Crop Yeild Prediction Using Machine Learning And Deep Learning

1st Md. Asifu Islam
Computer Science and Engineering
BRAC University

Dhaka, Bangladesh
md.asiful.islam@g.bracu.ac.bd

2nd Mahmudul Hasan Mitul
Computer Science and Engineering
BRAC University

Dhaka, Bangladesh
mahmudul.hasan.mitul@g.bracu.ac.bd

3rd Mahedi Hasan Shanto
Computer Science and Engineering
BRAC University

Dhaka, Bangladesh
mahmudul.hasan.mitul@g.bracu.ac.bd

4th Annajiat Alim Rasel Senior Lecturer
Computer Science and Engineering
BRAC University

Dhaka, Bangladesh
annajiat@bracu.ac.bd

***Abstract*- Bangladesh is an agrarian country. Though, a substantial portion of our economy and workforce depends directly or indirectly on agriculture. However, due to climate change, floods, insufficient incentives, and less grist our farmers are getting demotivated in farming. As a result, more and more farmers are leaving the agriculture sector every year and this can cause devastating effects for Bangladesh. Moreover, there is little or no research on improving Bangladesh agriculture using cutting-edge machine learning techniques. So, this research works on Crop yield prediction Using Machine learning and deep learning. This work explores the different state-of-the-art machine learning and deep learning techniques and relevant dataset to develop an effective Crop yield prediction system for Bangladeshi farmers. So that our farmers can decide which crop to cultivate for gaining the maximum yield by following our prediction system.**

Keywords: Crop yield prediction, Artificial Intelligence, Machine Learning, Deep Learning, Neural Network, Classification Models, Recurrent Neural Network

I. INTRODUCTION

Bangladesh has been an agricultural country since its dawn. However, with the rise of industrialization, agriculture is on the decline. But still, agriculture accounts for 42.7 % of the country's employment and 14-2 % of the GDP [2] which is one of the highest.

Moreover, in the past decade Bangladesh has become self-sufficient in most of the essential crops. It was possible due to extensive research and the green revolution. As a result, the production of the essential crops have increased 5-10 fold [10].

However, one thing that has not gone up is the condition of the farmers. It is still the same as the 90's. Their land has not increased nor their wealth. As a result, more and more farmers are shifting from farming to other sectors. According to a survey by the daily star, the number of farmers has fallen from around 73 percent of the population in 1983/84 to nearly 42.7 % by 2022 [12]. As a developing nation, it is an alarming signal. Because if we lose self sufficiency in the crop sector, our food import cost will increase. Moreover, it might as well cause famines.

The main reason behind farmers leaving agriculture is not getting enough price for their crops. For example, rice takes 3-4 months to grow and 1 kg of rice will require 3,000 liters of water to grow, not to mention the price of fertilizer and insecticides. Furthermore, growing rice is a labor intensive task too. But farmers only get 36 tk for one Kg of rice [11]. That might be ok for large farms which account for only 2 percent of the farms in Bangladesh. On the other hand, 98 percent of the farmers have limited farmland and labor. So, with lesser pay it's not possible for them to keep up with the price hikes. Moreover, there are middlemen involved in the trading of crops. The middlemans take a large chunk of the profit of the farmers. Lastly, there is labor shortage in the rural areas. So, more and more farmers are leaving their forefathers profession. Moreover, Bangladesh is one of the top producers of rice, potato, jute and fish [1], still it has very little farm productivity.

To solve the issue, farming productivity must be increased so that farmers do not have to leave their profession and can get more pay from their limited farmland with less labor. From our research, we have found, precision agriculture can solve the productivity issue and increase the profit of our farmers.

In order to support management decisions based on estimated variability for improved resource use efficiency, productivity, quality, and profitability, precision agriculture is a management strategy that collects, processes, and analyzes temporal, spatial, and individual data with the help

of different state of the art machine learning models. It then combines this data with other information and finds out which crop is better for production at a certain area. However, agriculture is different in every country and continent. As a result, not all precision agriculture techniques work for our country. Furthermore, it is crucial that the predictions made by the precision agriculture techniques have to be precise and error free. Otherwise, it might result in heavy capital loss and labor waste.

