasting. Machine learning algorithms, such as recurrent neural networks for weather prediction and the random forest classification algorithm for selecting suitable crops, were employed.

Devi et al. [18] proposed a model to select suitable crops and predict production rates using key weather parameters. The Random Forest algorithm was used for classification and prediction, and its performance was compared with the Support Vector Machine algorithm. The model achieved an average accuracy of 90%.

Rao et al. [19] aimed to identify the best crop prediction model to assist farmers in choosing crops based on climate and soil nutrients. The study compared K-Nearest Neighbor (KNN), Decision Tree, and Random Forest algorithms using Gini and Entropy criteria, finding that Random Forest achieved the highest accuracy.

Elbasi et al. [20] demonstrated the importance of integrating machine learning algorithms and IoT sensors in modern agriculture to optimize crop production and reduce waste through informed decision-making. The study identified challenges and opportunities in deploying these technologies, emphasizing the critical role of feature selection in achieving high accuracy. Experimental results showed that using features like Temperature, Humidity, pH, and Precipitation, the highest accuracy was 97.05% with Bayes Net and 97.32% with Random Forest.

Raja et al. [21] utilized efficient feature selection to preprocess raw data into a machine learning-friendly format, ensuring high precision. By using only features significantly relevant to the model's output, they reduced redundancies and enhanced accuracy. Optimal feature selection prevented unnecessary complications of the model and avoided increasing its time and space complexity. Their results indicated that ensemble techniques provided better prediction accuracy than existing classification methods.

Modi et al. [22] proposed a support vector machine(SVM) algorithm-based crop recommendation system for farmers. This system analyzed the profitability of specific crops to prevent losses and increase productivity. The SVM algorithm was used to classify various soil parameters and predict the most suitable crop. The proposed algorithm was simulated in Anaconda Navigator to analyze soil data and provide crop recommendations.

Murali wt al. [23] presented four models, including Linear Regression (LR) and Multi-Layer Perceptron (MLP), trained on historical agricultural data with climatic, soil, and geographical variables. The data was segmented seasonally to offer tailored crop suggestions. Model performance was evaluated with standard metrics, and an ensemble approach was considered for robustness. This framework provided farmers and agricultural professionals with a valuable tool for optimizing crop selection and enhancing productivity.

Patil and Mane [24] focused on studying water flow and its potential benefits and challenges, emphasizing advanced clustering systems and methods to enhance classification accuracy. The project aimed to predict crops and weather conditions (temperature, humidity, pH, rainfall) based on soil attributes (nitrogen, phosphorus, potassium), season, and region. They employed the Random Forest algorithm, selecting the configuration that provided the highest prediction accuracy, ultimately achieving an impressive 93.7% accuracy.

Nischitha et al. [25] used Decision Tree and SVM to predict rainfall and recommend seeds and fertilizers. They used three input features (pH, humidity, temperature) but noted that increasing input features could improve accuracy.

Karthikeya et al. [26] applied the K-NN algorithm with data from multiple regions but limited predictions to coconut and cocoa, without comparing other algorithms.

Cao et al. [27] demonstrated that Ensemble Learning with stacking outperforms single models, though they highlighted the need to reduce complexity to avoid overfitting.

It is important to include recent studies and discuss emerging trends and limitations in the literature review [5]. In table 1 we have included some recent studies on crop prediction using machine learning and illustrate their benefits as well as disadvantages.

## III. METHODS AND TOOLS

The main structure of the proposed system consists of 7 modules which are Data analysis, pre-processing, genetic algorithm, best recommendation, random forest algorithm, and performance evaluation. The main architecture of the proposed system is shown in Fig. 1.

### A. DATASET DESCRIPTION

The study's dataset is gathered from Indian Council of Agriculture Research(ICAR) [28]. The dataset has 7 input features of soil and environment. Table 2 shows examples of data tuples from the crop dataset. The seven input features are Nitrogen, Potassium, Phosphorus Temperature, Rainfall, pH, and Humidity. The values of Nitrogen, Phosphorus, and Potassium depended on their content in the soil, temperature measured in degrees Celsius (°C), humidity measured in percentage (%), rainfall in millimeters (mm), and pH has no unit. The output of the dataset is crop names and has a total of 22 categories. The 22 categories are Pigeon peas, Chickpea, Coffee, Pomegranate, Kidney beans, Apple, Muskmelon, Rice, Black gram, Cotton, Maize, Coconut, Grapes, Moth beans, Banana, Jute, Watermelon, Mung beans, Papaya, Lentil, Orange, and Mango.

### B. DATA ANALYSIS

In Machine Learning data analysis is a critical issue as it assists us in gaining insights into correlations and patterns among data. In this study, we visualize the distribution of input and categorical variables and the basic statistics of data.

1) Basic statistics of data

   The basic statistics represent the summary of input features in a dataset. Table 3 shows the basic statistics of the data where count represents the total number of data that exists in each column. It also shows the mean, max, std, min, and different percentages of data of input features that exist in the dataset.

**TABLE 1.** Comparative analysis of strengths and limitations of prior studies.

| Article | Methods | Advantages | Limitations |
|---|---|---|---|
| [15] | Machine Learning | predict the most suitable crops Analysis of all influencing parameters. | Didn't apply any Xplainable AI |
| [16] | RNN, RF and LSTM | Use both of weather and soil parameters. Determining appropriate sowing time. Achieved higher accuracy through the RF Model. | Didn't apply any Xplainable AI |
| [17] | RNN and RF | Maximize crop yield. Determining appropriate sowing time. RNN for weather prediction. | For massive data Random forest can take large time. |
| [18] | SVM and RF | Use of weather parameters.. achieving an average accuracy of 90%.. | Use of soil parameters along with weather parameters would be more effective. For massive data Random forest can take a long time. |
| [19] | Random Forest | They have used three different algorithms decision tree, random forest, and KNN. In particular, the study looks at the performance of Random Forests and Decision Trees using two distinct metrics: entropy and gini. | It can take a while to build many trees, particularly for massive amounts of data. |
| [20] | Naive Bayes and RF | Use of IoT sensors. Achiving higher accuracy. | More input features could impact on the model. For massive data RF model can take a long time. |
| [21] | Machine Learning | They have chooses optimal features. Implementation of ensemble techniques for the better accuracy. | They didn't use any Xplainable AI |
| [22] | SVM | The system analyzed profitability of specific crops. Use of different soil parameters. | Weather parameters also could be applied to change the impact of their model. |
| [23] | Linear Regression, Multi-layer perceptron | Use of soil, climate and geographical variables. Optimized productivity. | Didn't apply Xplainable AI |
| [24] | RF | Segmentation of data to offer tailored crop suggestions. Optimized productivity and achieved higher accuracy. | For massive data Random forest can take a long time. |
| [25] | Decision Tree and SVM | They individually predict rainfall and crops using two separate algorithms and recommend required seeds and fertilizer for agricultural land. | Used three input features (pH, humidity, and temperature) in their model. It would be more accurate and reliable if input features increased. |
| [26] | K-NN | A variety of data used in the model are collected from multiple areas. | Only two categories: Coconut and Cocoa. Other algorithms not applied except K-NN |
| [27] | Ensemble learning with stacking | shown that stacking numerous models yields higher accuracy than using only one model. | Minimize complexity in models to prevent overfitting. |

**TABLE 2.** Sample of data from dataset.

| Nitrogen | phosphorus | potassium | temperature(°C) | humidity(%) | ph | rainfall(mm) | Crop |
|---|---|---|---|---|---|---|---|
| 90 | 42 | 43 | 20.87974 | 82.00274 | 6.502985 | 202.9355 | rice |
| 67 | 60 | 25 | 24.92162 | 66.78627 | 5.750255 | 109.2162 | maize |
| 21 | 44 | 18 | 27.0691 | 86.89934 | 7.128511 | 50.46746 | mungbean |
| 105 | 14 | 50 | 26.21488 | 87.6884 | 6.419052 | 59.65591 | watermelon |
| 39 | 16 | 27 | 35.53845 | 52.94642 | 4.934965 | 91.5456 | mango |
| 86 | 40 | 39 | 25.72101 | 88.16514 | 6.20746 | 175.6087 | jute |
| 58 | 46 | 45 | 42.39413 | 90.79028 | 6.576261 | 88.46607 | papaya |
| 40 | 5 | 29 | 28.48445 | 97.76865 | 5.820979 | 160.3894 | coconut |

2) Visualize the distribution of input features
Data or information represented in a graphical or visual format is referred to as visualization [29]. We have seven different input features pH, rainfall, humidity, phosphorus, potassium, temperature, and nitrogen so we visualize the input data to seek patterns, trends, and insights within the data. The visualization of input features is shown in Figure 2.

