# First Draft: Multimodal Analysis of Soil Reports and Crop Text Descriptions for Crop Yield Prediction

## 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 [1] 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 essential crops has increased 5-10 fold [2].

However, one thing that has not gone up is the condition of the farmers. It is still the same as in 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 percent by 2022 [3]. 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 [4]. 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 a labor shortage in rural areas. So, more and more farmers are leaving their forefather's professions. Moreover, Bangladesh is one of the top producers of rice, potato, jute, and fish [5], but 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 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 using NLP and testing it on different machine learning models. This study develops a system that uses a custom dataset tailored to Bangladeshi crop data and uses this dataset to train and test state-of-the-art NLP methods and machine learning classifiers to compare the results.

## Literature Survey

Babu et al. [6] 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 in any other state. As a result, this model can be applied elsewhere in the world with some minor upgrades.

Khedr et. al [7] aim 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. [8] outline 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 of 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 naïve bayes, had improved predictive performance, it was determined. The proposed algorithms with the highest accuracy are multilayer perceptrons.

Yang et al. [9] attempt 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 precision 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. [10] demonstrate 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. [11] explore 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. [12] 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.

In their study, Dahikar et al. [13] explored  different feed-forward back propagation artificial neural networks for crop yield prediction. They created their own dataset. Their dataset contained features like phosphate, PH, potassium, nitrogen, organic carbon, calcium, magnesium, sulfur, manganese, copper, iron, depth, of soil along with the type of soil and climate contains like humidity, temperature, and rainfall. After that, they trained their custom ANN model with the dataset features and ANN with zero, one, and two hidden layers and found satisfactory results.

In their work,  Panda et al.[14] used NDVI, GVI, SAVI, and PVI measurements from aerial images to develop models for predicting corn crop yield before harvest. The authors used data mining techniques, satellite, and drone images to create the dataset. Then they transformed the image data into numbers using different encoding methods. Lastly, they fed their data into different BPNN crop yield prediction models. From their research on corn, they have bought extraordinary results and hope to apply this method to other crops as well.

Natural Language Processing (NLP) has gained prominence in the agricultural domain, offering innovative solutions for analyzing textual data related to crop cultivation. Several studies highlight the application of NLP techniques in agriculture:

Alvarez-Lopez et al. [15] conducted an extensive review of text mining applications in agriculture, emphasizing the extraction of valuable information from unstructured textual data. The study explores the potential of NLP in uncovering patterns and trends in agricultural reports, aligning with the goals of our research.

In a study by Wang et al. [16], NLP techniques were employed for extracting crop-related information from textual descriptions. The authors utilized advanced NLP algorithms for named entity recognition and relationship extraction, contributing insights into effective methods for handling diverse textual data in the context of crop cultivation.

Sentiment analysis, a subset of NLP, has been applied to assess the sentiments expressed in agricultural texts. Research by Singh et al. [17] demonstrates the utility of sentiment analysis in understanding the emotions conveyed in farmer narratives. This perspective is relevant to our study, where sentiment analysis can aid in gauging farmers' perceptions and experiences.

# Data Collection

Our data collection process commenced with the acquisition of relevant datasets from Bangladesh Agricultural University. This includes soil attributes data and textual descriptions of crop practices. The inclusion of NLP-focused data is crucial for a multimodal analysis approach.

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 BAU authority keeps the soil data in analog form. But we need the data in digital CSV format. So, our data collection process was slow and we could not manage a substantial amount of data.

Moreover, as the weather in Bangladesh is not diverse and farmers tend to follow other farmers to grow a certain type of crop in one district only. For example, Rajshahi and Chapainawabganj have similar weather and most mangos are produced there. As a result, the data points collected on a single crop are in close proximity.

However, general crop research data were also collected from Bangladesh Agriculture University. After collecting the data on different crops we labeled each data point.