Various studies are already done on precision agriculture to achieve a precise and effective system for crop yield recommendation. One such method is developing a custom dataset and testing it on different machine learning and deep learning models. This study develops a system that uses a custom dataset tailored with Bangladeshi crop data and uses this dataset to train and test state of the art machine learning classifiers and deep learning models and compare the results.

II. LITERATURE REVIEW

Babu et al. [4] explains the specifications and preparation required to design a system for precision farming. It studies the foundations of precision farming in detail. In order to exert some control over unpredictability, this research proposes a system that applies Precision Agriculture techniques to small, open farms at the level of the solo farmer and crop. The model's overall goal is to use the most accessible technology, such as SMS and email, to provide consultancy services to even the tiniest farmer at the level of the smallest plot of crops. This model was created to account for the situation in Kerala State where the typical farm size is smaller than any other state. As a result, this model can be applied elsewhere in the world with some minor upgrades. Khedr et. al [6] aims to find a solution to Egypt's food security issue. It suggests a structure that would forecast production and import for that specific year. WEKA builds the prediction using Artificial Neural Networks and Multi-Layer Perceptrons. As a result, the researchers would be able to monitor the quantity of cultivation, import, requirement and availability at the conclusion of the procedure. Therefore, it would be easier to decide whether or not food needs to be imported further. Ahamed et al. [5] outlines multiple classification techniques for the data set on liver disease. Because accuracy depends on the dataset and the learning process, the study underlines the importance for accuracy. These disorders were categorized using classifiers such VFI, ZeroR, ANN, and Naive Bayes. Then the efficacy and error rates were compared. The models' performance was evaluated in terms of precision and computational efficiency. All classifiers, with the exception of na¨ıve bayes, had improved predictive performance, it was determined. The proposed algorithms with the highest accuracy are multilayer perceptrons. Yang et al. [3] attempts to find a solution to the critical ensemble learning problem of classifier selection. There has been a method developed for choosing the best models from a pool of classifiers. Their method seeks to increase performance and accuracy. Based on categorization precition and accuracy, the SAD approach was suggested. The relationship between the most accurate and relevant classifiers is discovered using Q statistics. The ensemble was created by combining the classifiers that weren't selected. The goal of this measure is to ensure that the ensemble performs better and is more diverse. There were other ways found, including the No selection algorithm, the Selection by Accuracy, Selection by accuracy and Diversity algorithms. Finally, it is implied that SAD functions more effectively than others. Kumar et al. [7] demonstrates the significance of crop choice, and elements influencing crop choice are examined, such as production rate, market price, and governmental policy. This study suggests a crop selection method (CSM), which fixes the crop selection issue and raises the crop's net yield rate. It proposes that a season's worth of crops be chosen while taking the weather, crop type, soil type, and water density into account. The accuracy of CSM depends on the expected value of key factors. Consequently, a prediction approach with increased performance and accuracy must be included. Savla et al. [9] explores assortment algorithms in detail, including how well they function in predicting yield in precision husbandry. These algorithms are used to predict the production of a soybean crop using data that has been gathered over a number of years. In this study, Random Forest, SVM, Bayes, Neural Network, Bagging and REPTree were employed as the yield prediction techniques. The result reached in the end is that, of the algorithms mentioned above, bagging has the lowest error deviation with a mean absolute error of 18985, making it the best approach for yield prediction. The soil datasets in the study of Paul et al. [8] are examined, and a projected categorization is made. It is determined that the crop yield is a classification rule from the expected soil category. For predicting agricultural yield, naive Bayes and KNN algorithms are employed. The indicated future study involves developing effective models utilizing other classification methods, such as principal component analysis and support vector machines.After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

III. TERMINOLOGIES

For the preliminary analysis we have chosen SVM and KNN. The model descriptions are given below.

A. SVM

Support Vector Machines is a collection of supervised learning techniques used for outlier detection, regression analysis, and classification. It serves as a classifier. Here, we tend to depict each piece of information as an area in an n-dimensional house where n is the range of possibilities, with each feature's value being the value of a particular coordinate. It serves as a classification method. In this algorithmic rule, every information item is often plotted as a distance in an n-dimensional home (where n is the range of possibilities you have) with each feature's value being the value of a chosen coordinate. A separating hyperplane accurately delimits the discriminative classifier known as a Support Vector Machine (SVM). To put it another way, the algorithmic rule is applied to labeled coaching data (supervised learning). It produces the optimum hyperplane for classifying fresh samples.