3) Visualize the distribution of the categorical variable
Visualizing the distribution of categorical variables is a crucial step in exploratory data analysis. It assists in understanding class imbalance problems and frequency distribution in a dataset. The visualization of the categorical variable is shown in Figure 3.
Where it is shown that 22 classes of the dataset have equal numbers of data tuples in the dataset. It means that it has no class imbalance problem.

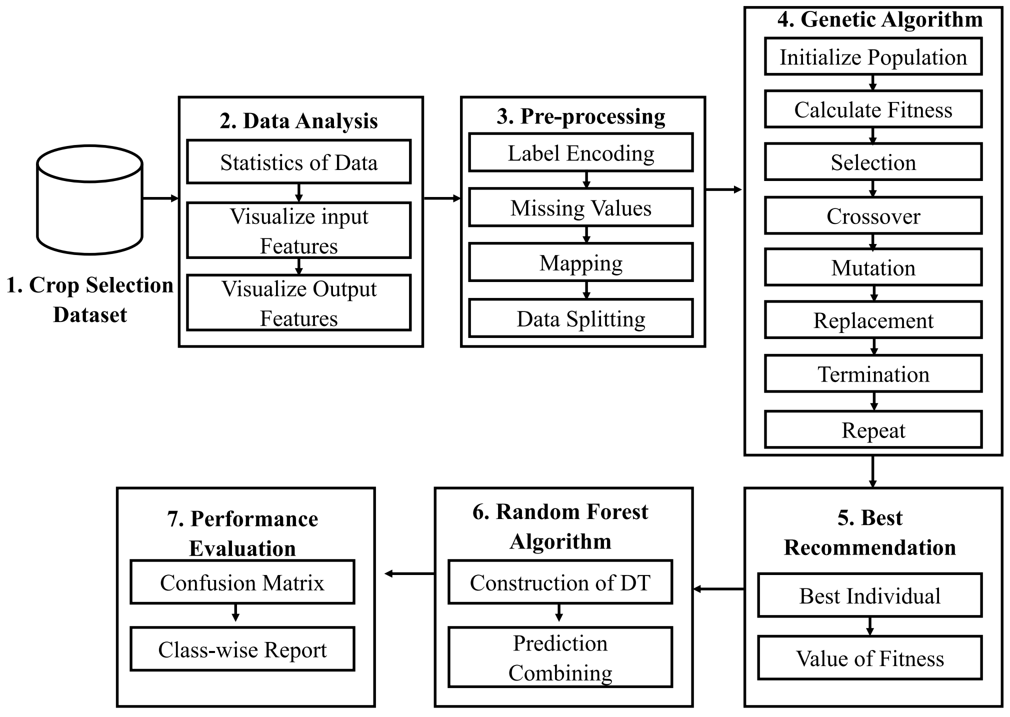


**FIGURE 1.** Architecture of the proposed system.

**TABLE 3.** Basic statistics of dataset.

| | Nitrogen | Phosphorus | Potassium | Temperature(°C) | Humidity(%) | pH | Rainfall(mm) |
|---|---|---|---|---|---|---|---|
| **Count** | 2200 | 2200 | 2200 | 2200 | 2200 | 2200 | 2200 |
| **Mean** | 50.55 | 53.36 | 48.14 | 25.61 | 71.48 | 6.46 | 103.46 |
| **Std** | 36.917 | 32.98 | 50.64 | 5.06 | 22.26 | 0.77 | 54.95 |
| **Min** | 0.00 | 5.00 | 5.00 | 8.823 | 14.25 | 3.5 | 20.21 |
| **25%** | 21.00 | 28.00 | 20.00 | 22.76 | 60.26 | 5.97 | 64.55 |
| **50%** | 37.00 | 51.00 | 32.00 | 25.59 | 80.47 | 6.42 | 94.86 |
| **75%** | 84.25 | 68.00 | 49.00 | 28.56 | 89.95 | 6.92 | 124.26 |
| **Max** | 140.00 | 145.00 | 205.00 | 43.67 | 99.98 | 9.93 | 298.56 |

### C. DATA PRE-PROCESSING

1) Label encoding
Label encoding is used for converting categorical data into numerical data [30]. In our dataset the categorical variable which refers to the name of the crop is in sting. So, we used label encoding to convert it to some numerical values.

2) Handling missing values
Missing data can affect the performance of a system. So, it is essential to remove missing data or replace it to increase the model accuracy. In this study, missing values are handled with mean imputation. Mean imputation means replacing a null value with its corresponding column mean.

3) Mapping dictionary
A mapping dictionary ensures that categorical labels are converted to numerical values accurately and consistently. A mapping dictionary makes it easier to make sure that the encoding procedure is consistent with the training data while working with fresh data. In this study, table 4 shows the mapping dictionary of the target variable.

4) Data splitting
Data splitting is nothing but dividing the dataset into training, validation, and testing subsets [31]. It is very crucial in machine learning as it guards against overfitting and guarantees that the model will generalize to new, untested data. As we have to train our model, we split the dataset into two parts where 80% of the data is used for training and the remaining 20% data are used for validation purposes.

### D. GENETIC ALGORTIHM

A genetic Algorithm is an optimization algorithm. It is based on the formula of natural selection and recommends the best individuals from a certain dataset [32], [33](See

