**GODFREY OKOYE UNIVERSITY, ENUGU STATE**

FACULTY OF NATURAL SCIENCES AND ENVIRONMENTAL STUDIES

DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS

# ENHANCED FARM YIELD PREDICTION SYSTEM

BY

**ANAYO NNENNA GRACE**

GOU/U22/CSC/792

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF

BACHELOR OF SCIENCE (B.SC.) IN COMPUTER SCIENCE

SUPERVISOR: DR. FRANK OKEBANAMA

MAY, 2026

## APPROVAL PAGE

This research titled "ENHANCED FARM YIELD PREDICTION SYSTEM" has been assessed and approved by the Department of Computer Science and Mathematics, Faculty of Computing and Information Technology, Godfrey Okoye University, Enugu State, Nigeria.

**Dr. Frank Okebanama**

Supervisor

Signature: ________________________ Date: ________________

External Examiner

Signature: ________________________ Date: ________________

HOD, Department of Computer Science and Mathematics

Signature: ________________________ Date: ________________

Dean, Faculty of Computing and Information Technology

Signature: ________________________ Date: ________________

## CERTIFICATION

I certify that I completed the research included herein and that it was primarily based on my own observations and investigation. I will fully accept the University's decision if I am found to have presented work that is not my own.

______________________________

**ANAYO NNENNA GRACE**

GOU/U22/CSC/792

# DEDICATION

This project is dedicated to God Almighty, whose grace and wisdom guided every step of this research, and to my uncle, Dr Smart Agwu Basile, whose unwavering love, sacrifice, and support saw me through every stage of this academic journey. I am forever grateful.

## ACKNOWLEDGEMENTS

I give all glory and honour to God Almighty for the strength, wisdom, and perseverance he granted me throughout this programme.

I sincerely thank my project supervisor, Dr. Frank Okebanama, for his invaluable guidance, constructive feedback, and patient direction throughout the course of this research. Your expertise and encouragement shaped this work in immeasurable ways.

My sincere appreciation also goes to Engr Kingsley and Mr Camillus Njoku, lecturers in the Department of Computer Science and Mathematics, whose thorough review of this work and constructive corrections significantly improved the quality and academic rigour of this project. Their time and dedication are deeply valued.

I also extend my appreciation to the Dean of the faculty, The HOD and all the lecturers in the Department of Computer Science and Mathematics, Godfrey Okoye University, for their dedication to teaching and their investment in my academic development.

My heartfelt gratitude goes to my uncle and my entire family for their prayers, financial support, and constant encouragement. Their belief in me gave me the strength to see this project through to completion.

Finally, I acknowledge my colleagues and friends whose insights and shared learning enriched my understanding of software engineering and kept me motivated during challenging periods.

## ABSTRACT

Despite agriculture being the mainstay of the economy in Nigeria, small scale farmers in the country still depend on subjective and manual methods for making predictions on their yield, hence ineffective plans leading to food insecurity. In this paper, an enhanced system for predicting the farm yield of Nigeria's native crops was designed using two catboost algorithms; CatBoost classifier that helps in recommending crops and a CatBoost regressor that makes yield prediction for those crops, which were all hosted on a website developed using React 19 and Flask framework. A dataset containing 20,000 records from Nigeria's agricultural sector that had data of nine crop types, six geopolitical zones, 36 states and the Federal Capital Territory, and five agro-ecological zones was downloaded from Kaggle and preprocessed using a six step pipeline comprising of column dropping, stratified median imputation, and creation of seven new columns. The CatBoost Classifier gave a classification accuracy of 89.5 percent in a test set of 4,000 records, giving a Macro F1-Score of 0.8132 and Weighted Precision of 90.0 percent. The CatBoost Regressor gave an R2 of 98.48 percent, RMSE of 435.9 kg/ha, MAE of 308.8 kg/ha, and MAPE of 11.36 percent. Feature Importance Analysis validated that Humidity and Potassium levels in soil are the key feature for crop recommendation and Crop Type explains 60 percent variability in the yield data. This system is the only one mentioned in the relevant literature which has features like full deployment on web, covergae for nine crops in Nigeria, dual task catboost and agronomy recommendations all together in one platform.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background of the Study