B. KNN

K-NN A data mining method known as the k-nearest neighbor (k-NN) method is regarded as one of the top five methods. The value an observation has for a particular characteristic is taken to be its coordinate in that dimension, resulting in a set of points in space. We do this by treating each feature in our training set as a separate dimension in some space. The distance between two points in this space can then be determined by some appropriate metric by comparing how similar two points are to one another. The algorithm chooses the k th point from the training set that is comparable enough to be taken into account when deciding the class to forecast for a new observation. The k data points that are the closest to the new observation, and to select the most prevalent class among them. It is known as the k Nearest Neighbors algorithm for this reason. The implementation of algorithm can be noted as below.

We load the data first. Initialize K to the specified number of neighbors after that. We determined the distance between each example in the data and the query example using the data. An ordered collection's index and distance are added. The results are then sorted by distances into an ordered collection of distances and indices, from smallest to largest (ascending order). The labels of the first K entries are then obtained by selecting them from the sorted collection as shown in the following example. Return the mean of the K labels if regression Return the K labels' mode in the categorization case.

In future work we plan to introduce our hybrid model for the prediction.

IV. METHODOLOGIES

This flowchart describes our work procedures step by step

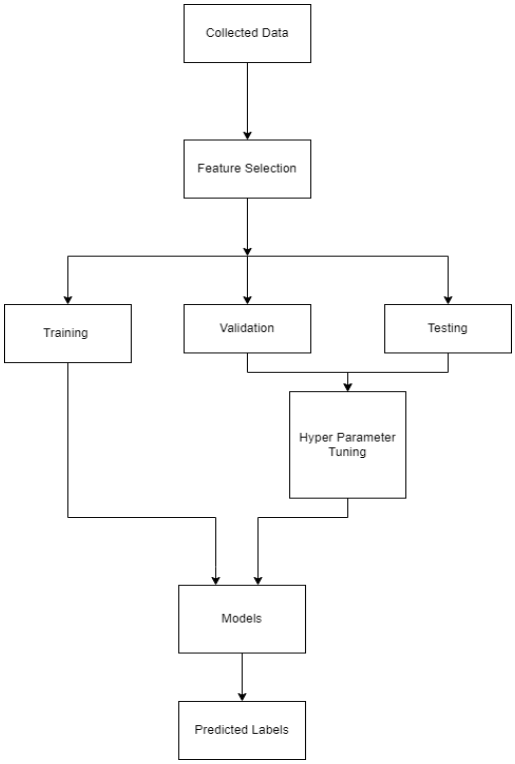

Figure 3.1: Working Procedure

The data collection process was one of the toughest for us till now. Because for our prediction system to be precise and efficient we would need accurate data. Moreover, we need specific soil attributes to train our models which is hard to get because the authority keeps the soil data in analog form. But we need the data in digital CSV format. Furthermore, to access the soil data we would need special permission from the Agriculture Bureau. So, our data collection process was slow and we could not manage a substantial amount of data.
However, we managed to collect soil specific attribute data from Bangladesh Agriculture Research Institute. In addition, general crop research data were also collected from Bangladesh Agriculture University respiratory. After collecting the data on different crops we have labeled each data point.
The fruits and crops examined in our models are coconut, rice, apple, jute, pigeon peas, pomegranate, papaya, orange, banana, maize, mango. The number of data points for each crop available in the dataset is shown in the below table.

Crop and Fruit	Data Points
Coconut	62
Rice	59
Apple	56
Jute	54
Pigeon peas	51
Orange	40
Banana	32
Papaya	32
Maize	17
Mango	5
Pomegranate	41
Total	449

Table 3.1: Data Points on different crops of Bangladesh

After labeling the data we have looked for certain errors in the dataset. We have found errors like duplicate data, symbols, punctuations, wrong input etc. After that we started to clean the data and make it more efficient for our models to understand.

First, we removed all the symbols and punctuation marks. Then we eliminated all the wrong inputs. Lastly, we removed all the duplicate data.
The data cleaning process was done with the help of library functions.