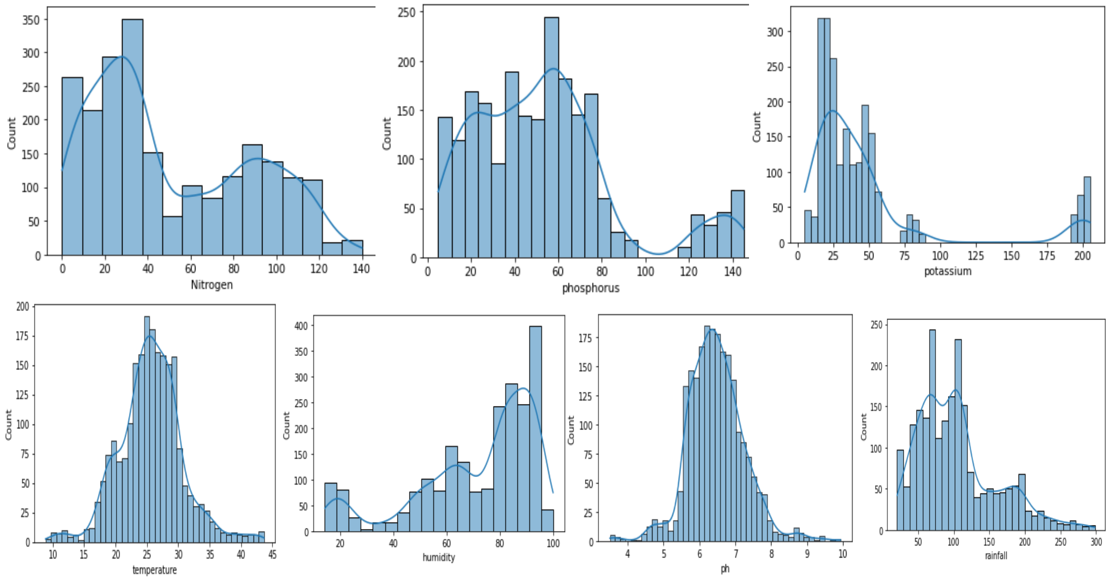


**FIGURE 2.** Visualize the distribution of input features.

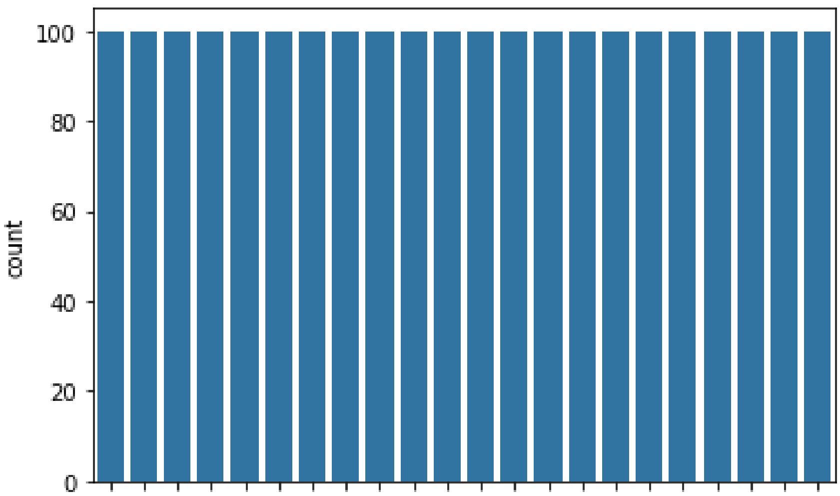


**FIGURE 3.** Visualize the distribution of the categorical variable.

**TABLE 4.** The mapping dictionary of the target variable.

| Index | Label | Index | Label | Index | Label | Index | Label |
|---|---|---|---|---|---|---|---|
| 0. | Chickpea | 6. | Pomegranate | 12. | Coffee | 18. | Muskmelon |
| 1. | Kidney beans | 7. | Rice | 13. | Cotton | 19. | Orange |
| 2. | Maize | 8. | Apple | 14. | Grapes | 20. | Papaya |
| 3. | Moth beans | 9. | Banana | 15. | Jute | 21. | Watermelon |
| 4. | Mung beans | 10. | Black gram | 16. | Lentil | | |
| 5. | Pigeon peas | 11. | Coconut | 17. | Mango | | |

Table 5). In this study, we have used a Genetic Algorithm to choose the best individuals from the dataset. The steps related to choosing the best individuals are illustrated in Table 6.

The parameter list, training time, and total memory usage of the proposed model are illustrated in table 7.

From table 7 we have seen that, the crossover probability of the model is 0.7; the mutation probability is 0.2; the

**TABLE 5.** Steps of genetic algorithm.

| Steps | Description | Purpose | Working process of GA |
|---|---|---|---|
| **Initialization** | It is the beginning stage of the Genetic Algorithm in which a starting population of individuals(solution) is generated. | Offer a wide range of probable options as a starting point. | Problem - Provide five arbitrary solutions to the traveling salesman problem. Let's assume – A, B, C, D, and E. |
| **Selection** | It is the process after initialization that Determine which individuals are the fittest for reproduction by evaluating fitness. | The purpose of this stage is to ensure a larger likelihood of reproducibility for superior solutions. | Out of the five alternatives, choose the top two shortest paths. (E.g A B) |
| **Crossover** | Mix and match pairings of chosen individuals to produce offspring. | The purpose of this stage is to produce new solutions by combining genetic information. | New route [offspring] = Combine (Route A Route B). It will produce two new routes (E.g A1 and B1) |
| **Mutation** | The next stage of the genetic algorithm is mutation in which some genes of the offspring are altered randomly. | It prevents early convergence and preserves genetic diversity. | Introduce small changes in A1 by swapping the order of two cities. |
| **Evaluation** | Check the new generation's level of fitness. | It aims to analyze the performance of the novel solutions. | Calculate the total distance for A1 and B1. |
| **Replacement** | Replace the old population with the new to create a new population. | Proceed with the evolutionary procedure. | Replace the two worst routes with A1 and B1. |
| **Termination** | It is the last stage of the genetic algorithm in which based on a predetermined circumstance, terminate the algorithm. | This stage aims to close the search and provide the best result. | If the shortest path (stopping condition) is met stop the algorithm and return the best route. |

number of generations is 20; and the number of trees is 86 with a maximum depth of 59. The training time of the model is 30 seconds whereas the memory usage is 1 GB(approximately).

Algorithm 1 represents the Genetic Algorithm for feature optimization. It initializes the population first with the value of fitness level. While a certain condition is not satisfied it selects the individuals based on their fitness value, then applies crossover and mutation respectively. After that evaluate the fitness of the new individuals. Finally, from a new population based on the best individuals.

### *E. BEST RECOMMENDATION*

The best Recommendation is the result of an individual who returns the best fitness values. After successfully compiling 20 generations, the genetic algorithm returns the best individual with a fitness value.

Table 8 shows the generation-wise number of evaluations along with the best individuals and fitness function. Where it is shown that the number of evaluations started from 50 and visited up to 96 evaluations to get the best individuals. The best individual in this study is [86.047, 59.0901] along with the best fitness function 0.993.

### *F. RANDOM FOREST ALGORITHM*

Generating numerous decision trees and aggregating their forecasts is the method for creating a Random Forest(RF) classifier, which enhances the accuracy and reliability of crop prediction models. We train the Random Forest, particularly with a portion of features picked by the genetic algorithm for a particular individual in the overall population, using specific attributes that we have chosen from the genetic algorithm.

1) Construction of decision tree(DT)
   In this study, the genetic algorithm determined the features of individuals. To form the decision tree random forest algorithm utilizes a random subset of those features along with a subset of training data. The number of decision trees and their depth depend on the value of the best individuals. In this study,
   The values of the best individual = [86.047, 59.0901]
   so, The number of n_estimators = 86
   and max_depth = 59
   Here, n_estimators define the number of trees and max_depth refers to the number or depth of the node.
2) Prediction Combining
   Based on the chosen attributes, each Random Forest decision tree forecasts the crop types during the prediction phase. The ultimate forecast of this study is frequently determined by the average or majority vote of 86 individual decision trees.

Algorithm 2 shows the execution process of the proposed system step-wise. In the very first data features and categorical variables are input to the system. Then the input data is split into training and testing sets. After that, the genetic algorithm is initialized. The size of the population, number of generations, crossover, and mutation probability need to pass through a genetic algorithm. For a certain number of generations, the genetic algorithm will do selection, crossover, and mutation. After the successful compilation of the genetic algorithm, the best solution will come out. After that, The features of the best solution will be selected and the Random Forest Algorithm will initialized. The random forest

**TABLE 6.** Result of crossover and mutation.

| | |
|---|---|
| **Initial Population** | [[96, 62], [76, 71], [30, 25], [81, 98], [33, 80], [83, 88], [98, 13], [19, 97],[68, 15], [78, 22], [79, 90], [68, 16], [97, 23], [25, 67], [95, 46], [15, 85], [90, 99], [19, 31], [27, 18], [89, 72], [31, 46], [82, 73], [47, 78], [14, 35], [49, 70], [57, 57], [99, 64], [63, 30], [88, 95], [19, 27], [20, 38], [23, 24], [83, 62], [91, 69], [58, 87], [82, 92], [39, 94], [46, 84], [77, 89], [22, 45], [87, 60], [99, 93], [42, 99], [33, 35], [14, 60], [41, 22], [89, 75], [81, 60], [73, 38], [11, 63]] |
| **Generation 1** | |
| **Fitness Value** | [(0.993), (0.993), (0.990), (0.9931), (0.9909,), (0.9931,), (0.9931), (0.990), (0.993), (0.9931), (0.993), (0.993), (0.993), (0.990), (0.993,), (0.986), (0.993), (0.990), (0.990), (0.993), (0.990), (0.993), (0.990), (0.988), (0.990), (0.990), (0.993), (0.990), (0.993), (0.990), (0.990), (0.990), (0.993), (0.993), (0.993), (0.993), (0.990), (0.990), (0.993),(0.990), (0.993), (0.993), (0.990), (0.990), (0.988), (0.990), (0.993), (0.993), (0.9931) (0.9909)] |
| **Selected Offspring** | [[82, 92], [68, 16], [77, 89], [90, 99], [81, 60], [99, 64], [23, 24], [73, 38],[76, 71], [77, 89], [58, 87], [68, 16], [82, 73], [99, 64], [78, 22], [68, 15], [83, 88], [81, 98], [87, 60], [96, 62], [81, 60], [88, 95], [76, 71], [57, 57], [31, 46], [83, 62], [99, 93], [96, 62], [23, 24], [49, 70], [76, 71], [89, 72], [82, 73], [96, 62], [77, 89], [97, 23], [78, 22], [82, 73], [20, 38], [83, 88], [81, 98], [68, 16], [89, 75], [19, 27], [91, 69], [23, 24], [78, 22], [99, 64], [23, 24], [68, 15]] |
| **After Crossover and Mutation** | [[74.58, 6.95], [75.41, 101.047], [89.13, 86.11], [77.86, 101.88], [81, 60], [99, 64], [57.55, 37.93], [38.44, 24.06], [76.18, 75.70], [78.14, 84.29], [58, 87], [68, 16], [82.04, 61.39], [98.95, 75.60], [71.69, 20.34],[74.30, 16.65], [81.74, 86.43], [82.25, 99.56], [95.94, 60.30],[87.05, 61.69], [80.61, 98.03], [88.38, 56.96], [71.09, 55.42], [61.90, 70.91], [31, 46], [83, 62], [99, 93], [96, 62], [43.09, 91.49], [28.90, 2.50], [80.18, 72.35], [84.81, 70.64], [101.10, 68.45], [76.89, 66.54], [105.97, 2.54], [68.02, 109.45], [78, 22], [82, 73], [17.88, 101.64], [85.11, 24.35], [64.04, -16.80], [84.95, 130.80], [47.83, 6.88], [60.16, 95.11], [91, 67.79], [23, 24], [85.76, 31.59], [91.23, 54.40], [25.01, 27.62], [65.98, 11.37]] |
| **Generation 2 - Generation 20** | |
| **Calculate Fitness Value, Offspring, Crossover and Mutation.** | |