The agricultural sector is still the major source of livelihoods in Africa, forming about 22 percent of GDP and more than 60 percent of the labor force in Africa (Lere, 2024). In Nigeria, for example, agricultural activities occupy not less than 36 percent of the national labor force, and are fundamental to food security across all the six geopolitical zones of Nigeria (Achoja and Michael, 2024). There are four sub-sectors in agriculture: crops production, fishing, livestock, and forestry, where crops production accounts for about 87.6 percent of total agricultural production in Nigeria. This notwithstanding, agriculture in Nigeria is known for its inefficiencies and low output levels, especially for small-scale farmers working on less than two hectares of land who make up more than 90 percent of the total production level of food products in Nigeria.

Yield in farming is the output of a crop measured in terms of volume or weight per area of land or input employed for farming, usually stated as kilograms per hectare (Minet et al., 2017). It is important to accurately estimate the yield because it helps with various decision-making processes such as inputs sourcing, storage, selling strategies, loans, and even food policies of the country. The current techniques of estimating yield employed by the smallholder farmers in Nigeria are rudimentary, imprecise, and have not been changed in the last few decades. Usually, farmers rely on visual observation, remember past seasons' occurrences, and then roughly estimate yields either with or without the help of an extension agent. They are subjective, delayed, and utterly incapable of estimating the complex interaction between weather patterns, soil chemistry, and farming practices (Paudel et al., 2021).

The agriculture industry in Nigeria is faced with several complex problems, which not only make it difficult for farmers to predict their yields but also make such predictions highly critical. This is because of the wide variety of agro-ecological zones found within Nigeria. Such zones include the humid tropical forest zone in the South as well as the Sahel Savanna zone in the North. In addition to this, soil erosion, unpredictable rainfall caused by climate changes, and erratic input availability make it extremely hard for farmers to predict their yields. It is mainly the small-scale farmers in Nigeria who suffer greatly from these problems since they have no economic cushion to protect themselves from unexpected low crop yields. They often make decisions based on hunches and not facts.

Machine learning has developed into a powerful tool in the area of predicting agricultural yields worldwide. Machine learning refers to an area of artificial intelligence that is capable of analyzing previous information and making predictions based on the learned patterns on future, unknown data sets (Liakos et al., 2018). Applications of machine learning technology in agriculture include predicting yields of various crops, suggesting the best types of crops to grow under specific soil conditions, detecting plant diseases, and optimizing watering programs. All these applications have proven that the prediction of data is able to significantly surpass traditional estimates when more than one predictor variable is taken into account at once.

The machine learning algorithms which can be applied for the purposes of agriculture prediction include many different types. However, the most successful among all of them when working with structured tabular data have proven to be the gradient boosting algorithms. Among those gradient boosting algorithms, CatBoost created by Yandex in 2017 demonstrates superior results on the data containing both numerical and categorical features. Data from agriculture in Nigeria contains plenty of categorical variables like type of soil, agro-ecological zone, region, state, crop variety, as well as farming practices. All of them can be dealt with through CatBoost's Ordered Target Statistics encoding without any additional processing (Jabed and Murad, 2024). The superiority of CatBoost over other similar gradient boosting algorithms such as XGBoost and LightGBM has been experimentally proven for the Nigerian and West African contexts (Adegoke et al., 2023).

Although considerable advances have been made in machine learning systems that can predict agricultural yields, one very important area has yet to be addressed. The vast majority of such machine learning systems remain prototype academic solutions, designed, tested, published in journals, and thereafter left out of reach of farmers who would benefit from their implementation. Although some machine learning systems do possess deployment solutions, these deployment solutions cater to just one or two types of crops and do not take into account the unique soil and climatic factors that are inherent to Nigeria.

The current research fills the existing gap by designing an Enhanced Farm Yield Prediction System for Nigerian indigenous crops. The proposed system utilizes two CatBoost models: the first one is a CatBoostClassifier used for selecting the most appropriate crop according to specific soil and weather conditions, while the other model is a CatBoostRegressor for predicting farm yield in kilogram per hectare. All the models are provided via a React 19 application running on Vercel platform and Flask REST API running on the Render cloud hosting service, allowing any farmer to use the system with a smartphone and the Internet access. The current research considers nine Nigerian indigenous crops in six geopolitical zones, 36 states and FCT, as well as five agro-ecological zones.