For our models to be more efficient and provide better recommendations we selected certain features of our dataset. The crop data in Table-3.1 contains the optimal level of nitrogen, phosphorus, pH, temperature, rainfall and humidity needed to harvest the crops in order to achieve high yield. The below table provides an initial idea of the data

	Nitrogen	Phosphorus	Potassium	Temperature	Humidity	pH	Rainfall
Count	449	449	499	499	499	499	499
Mean	44.75	53.97	64.29	25.99	82.42	6.32	153.95
Std	32.22	37.24	56.33	3.93	14.90	0.64	50.29
Min	0	5.0	5.0	20.03	30.40	4.56	100.19
Max	119.0	145.00	205.00	38.41	99.96	7.87	298.56

Table 3.2: Different Features of The dataset

From our initial research on different crops we have found that Nitrogen helps the growth of leaves on the plant, where Phosphorus accounts for root growth and flower and fruit development. On the other hand, Potassium helps the overall functions of the plant perform correctly. Moreover, different plants require different optimal temperatures, humidity, pH, and rainfall. This is why we have chosen these features for our analysis.

All the above criteria in soil are required for crops to grow perfectly and maximize the profit of the farmers.

V. RESULT AND ANALYSIS

After cleaning the data and selecting certain features from the data we have fed the data to our machine learning algorithms. For training and testing we have split the data into 80:20 ratio, here we have used 80% of our data for training our models and 20% for testing.
After training SVM and KNN, we have tested the accuracy of the models for predicting crop yield and the results were satisfactory. The below graphs show a glimpse of our analysis.

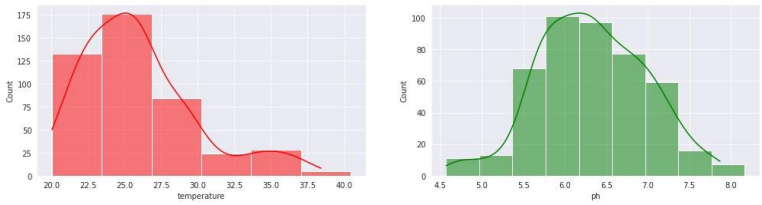

Figure 4.1: Predicted Temperature and pH level for selected crops

The above graph indicates that our models show a high number of the crops in the dataset needing an optimal temperature of 25 degree Celsius and a pH level of 6-6.5 to grow.

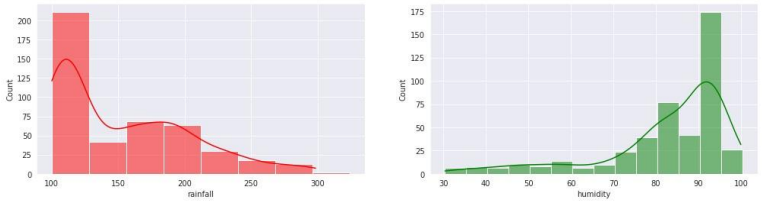

Figure 4.2: Predicted Rainfall and humidity level for selected crops

In figure 4.2, we find that most of our crops prefer a rainfall of 100mm and humidity between 90-100 to grow. Now, let's have a look at the predictions of our models.

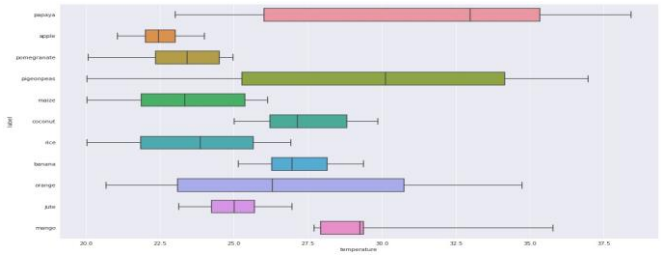

Figure 4.3: Predicted optimal Temperature for selected crops

In figure 4.3 we can see the predicted optimal temperature of our models for different crops.

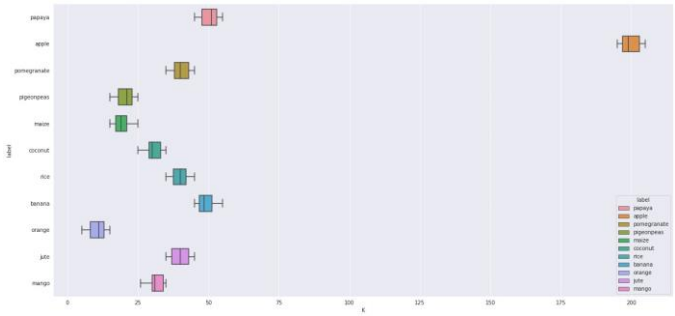

Figure 4.4: Predicted optimal potassium level in soil for selected crops

Figure 4.4 demonstrates the predicted optimal potassium level for our selected crops.