algorithm will combine the selected features and the training data to generate the decision trees. After the successful compilation of that step, the performance of the system is evaluated.

## IV. RESULT AND DISCUSSION

In this section, we used our proposed system to predict crop from used input and compare our model with some other existing methods that have been developed by several researchers.

### A. PERFORMANCE EVALUATION

The classification report is a representation of the result of system performance. In this section, the result of the proposed system has been measured on a confusion matrix and categorical class-wise report of classification.

1) Confusion Matrix
A confusion matrix is nothing more than a matrix that is utilized to represent and illustrate the classification performance. Acquiring true positive, false positive, true negative, and false negative rates is aided by it. Figure 4 shows the confusion matrix of classification. From the confusion matrix, it is shown that most of the validation data are correctly classified except for a small number of validation data.
2) Class-wise Report
The final report of classification is based on the precision, recall, f1-score, and support of each class. The classification report of crop prediction is shown in Table 9. We showed here individual values of precision, recall, and f1-score for 22 categories of our dataset. For almost every class we have achieved a better outcome.

**TABLE 7.** List of parameters, training time, and memory usage of the proposed model.

| Parameter Name | Values |
|---|---|
| Crossover Probability | 0.7 |
| Mutation Probability | 0.2 |
| Generations (Steps number) | 20 |
| Number of Trees | 86 |
| Maximum Depth | 59 |
| Size of population | 50 |
| Number of Records | 2200 |
| Training Time | Training Time = GA Optimization + ML Model Training<br>= Each generation time * generation number + ML model training<br>= (1 * 20) seconds + 10 seconds<br>= 30 seconds (approximately) |
| Total Memory Usage | Total Memory Usage = GA Optimization + ML Model Training + SHAP Computation<br>= 200 MB + 500 MB + 300 MB<br>= 1000 MB ( 1GB approximately) |

**Algorithm 1** Genetic Algorithm for Feature Optimization

**Step 1:** Initialize the population of individuals.
**Step 2:** Determine each individual's fitness level.
**while** the termination criterion is not met **do**
  **Step 3:** Select individuals based on fitness level.
  **Step 4:** Apply crossover with probability $p_c$.
  **Step 5:** Apply mutation with probability $p_m$.
  **Step 6:** Evaluate the fitness of the new individuals.
  **Step 7:** Form a new population by selecting the best individuals.
**end while**
=0

```
[[23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 26  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 14  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 23  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 21  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 19  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  0 23  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0 19  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 14  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 23  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 23  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 23  0  0]
 [ 0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0 17  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 19]]
```

**FIGURE 4.** Confusion matrix of classification.

The average values for recall, precision, and f1-score here 99.3%.

3) Generation-wise Performance
Generation-wise performance based on Genetic Algorithm and Random Forest Algorithm is shown in Table 10. We trained our Genetic Algorithm for 20 generations and achieved the best individual [86.047, 59.0901] along with a fitness value of 0.993. After that, we pass the values to the RF classifier and RF takes the features of that individual and finally gives

**Algorithm 2** Crop Prediction Algorithm

**Require:** Data features ($X$), Categorical Variable ($y$)
**Ensure:** RF Model with GA for Crop Prediction
**Step 1:** Split Dataset
($X\_train$, $X\_test$, $y\_train$, $y\_test$) ← train_test_split($X$, $y$)
**Step 2:** Initialize Genetic Algorithm
Define `n_estimators`, `max_depth`
Set fitness function
Set up the initial population
`GeneticAlgorithm(pop_size, gen, crossProb, mutProb)`
**for** $i \leftarrow 1$ to gen **do**
Find the individual's level of fitness
Select the individuals to propagate
Employ crossover and mutation
**end for**
**Step 3:** Get the best solution and the best fitness
`Selected_features = SelectFeatures(best_solution)`
**Step 4:** Initialize the RF Classifier
`Final_Model = TrainRandomForest`(X_train[`selected_features`], y_train)
**Step 5:** Evaluate performance
Make predictions on $X\_test$
Calculate Accuracy
Output the Model performance metrics
**Output:** RF Model with GA for Crop Prediction
=0

**TABLE 8.** Generation, evaluation number and best individual result.

| Gen | nevals | Gen | nevals | Gen | nevals | Gen | nevals |
|---|---|---|---|---|---|---|---|
| 0 | 50 | 5 | 88 | 10 | 93 | 15 | 93 |
| 1 | 93 | 6 | 91 | 11 | 87 | 16 | 90 |
| 2 | 91 | 7 | 89 | 12 | 95 | 17 | 86 |
| 3 | 92 | 8 | 89 | 13 | 96 | 18 | 93 |
| 4 | 91 | 9 | 91 | 14 | 94 | 19 | 94 |
| **Best Individual** | | | **[86.047, 59.0901]** | | | | |
| **Best Fitness** | | | **0.993** | | | | |

us a model with better accuracy. In this table, we show a report of what happens if the number of generations we changed several times. It is shown that for the number of 20 generations, we have achieved a better result.

4) ROC Curve
A graphics representation called the Receiver Operating Characteristic (ROC) curve is used to evaluate the classification model's performance. It shows how a classifier may be used to diagnose problems at different threshold levels. Figure 5 shows the ROC Curve of our proposed classifier.

The ROC curve works based on the True positive rate (TPR) and the False Positive Rate (FPR).

$$TPR = TP/(TP + FN) \tag{1}$$

$$FPR = FP/(FP + TN) \tag{2}$$

**TABLE 9.** Final report of classification.

| Index | Class | Precision | Recall | F1-score |
|---|---|---|---|---|
| 0 | Chickpea | 1.00 | 1.00 | 1.00 |
| 1 | Kidney beans | 1.00 | 1.00 | 1.00 |
| 2 | Maize | 1.00 | 1.00 | 1.00 |
| 3 | Moth beans | 1.00 | 1.00 | 1.00 |
| 4 | Mung beans | 1.00 | 1.00 | 1.00 |
| 5 | Pigeon peas | 1.00 | 1.00 | 1.00 |
| 6 | Pomegranate | 1.00 | 1.00 | 1.00 |
| 7 | Rice | 1.00 | 1.00 | 0.96 |
| 8 | Apple | 0.92 | 1.00 | 1.00 |
| 9 | Banana | 1.00 | 1.00 | 0.96 |
| 10 | Black gram | 0.92 | 1.00 | 1.00 |
| 11 | Coconut | 1.00 | 1.00 | 1.00 |
| 12 | Coffee | 1.00 | 1.00 | 1.00 |
| 13 | Cotton | 1.00 | 0.96 | 0.98 |
| 14 | Grapes | 1.00 | 1.00 | 1.00 |
| 15 | Jute | 1.00 | 1.00 | 1.00 |
| 16 | Lentil | 1.00 | 1.00 | 1.00 |
| 17 | Mango | 1.00 | 1.00 | 1.00 |
| 18 | Muskmelon | 1.00 | 1.00 | 1.00 |
| 19 | Orange | 1.00 | 1.00 | 1.00 |
| 20 | Papaya | 1.00 | 0.89 | 0.94 |
| 21 | Watermelon | 1.00 | 1.00 | 1.00 |
| **Average** | | **0.993** | **0.993** | **0.993** |

where, TP = True Positive,
FN = False Negative,
FP = False Positive and,
TN = True Negative.