### 1.2 Statement of the Problem

1. Despite the advancements in the agricultural machine learning research arena, the problem of yield estimation is still out of reach for the Nigerian small-scale farmers because of the following factors:
2. The process of farm yield estimation in Nigeria still relies heavily on conventional methods of estimating the yield which are purely subjective, unreliable, and fail to consider the non-linear correlation between soil nutrients, climate, and management practices.
3. Agricultural machine learning algorithms used for farm yield estimation still lack proper implementation in real-world situations as they are mostly found only in laboratories without any form of usability by the farmers.

4. There are very few available farm yield estimation platforms and those existing do not cater for the major crops in Nigeria due to non-customization to Nigerian agro ecological zones.
5. None of the available systems provide crop recommendation along with yield estimation.

### 1.3 Aim and Objectives of the Study

The goal of this research is to design and implement a Farm Yield Prediction System for indigenous crops in Nigeria. To meet this goal, the following specific objectives have been set:

1. Collection and analysis of relevant agricultural and climatic data on indigenous Nigerian crops.
2. Data preprocessing and feature engineering in order to prepare data for modeling using machine learning algorithms.
3. Modeling and implementation of crop recommendation and farm yield prediction models using CatBoost algorithm.
4. Assessment of the built models based on different performance measures for classification and regression problems.
5. Implementation of web application which will integrate and provide access to prediction models implemented in previous objective.

### 1.4 Significance of the Study

Crop yield prediction is one of the vital capabilities of predictive analytics in contemporary agriculture. An accurate and easy-to-use predictive model allows for more efficient farming operations, from optimal planning of plantations and fertiliser use to proper crop harvest, which results in higher agricultural efficiency and food security. The proposed Farm Yield Prediction System is significant to the following groups:

1. **Smallholder farmers** — gain access to data-driven crop recommendations and yield estimates without requiring any prior technical knowledge, enabling better planting and input procurement decisions.
2. **Agricultural extension agents** — can use the system as a digital advisory tool to provide faster, more accurate guidance to farmers across different agro-ecological zones.
3. **Agricultural planners and policymakers** — gain access to aggregated yield trend data across nine crops and six geopolitical regions, supporting evidence-based food security planning and resource allocation.
4. **Academic researchers** — can replicate and extend the proposed dual-task CatBoost architecture combined with the React and Flask deployment framework as a foundation for future work in Nigerian and African agricultural machine learning.
5. **Institutions and development agencies** — organisations such as FMARD, NiMet, and IITA can integrate the system's open REST API into their existing digital farmer advisory platforms to extend its reach to a larger farmer population.

### 1.5 Scope of the Study

The research involves the design and assessment of an advanced farm yield prediction model that includes nine native crops found in Nigeria. This crop prediction model uses a dataset consisting of 20,000 records obtained from Kaggle. The dataset used consists of six geopolitical zones, 36 states, and FCT, five agro-ecological zones. The input variables will involve macronutrients in the soil (nitrogen, phosphorus, and potassium), pH of the soil, humidity, rainfall, temperature and other categories in agriculture. Historical datasets will be used, but not any real-time dataset generated from sensors.

### 1.6 Limitation of the Study

1. The study relies heavily on historical agricultural data, obtained from an open-source data repository. The poor quality of the data can potentially impact the generalisability of the trained models when applied in the real world.

2. The models were trained using data relating to nine types of crops only. This means that there is an inherent risk that the models will not be accurate when predicting yields of crop varieties/agroecological zones not included in the training set.

3. There was no provision for real-time variables, such as pests or weather conditions, that can adversely affect the predictions. The models did not include satellite-based/remote sensing indices like NDVI/EVI (as demonstrated by Muhammad et al., 2026).

4. The Onions class consisted of very few test data points; thus, its metrics cannot be statistically significant compared to those of the other classes.

5. There was no provision for real-time variables, such as pests or weather conditions, that can adversely affect the predictions. The models did not include satellite-based/remote sensing indices like NDVI/EVI (as demonstrated by Muhammad et al., 2026).

6. The Onions class consisted of very few test data points; thus, its metrics cannot be statistically significant compared to those of the other classes.