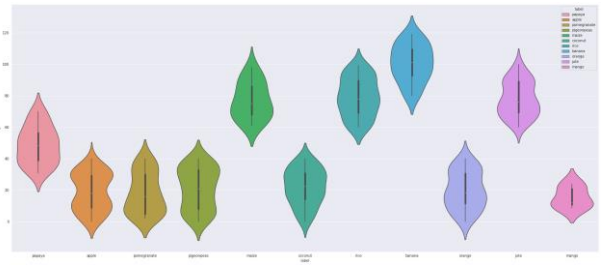

Figure 4.5: Predicted optimal Nitrogen level in soil for selected crops

In figure 4.5 we can see the predicted required nitrogen level for the crops in the dataset.

In figure 4.6 we can see the optimal rainfall and humidity of different crops in our dataset.

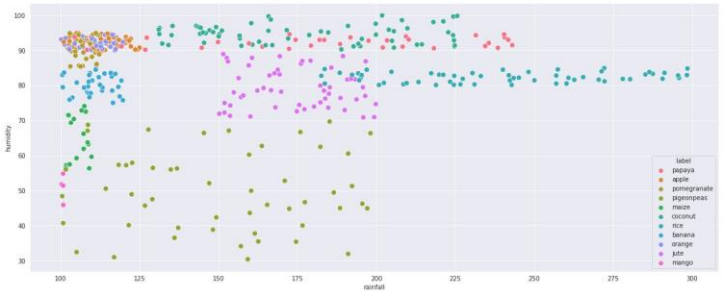

Figure 4.6: Predicted optimal Rainfall and Humidity for selected crops

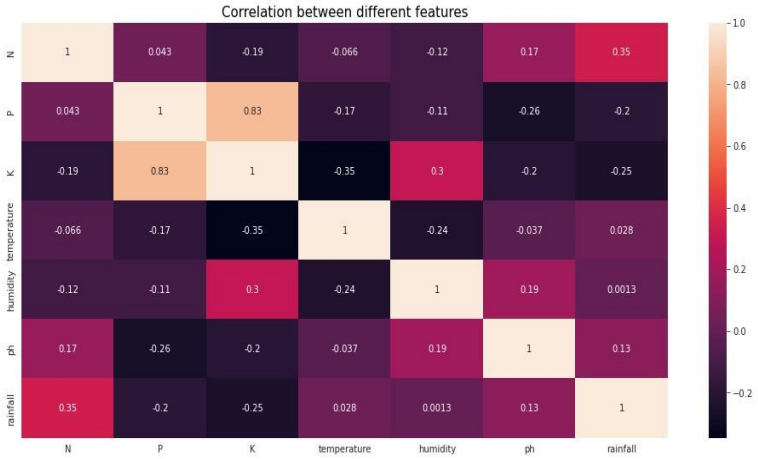

Figure 4.7: Correlation between different features

Figure 4.7 describes the correlation between different features of our dataset.

The below table describes that in a supervised environment, our chosen model's gained accuracy of selecting the accurate crop.

Model	Accuracy
SVM	0.9667
KNN	0.9778

Table 4.1: Results of different models

Meaning, with our dataset SVM has gained 96.67% at predicting the right crop. On the other hand, KNN has achieved a 97.78% accuracy at predicting the accurate crop. With more data we can achieve more efficient and precise results.

Figure 4.8 demonstrates the confusion matrix for the SVM classifier. Meaning, it shows the summary of correct and incorrect predictions by the model.