TPR is plotted on the y-axis whereas FPR is plotted on the x-axis. The ROC Curve (area = 1.00) for a class like

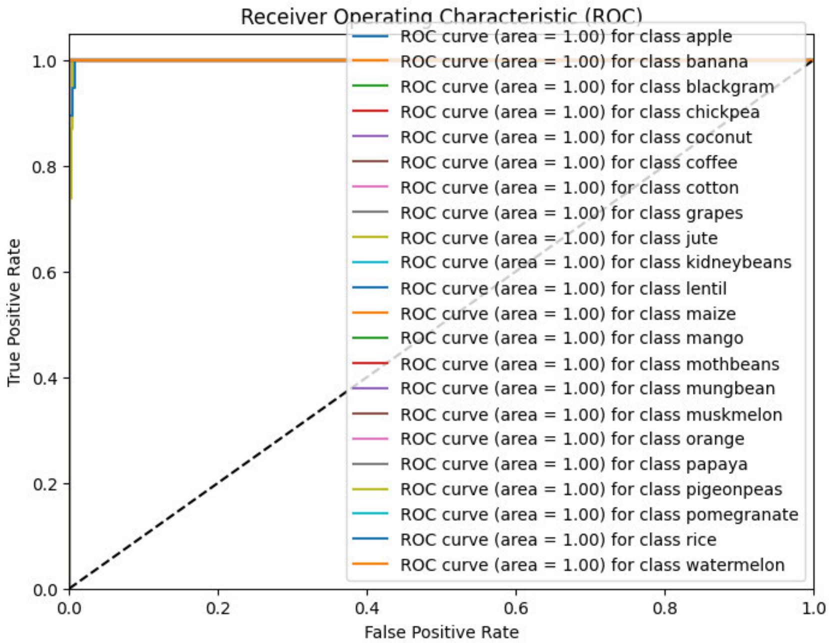


**FIGURE 5.** ROC Curve of proposed classifier.

**TABLE 10.** Generation-wise performance of GA Vs RF.

| | Genetic Algorithm | | Random Forest | |
|---|---|---|---|---|
| Generation Number | Best Individuals | Fitness | Precision | Recall |
| 05 | [75.657,31.553] | 0.913 | 90.3% | 91.2% |
| 10 | [66.672,31.658] | 0.927 | 92.67% | 92% |
| 15 | [65.996, 44.874] | 0.957 | 95.6% | 97.5% |
| 20 | [86.047, 59.0901] | 0.993 | 99.3% | 99.3% |
| 25 | [80.315,58.091] | 0.976 | 98.7% | 97.7% |

apple means that the model is exhibiting excellent accuracy and precision in its predictions for this particular class and is working perfectly in discriminating apples against non-apples. It produces a perfect ROC curve that crosses the upper-left corner of the plot, with a TPR of 1.00 (100%) and an FPR of 0.00 (0%).

### B. INTERPRETING PREDICTION: UNDERSTANDING SHAP

SHAP refers to SHapley Additive exPlanations. SHAP frequently uses plots such as summary plots, force plots, and dependency plots to illustrate feature attributions and show the relative contributions of each feature to a specific prediction. It assists users to understand the impact of each feature and how they interact with each other to influence the prediction [34].

Figure 6 shows the impact of features on model magnitude with SHAP. From Figure 6, it is clear that for the class of apple, phosphorus has the most impact on the prediction. The rest of the features impact as follows potassium, humidity, rainfall, temperature, nitrogen, ph. Similarly, for the class of papaya, the feature impacts are as follows - humidity, rainfall, potassium, phosphorus, nitrogen, pH, and temperature. The impact of features for the last class of rice is rainfall, humidity, phosphorus, nitrogen, potassium, temperature, ph respectively.

### C. EXPLAINING PREDICTIONS WITH LIME

LIME stands for Local Interpretable Model-agnostic Explanation which attempts to provide case-by-case explanations for machine learning models' predictions, offering insights into the reasoning behind a given prediction's creation for a certain instance [35]. LIME concentrates on explaining the decision-making process surrounding a particular instance rather than providing a global explanation of the model.

Figure 7 shows the explanation of predictions using LIME. From Figure 7 it is clear that the classes apple, papaya, and rice have prediction probabilities around 1. The crop selection based on the input features can contribute to maximizing the crop yield. If the values of rainfall $<= 64.09$; nitrogen $> 85.00$; humidity $> 89.90$; phosphorus $<= 28$; Potassium $> 31$; temperature $> 28.52$; $5.96 < \text{pH} <= 6.93$; then this environment will be good for selecting the apple to maximize yield. Similarly, If the values of rainfall $<= 94.30$; nitrogen $> 37$; humidity $> 89.90$; phosphorus $> 51$; Potassium $> 48$; pH $> 6.93$; then this environment will be good for selecting the Papaya to maximize yield. The third row in the figure shows that if the values of phosphorus $> 68$; potassium $> 48$; humidity $> 89.90$; rainfall $> 94.30$; pH $<= 5.96$; Nitrogen $> 20$ and temperature $> 22.76$ then this environment will be good for selecting the Rice to maximize yield. In this way, we can fix the soil and weather environment and select the appropriate crop to maximize the crop yield.