# Methodology

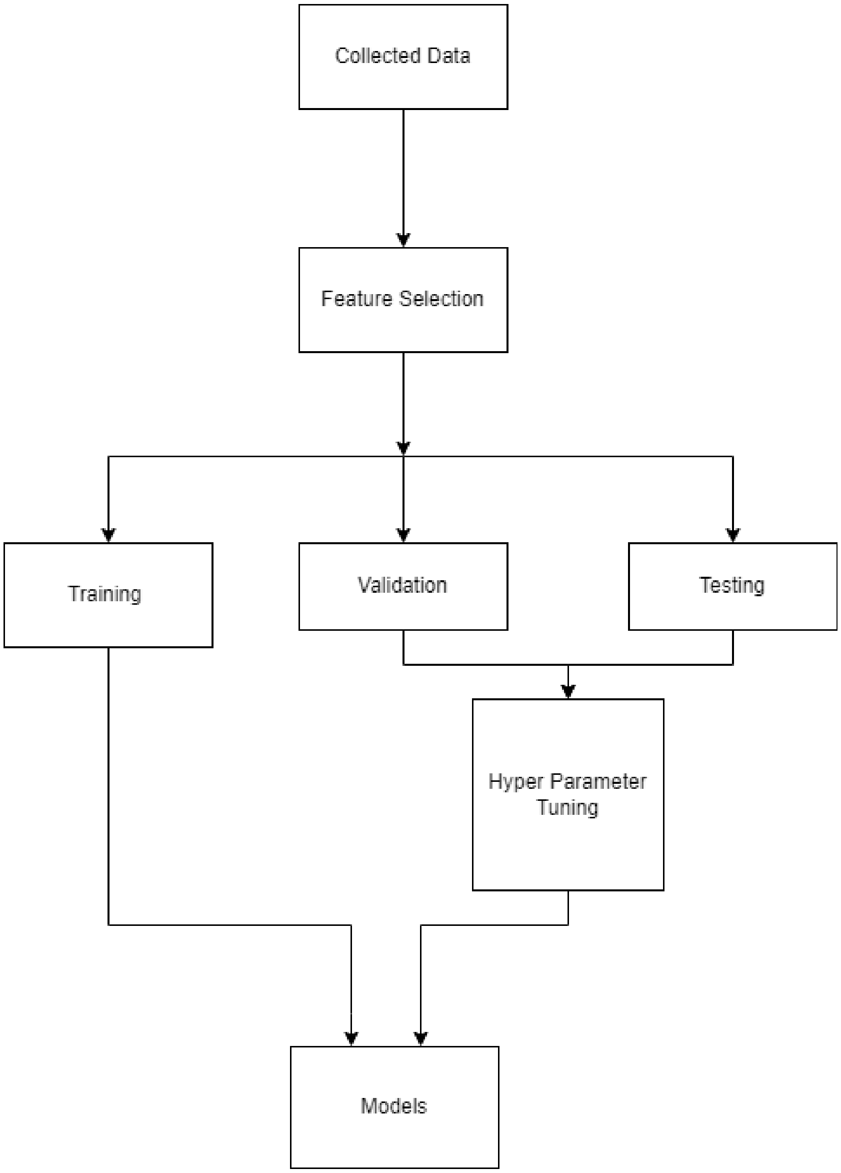
### Initial Research Exploration

Our research began with a thorough exploration of diverse NLP approaches applicable to crop yield prediction. We investigated existing methodologies, including sentiment analysis, named entity recognition (NER), and text mining techniques, to extract meaningful insights from textual descriptions of crop practices.

### Methodology Development

Building on the insights gained from the initial NLP exploration, we formulated a methodology that seamlessly integrates NLP techniques with traditional crop yield prediction models. This methodology serves as the framework for leveraging NLP insights for improved crop yield prediction accuracy.

This flowchart describes our work after NLP approach procedures step by step



In the initial research period we have explored different approaches of crop yield prediction. According to our research, we developed our own methodology for this research. Then we started collecting data from Bangladesh Agricultural University.

Once the dataset was ready, we initiated the data cleaning process and then started the feature selection. After feature selection, we have splitted the data into training, validation and testing.Lastly, we trained and tested our models and bought predicted results.