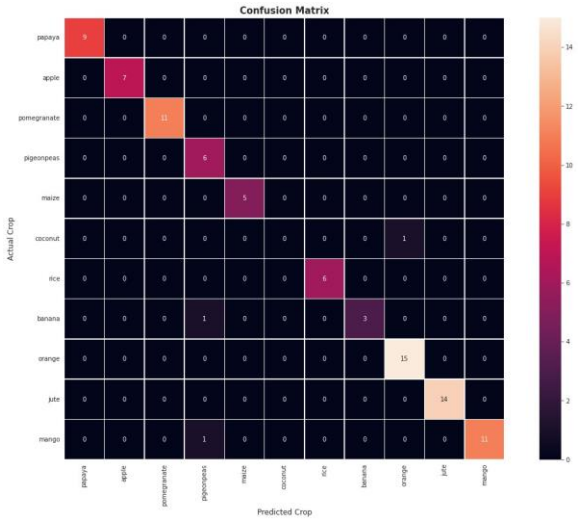

Figure 4.8: Confusion Matrix of SVM

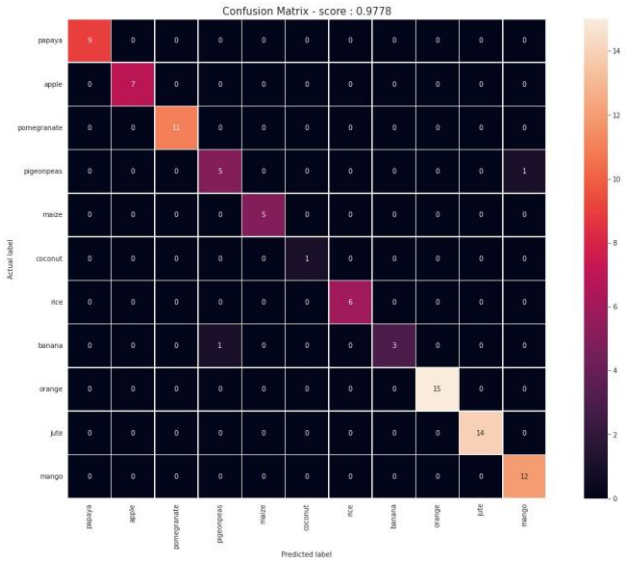

Figure 4.9: Confusion Matrix of KNN

Figure 4.9 demonstrates the confusion matrix for the KNN. Meaning, it shows the summary of correct and incorrect predictions by the KNN model.

VI. CONCLUSION

In conclusion, building an effective prediction system for crop yields in Bangladesh using Machine Learning and deep learning approaches is one of the most difficult tasks. As our country has not seen notable association between agriculture and state of the art machine learning classifiers before. So, our big challenge is the successful integration of agriculture with Machine Learning and deep learning. However, we take these challenges as our inspiration to develop the most efficient prediction system for crop yields using different Machine Learning approaches. Though this study is still at its beginning level. But our step might be the start of something which will revolutionize the entire agricultural sector of Bangladesh.

REFERENCES

[1] A. E. Hossain and M. A. Hassan, “Agriculture in bangladesh,” 1991.

[2] S. Scarpetta, A. Bassanini, D. Pilat, and P. Schreyer, “Economic growth in the oecd area: Recent trends at the aggregate and sectoral level,” *Available at SSRN 241568*, 2000.

[3] L. Yang, “Classifiers selection for ensemble learning based on accuracy and diversity,” *Procedia Engineering*, vol. 15, pp. 4266–4270, 2011.

[4] S. Babu, “A software model for precision agriculture for small and marginal farmers,” in *2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS)*, IEEE, 2013, pp. 352–355.

[5] A. M. S. Ahamed, N. T. Mahmood, N. Hossain, *et al.*, “Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh,” in *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, IEEE, 2015, pp. 1–6.

[6] A. E. Khedr, M. Kadry, and G. Walid, “Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector applied case on food security information center ministry of agriculture, egypt,” *Procedia Computer Science*, vol. 65, pp. 633–642, 2015.

[7] R. Kumar, M. Singh, P. Kumar, and J. Singh, “Crop selection method to maximize crop yield rate using machine learning technique,” in *2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, IEEE, 2015, pp. 138–145.

[8] M. Paul, S. K. Vishwakarma, and A. Verma, “Analysis of soil behaviour and prediction of crop yield using data mining approach,” in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, 2015, pp. 766–771.

[9] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada, and S. Bhardwaj, “Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture,” in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, IEEE, 2015, pp. 1–7.

[10] S. Sharmin, S. Mitra, and M. Rashid, “Production, yield and area growth of major winter vegetables of bangladesh,” *J Bangladesh Agril Univ*, vol. 16, no. 3, pp. 492–502, 2018.

[11] N. Banu, “Governance for covid-19 in bangladesh,” in *5th World Congress on Disaster Management: Volume I*, Taylor & Francis, 2022.