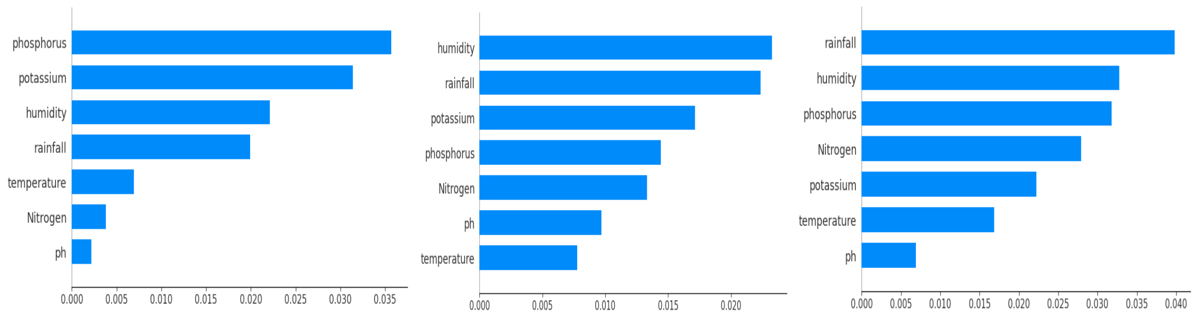


**FIGURE 6.** **Interpreting prediction: understanding SHAP for (Apple, Papaya, Rice).**

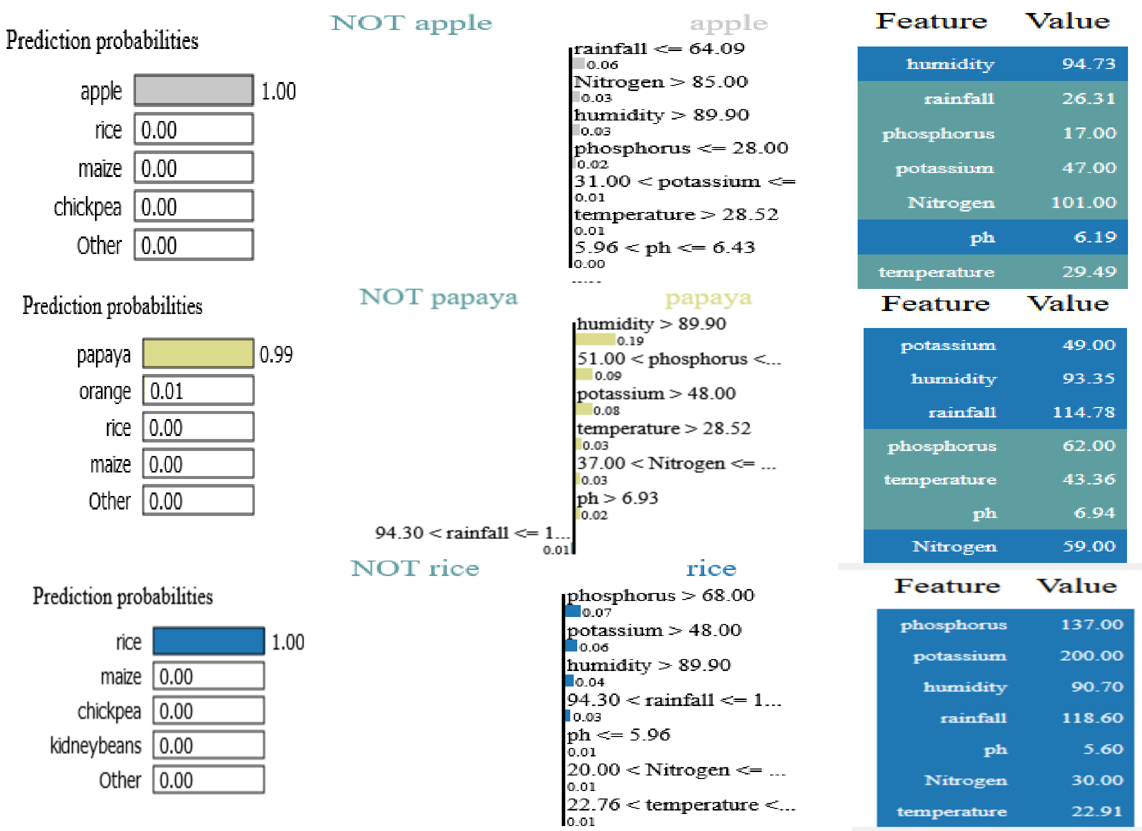


**FIGURE 7.** **Local interpretability with LIME.**

### *D. TEST WITH USER INPUT*

The system can suggest the best crop based on the seven input features nitrogen, pH, rainfall, potassium, phosphorus, humidity, and temperature. Figure 8 shows the result of prediction where input is taken from the user. Where we have shown four individual user inputs case-1, case-2, case-3, and case-4. Each of the case inputs consists of different input values and finally based on the input values the system provides the name of the predicted crop. For case-1 the predicted crop is banana, for case-2 predicted crop is Rice, case-3 the predicted crop is Lentil, for case-4 the predicted crop is Jute.

### *E. COMPARISON WITH EXISTING METHODS*

Our study introduces an innovative approach by integrating Genetic Algorithms with a Random Forest (RF) Classifier and utilizing SHAP and LIME for Explainable AI (XAI). This method has achieved an accuracy of 99.3%, significantly improving existing methods in the literature. For instance, K-Nearest Neighbor (KNN), as seen in [19], achieved an accuracy of 99.32%, comparable to our study(See Table 11). However, KNN lacks the interpretability provided by XAI methods such as SHAP and LIME. Similarly, the MRFE feature selection method combined with RF in [21] reached an accuracy of 97.29%. While effective, this approach does

| Crop selection using user input | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Case -1 input** | | **Case-2 input** | | **Case-3 input** | | **Case-4 input** | |
| Nitrogen | 12 | Nitrogen | 91 | Nitrogen | 10 | Nitrogen | 84 |
| Phosphorus | 61 | Phosphorus | 35 | Phosphorus | 75 | Phosphorus | 36 |
| Potassium | 19 | Potassium | 39 | Potassium | 17 | Potassium | 42 |
| Temperature | 19.33 | Temperature | 23.22 | Temperature | 18.1 | Temperature | 25.5 |
| Humidity | 24 | Humidity | 81.12 | Humidity | 68 | Humidity | 81.2 |
| pH | 5 | pH | 6.5 | pH | 7.1 | pH | 6.8 |
| Rainfall | 68 | Rainfall | 206 | Rainfall | 52 | Rainfall | 170 |
| **Predicted Crop** | **Banana** | **Predicted Crop** | **Rice** | **Predicted Crop** | **Lentil** | **Predicted Crop** | **Jute** |

**FIGURE 8.** User input test result using proposed method.

not match our method's performance and does not offer model interpretability. The Support Vector Machine (SVM) applied in [22] achieved 97% accuracy, again falling short of our results and lacking the insight into feature importance that SHAP and LIME provide. Additionally, the combination of Multi-Layer Perceptron (MLP) and Linear Regression (LR) classifiers in [23] resulted in an accuracy of 81%, which is significantly lower than our integrated approach.

Several studies have utilized Random Forest as the primary classifier, achieving varying degrees of success. For example, [24] reported an accuracy of 93.7%, [16] reached 97.235%, [18] obtained 90%, and [20] achieved 97.32%. Gosai et al. [36] combined Logistic Regression and RF, achieving 95.22% and 99% accuracy, respectively. ML models (RF, Naive Bayes) in Anguraj et al. [37] obtained an accuracy of 96.89%. RF and Decision Trees (DT), as applied by Thilakarathne et al. [38], achieved accuracies of 97.18% and 86.64%, respectively. Hossain et al. [39] utilized ML (CRS) models and achieved 99% accuracy. While these results demonstrate the robustness of the RF algorithm, our study's accuracy of 99.3% surpasses these findings. This highlights the efficacy of combining Genetic Algorithms with RF. Moreover, the inclusion of SHAP and LIME enhances our model's transparency and interpretability, which is not addressed in these previous works. By achieving higher accuracy and providing clear insights into the model's decision-making process, our approach represents a significant advancement in crop prediction technology, offering farmers a valuable tool for informed decision-making.

### F. PRACTICAL IMPLICATIONS

**Use-case:** Crop prediction in smart agriculture.

**Implementation:** The dataset of historical crop prediction, weather, soil quality, and other pertinent variables is subjected to the integrated GA-ML technique. The crop prediction model produced by the GA is extremely accurate due to its optimization of the feature set and ML model parameters.

**Outcomes:** When compared to conventional approaches, the innovative methodology produces greater prediction accuracy. Additionally, it offers comprehensible insights into the variables that most affect crop yields.

**Benefits:** These forecasts can help farmers make well-informed choices on planting, irrigation, and resource allocation, which will eventually result in higher yields and lower waste.

The model of the proposed crop recommendation system is available publicly at GitHub repository for Crop Recommendation Model. Anyone can use this model from this source.

### G. ECONOMIC IMPLICATIONS

The integration of genetic algorithms and machine learning techniques for crop prediction in agriculture presents a promising avenue to address the challenges faced by the agricultural sector.

Despite government subsidies, the sustainability of agriculture remains a pressing concern, especially with changing climatic patterns and increasing water and land scarcity [40], [41]. The economic impact of this research is profound, as it aims to mitigate crop losses and enhance agricultural productivity. By accurately predicting crop outcomes using machine learning techniques and genetic algorithms, farmers can make informed decisions about crop selection, resource allocation, and risk management. With an impressive accuracy rate of 99.3%, the proposed predictive model offers valuable insights for farmers and policymakers alike, enabling them to optimize agricultural practices and maximize efficiency [41].

Furthermore, improved crop prediction can lead to better market planning and reduced food waste, as farmers can align their production with market demands more effectively [42]. This alignment not only benefits farmers by increasing their

**TABLE 11. Comparison with state-of-art.**

| Reference | Method Name | XAI | Accuracy |
|---|---|---|---|
| Rao et al. [19] | KNN | Not Apply | 99.32% |
| Raja et al. [21] | MRFE with the RF | Not Apply | 97.29% |
| Modi et al. [22] | SVM | Not Apply | 97% |
| Murali et al. [23] | MLP and LR classifier | Not Apply | 81% |
| Patil et al. [24] | RF | Not Apply | 93.7% |
| Rani et al. [16] | RF | Not Apply | 97.235% |
| Devi et al. [18] | RF | Not Apply | 90% |
| Elbasi et al. [20] | RF | Not Apply | 97.32% |
| Gosai et al. [36] | Logistic Regression, RF | Not Apply | 95.22%, 99% |
| Anguraj et al. [37] | ML model(RF, Naive Bayes) | Not Apply | 96.89% |
| Thilakaranthne et al. [38] | RF, DT | Not Apply | 97.18%, 86.64% |
| Hossain et al. [39] | ML(CRS) | Not Apply | 99% |
| **Our Study** | **Genetic Algorithm, RF Classifier** | **SHAP and LIME** | **99.3%** |

income and reducing financial uncertainty but also supports national food security goals by ensuring a stable supply of agricultural products [43]. The ripple effects extend to the broader economy, where increased agricultural productivity can stimulate growth in related industries, such as food processing, distribution, and retail [44]. Overall, this study has the potential to revolutionize the agriculture sector, drive economic growth, and enhance the livelihoods of millions of farmers.

## V. CONCLUSION AND FUTURE WORK

### A. CONCLUSION

In this study, we presented a hybrid strategy that integrates machine learning and genetic algorithms to improve crop prediction in agriculture. Given the critical role of agriculture in employing a large percentage of the workforce in South Asian countries, enhancing productivity and efficiency in this sector is vital for both individual livelihoods and national GDP growth.

Our system demonstrates an impressive accuracy rate of 99.3% in predicting the outcomes for 22 different crops based on input data from weather and soil parameters. By combining the Random Forest Classifier with Genetic Algorithms, we have developed a robust predictive model that surpasses the accuracy of previous methods.

The practical implications of our system extend beyond academic research. By providing predictive insights, our system can serve as a valuable tool for farmers, enabling them to make informed decisions before commencing cultivation activities. This, in turn, can help farmers optimize resource allocation, reduce risks, and ultimately maximize crop yields and profitability.

### B. FUTURE WORK

In future work, we plan to further enhance the impact of our approach on sustainability by integrating climate change scenarios into our crop prediction model. This will enable us to anticipate long-term environmental impacts and provide guidance to farmers on adopting more sustainable agricultural practices. The incorporation of real-time data from IoT sensors and other emerging agricultural technologies will allow for more immediate, data-driven decision-making, which can help farmers optimize resource use and reduce waste. Additionally, focusing on renewable energy sources for powering IoT devices, and promoting environmentally friendly farming techniques will contribute toward the broader goal of agricultural sustainability. Creating a user-friendly mobile application will also help ensure that these insights are accessible to all farmers, regardless of their technological background, further promoting sustainable farming across diverse communities.

## ACKNOWLEDGMENT

The authors would like to thank the Head of the Laboratory for providing invaluable guidance, continuous evaluation, and steadfast support throughout the research process and also would like to thank the dedicated laboratory members whose active participation significantly aided in the collection and refinement of the dataset. Furthermore, they utilized an AI tool to improve the manuscript's quality by refining the grammar and adopting a more technical writing style. The research was carried out with the Artificial Intelligence Laboratory (AI Laboratory), Rangamati Science and Technology University, Bangladesh.

## REFERENCES

[1] *The World Factbook*. Accessed: Aug. 10, 2023. [Online]. Available: https://www.cia.gov/the-world-factbook/countries/bangladesh/
[2] *Bureau of South and Central Asian Affairs*. Accessed: Aug. 10, 2023. [Online]. Available: https://2009-2017.state.gov/r/pa/ei/bgn/3452.htm
[3] *India Economic Survey 2018*. Accessed: Aug. 10, 2023. [Online]. Available: https://www.financialexpress.com/budget/india-economic-survey-2018-for-farmers-agriculture-gdp-msp-1034266/
[4] M. H. Imam, N. Nahar, R. Bhowmik, S. B. S. Omit, T. Mahmud, M. S. Hossain, and K. Andersson, "A transfer learning-based framework: MobileNet-SVM for efficient tomato leaf disease classification," in *Proc. 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2024, pp. 693–698.
[5] Y. Akkem, S. K. Biswas, and A. Varanasi, "Smart farming using artificial intelligence: A review," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105899.
[6] T. Mahmud, K. Barua, A. Barua, N. Basnin, S. Das, M. S. Hossain, and K. Andersson, "Explainable AI for tomato leaf disease detection: Insights into model interpretability," in *Proc. 26th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2023, pp. 1–6.

[7] T. Mahmud, M. Ptaszynski, and F. Masui, "Exhaustive study into machine learning and deep learning methods for multilingual cyberbullying detection in Bangla and Chittagonian texts," *Electronics*, vol. 13, no. 9, p. 1677, Apr. 2024.

[8] A. S. Khan, F. T. Khan, T. Mahmud, S. K. Khan, N. Sharmen, M. S. Hossain, and K. Andersson, "Integrating BERT embeddings with SVM for prostate cancer prediction," in *Proc. 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2024, pp. 1–6.

[9] S. U. Habiba, F. Tasnim, M. S. H. Chowdhury, M. K. Islam, L. Nahar, T. Mahmud, M. S. Kaiser, M. S. Hossain, and K. Andersson, "Early prediction of chronic kidney disease using machine learning algorithms with feature selection techniques," in *Proc. Int. Conf. Appl. Intell. Inform.* Cham, Switzerland: Springer, 2023, pp. 224–242.

[10] A. D. Bappy, T. Mahmud, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A BERT-based chatbot to support cancer treatment follow-up," in *Proc. Int. Conf. Appl. Intell. Inform.* Cham, Switzerland: Springer, 2023, pp. 47–64.

[11] S. Das, T. Mahmud, D. Islam, M. Begum, A. Barua, M. T. Aziz, E. N. Showan, L. Dey, and E. Chakma, "Deep transfer learning-based foot no-ball detection in live cricket match," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, Jan. 2023, Art. no. 2398121.

[12] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Inf. Process. Manage.*, vol. 60, no. 5, Sep. 2023, Art. no. 103454.

[13] L. Bi and G. Hu, "A genetic algorithm-assisted deep learning approach for crop yield prediction," *Soft Comput.*, vol. 25, no. 16, pp. 10617–10628, Aug. 2021.

[14] V. Geetha, A. Punitha, M. Abarna, M. Akshaya, S. Illakiya, and A. P. Janani, "An effective crop prediction using random forest algorithm," in *Proc. Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Jul. 2020, pp. 1–5.

[15] N. Jain, A. Kumar, S. Garud, V. Pradhan, and P. Kulkarni, "Crop selection method based on various environmental factors using machine learning," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 551–558, 2017.

[16] S. Rani, A. K. Mishra, A. Kataria, S. Mallik, and H. Qin, "Machine learning-based optimal crop selection system in smart agriculture," *Sci. Rep.*, vol. 13, no. 1, p. 15997, Sep. 2023.

[17] S. Jain and D. Ramesh, "Machine learning convergence for weather based crop selection," in *Proc. IEEE Int. Students' Conf. Elect., Electron. Comput. Sci. (SCEECS)*, Feb. 2020, pp. 1–6.

[18] M. A. Devi, D. Suresh, D. Jeyakumar, D. Swamydoss, and M. L. Florence, "Agriculture crop selection and yield prediction using machine learning algorithms," in *Proc. 2nd Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, Feb. 2022, pp. 510–517.

[19] M. S. Rao, A. Singh, N. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *J. Phys., Conf. Ser.*, vol. 2161, no. 1, 2022, Art. no. 012033.

[20] E. Elbasi, C. Zaki, A. E. Topcu, W. Abdelbaki, A. I. Zreikat, E. Cina, A. Shdefat, and L. Saker, "Crop prediction model using machine learning algorithms," *Appl. Sci.*, vol. 13, no. 16, p. 9288, Aug. 2023.

[21] S. P. Raja, B. Sawicka, Z. Stamenkovic, and G. Mariammal, "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," *IEEE Access*, vol. 10, pp. 23625–23641, 2022.

[22] D. Modi, A. V. Sutagundar, V. Yalavigi, and A. Aravatagimath, "Crop recommendation using machine learning algorithm," in *Proc. 5th Int. Conf. Inf. Syst. Comput. Netw. (ISCON)*, Oct. 2021, pp. 1–5.

[23] M. Suguna K, P. Murali, P. Ayyasamy, and O. Obuli, "Crop recommendation system using machine learning algorithm," *Int. Res. J. Adv. Eng. Hub*, vol. 2, no. 5, pp. 1237–1242, May 2024.

[24] N. A. Patil and S. Mane, "Crop recommendation in precision agriculture using machine learning techniques," *Res. Square*, 2024, doi: 10.21203/rs.3.rs-3834326/v1.

[25] K. Nischitha, D. Vishwakarma, M. N. Ashwini, and M. Manjuraju, "Crop prediction using machine learning approaches," *Int. J. Eng. Res. Technol.*, vol. 9, no. 8, pp. 23–26, 2020.

[26] H. Karthikeya, K. Sudarshan, and D. S. Shetty, "Prediction of agricultural crops using KNN algorithm," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 5, pp. 1422–1424, 2020.

[27] H. Cao, Y. Gu, J. Fang, Y. Hu, W. Ding, H. He, and G. Chen, "Application of stacking ensemble learning model in quantitative analysis of biomaterial activity," *Microchem. J.*, vol. 183, Dec. 2022, Art. no. 108075.

[28] *Indian Council of Agricultural Research*. Accessed: Aug. 10, 2023. [Online]. Available: https://icar.org.in

[29] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.

[30] R. Guedrez, O. Dugeon, S. Lahoud, and G. Texier, "Label encoding algorithm for MPLS segment routing," in *Proc. IEEE 15th Int. Symp. Netw. Comput. Appl. (NCA)*, Oct. 2016, pp. 113–117.

[31] D. R. Cox, "A note on data-splitting for the evaluation of significance levels," *Biometrika*, vol. 62, no. 2, pp. 441–444, Aug. 1975.

[32] T. V. Mathew, "Genetic algorithm," IIT, Bombay, India, 2012, vol. 53.

[33] M. T. Aziz, T. Mahmud, M. K. Uddin, S. N. Hossain, N. Datta, S. Akther, M. S. Hossain, and K. Andersson, "Machine learning-driven job recommendations: Harnessing genetic algorithms," in *Proc. Int. Congr. Inf. Commun. Technol.* Cham, Switzerland: Springer, 2024, pp. 471–480.

[34] S. Das and S. Chatterjee, "Explainable machine learning for crop recommendation from agriculture sensor data-a new paradigm," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, vol. 7, Jul. 2023, pp. 1–7.

[35] P. N. Srinivasu, M. F. Ijaz, and M. Woźniak, "XAI-driven model for crop recommender system for use in precision agriculture," *Comput. Intell.*, vol. 40, no. 1, Feb. 2024, Art. no. e12629.

[36] D. Gosai, C. Raval, R. Nayak, H. Jayswal, and A. Patel, "Crop recommendation system using machine learning," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 7, no. 3, pp. 558–569, 2021.

[37] K. Anguraj, B. Thiyaneswaran, G. Megashree, J. P. Shri, S. Navya, and J. Jayanthi, "Crop recommendation on analyzing soil using machine learning," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 6, pp. 1784–1791, 2021.

[38] N. N. Thilakarathne, M. S. A. Bakar, P. E. Abas, and H. Yassin, "A cloud enabled crop recommendation platform for machine learning-driven precision farming," *Sensors*, vol. 22, no. 16, p. 6299, Aug. 2022.

[39] S. Hossain, N. N. I. Prova, M. R. Sadik, and A. A. Maruf, "Enhancing crop management: Ensemble machine learning for real-time crop recommendation system from sensor data," in *Proc. Int. Conf. Smart Syst. Appl. Electr. Sci. (ICSSES)*, vol. 8, May 2024, pp. 1–6.

[40] J. N. Pretty, J. I. L. Morison, and R. E. Hine, "Reducing food poverty by increasing agricultural sustainability in developing countries," *Agricult., Ecosyst. Environ.*, vol. 95, no. 1, pp. 217–234, Apr. 2003.

[41] J. Beddington, "Food security: Contributions from science to a new and greener revolution," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 365, no. 1537, pp. 61–71, Jan. 2010.

[42] S. P. Dash and K. Priyashantha, "For sustainable farming in India: A data analytics perspective," *Digit. Agricult. Ecosyst., Revolutionary Advancements Agricult.*, pp. 307–318, Apr. 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394242962.ch17

[43] W. Zhu, J. Chen, X. Liang, D. Li, and K. Chen, "Government regulations, benefit perceptions, and safe production behaviors of family farms—A survey based on Jiangxi Province, China," *J. Cleaner Prod.*, vol. 450, Apr. 2024, Art. no. 141824.

[44] C.-C. Lee, M. Zeng, and K. Luo, "The impact of urbanization on food security in China," *Int. Rev. Econ. Finance*, vol. 93, pp. 1159–1175, Jun. 2024.

**TANJIM MAHMUD** (Member, IEEE) received the B.Sc. degree in engineering and in computer science and the M.Sc. degree in engineering from the University of Chittagong, Bangladesh. He is currently pursuing the Ph.D. degree in engineering and in natural language processing with the Kitami Institute of Technology, Japan. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Rangamati Science and Technology University, Bangladesh. His research interests include artificial intelligence, natural language processing, AI in healthcare, machine learning, and image processing.

**NIPPON DATTA** is currently pursuing the degree in computer science and engineering with Chittagong University of Engineering and Technology, Chittagong, Bangladesh. His research interests include machine learning, deep learning, and natural language processing.

**RISHITA CHAKMA** received the B.Sc. and M.Sc. degrees in engineering from Chittagong University of Engineering and Technology, Chittagong, Bangladesh. She is currently an Assistant Professor in computer science and engineering with the Rangamati Science and Technology University, Bangladesh. Her research interests include machine learning, deep learning, and the IoT.

**UTPOL KANTI DAS** received the B.Sc. and M.Sc. degrees in engineering from Rangamati Science and Technology University, Bangladesh. His research interests include machine learning, deep learning, and image processing.

**MOHAMMAD TAREK AZIZ** is currently pursuing the degree in computer science and engineering with Chittagong University of Engineering and Technology, Chittagong, Bangladesh. His research interests include machine learning and natural language processing.

**MUSADDIKUL ISLAM** is currently pursuing the degree in economics with the University of Chittagong, Bangladesh.

**ABUL HASNAT MUHAMMED SALIMULLAH** received the B.S.S. (Hons.) and M.S.S. degrees in economics from the University of Chittagong, Bangladesh, in 2002, and the M.A. degree in financial economics and the Ph.D. degree in econometrics and quantitative economics from Middle Tennessee State University (MTSU), USA, in 2014 and 2018, respectively. He is currently a Professor in economics with the University of Chittagong. He has published twelve articles in peer-reviewed journals, both nationally and internationally. His research interests include labor economics, dynamic macroeconomics, financial economics, and environmental economics.

**MOHAMMAD SHAHADAT HOSSAIN** (Senior Member, IEEE) received the M.Phil. and Ph.D. degrees in computation from the Institute of Science and Technology (UMIST), University of Manchester, U.K., in 1999 and 2002, respectively. He is currently a Professor in computer science and engineering with the University of Chittagong, Bangladesh, and a Visiting Professor with Luleå University of Technology, Sweden. His current research interests include e-government, the modeling of risks and uncertainties using evolutionary computing techniques, investigation of pragmatic software development tools and methods, information systems in general, and expert systems.

**KARL ANDERSSON** (Senior Member, IEEE) received the M.Sc. degree in computer science and technology from the Royal Institute of Technology, Stockholm, Sweden, and the Ph.D. degree in mobile systems from Luleå University of Technology, Sweden. After pursuing postdoctoral research with the Internet Real-Time Laboratory, Columbia University, New York City, USA, and the National Institute of Information and Communications Technology, Tokyo, Japan, he is currently a Professor with the Cybersecurity Laboratory with Luleå University of Technology. His research interests include green and mobile computing, the Internet of Things, cloud technologies, and information security.