### 1.7 Operational Definition of Terms

**Yield**: Quantity of agricultural produce harvested per unit area of land, measured in kilograms per hectare (kg/ha).

**Machine Learning**: The science of artificial intelligence that involves using algorithms trained with historical data to predict results for previously unknown data.

**CatBoostClassifier**: The classifier version of the CatBoost gradient boosting framework that encodes categorical inputs using Ordered Target Statistics, used in this study for classifying crops.

**CatBoostRegressor**: The regressor version of the CatBoost gradient boosting framework, used in this study for predicting yield values in kg/ha.

**Feature Engineering**: The process of creating additional features from the initial raw input data based on domain knowledge about how the output value was created.

**Agro-Ecological Zone**: Geographic region with similar soil, climate and vegetation properties, used in this study as a categorical feature for crop yield estimation.

**CRISP-DM**: Cross-Industry Standard Process for Data Mining — the six-phased approach to machine learning used to conduct this project.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Introduction

In this chapter, an organized literature review will be conducted based on available resources concerning the Farm Yield Prediction System. The goal of this literature review will be to map out the progress of development of crop yield prediction techniques from its initial manual forms to current use of gradient boosting approach, and also to find out what gaps the new system seeks to fill. This chapter comprises theoretical review in section 2.2, a review of twenty literature articles in section 2.3, a summary table and comparison table in section 2.4, and finally research gaps in section 2.5.

## 2.2 Theoretical Background

### 2.2.1 Concept of Crop Yield Prediction

The prediction of crop yields is the act of determining the amount of agricultural production that is expected to be obtained at the end of the growing period for a particular field (Jabed and Murad, 2024). This depends on the evaluation of historical and present information, which includes weather, soil, and crops in order to provide a quantitative prediction of the expected output. It is necessary to note that there is a distinction between crop yield prediction and crop recommendation. While the former determines the expected amount of a crop that will be produced, the latter suggests what type of crop should be planted given certain conditions (Sundari et al., 2022).

### 2.2.2 Overview of Predictive Analytics in Agriculture

Predictive analysis within the agriculture sector refers to the utilization of analytical models based on statistical algorithms and machine learning to predict future trends based on existing or current data concerning things like crop productivity, crop diseases, and pricing information (Bohra and Gupta, 2025). Traditional methods of crop productivity estimation that have been widely used in the sub-Saharan African region, for example, include farmer memory, visual inspections, crop-cut tests, and the use of historical crop average numbers adjusted through the help of agricultural extension workers. Digitization of agricultural data along with the availability of open-source machine learning software, which does not need any specific hardware to perform, is what has made the shift towards predictive analysis possible.

### 2.2.3 Machine Learning Techniques for Yield Prediction

Linear Regression is the oldest and simplest method in machine learning for the prediction of crop yields but does not consider non-linear relationships existing within agro-ecological systems (Abbas et al., 2020). Decision Trees and Ensemble methods are a step forward in the prediction of crop yields and have shown good results; however, in case of using Random Forests, many decision trees are trained independently, and the outcomes of different trees are combined to achieve maximum accuracy (Jhajharia et al., 2023). The Gradient Boosting models, including XGBoost, LightGBM, and CatBoost, further enhance the process by sequentially growing trees in such a way that every next tree counteracts the mistakes made by all preceding trees. CatBoost ensures the superiority on the dataset with categorical values due to Ordered Target Statistics encoding used in the model, which makes the manual encoding unnecessary and prevents data leakage (Jabed and Murad, 2024).

### 2.2.4 Web-Based and Deployed Agricultural Prediction Tools

Most research work in machine learning for agricultural crop yield forecasting is stuck at the prototype level without ever becoming available for use by the general public. This is among the most crucial shortcomings of current literature (Bohra and Gupta, 2025). A properly designed web application which is not only lightweight and responsive but also capable of being executed in any browser can solve the problem of limited access caused by poor internet connectivity and poor computer literacy of small-scale farmers in Nigeria more efficiently compared to the use of desktop or CLI tools. The availability of cloud-based deployment systems like Vercel and Render makes it technically easier to deploy machine learning-based web apps.

## 2.3 Review of Related Works

The following twenty publications were chosen based on their relevance to machine learning techniques used for crop yield predictions, crop recommendations based on soil nutrient composition, prediction of agriculture in Nigeria or Africa, web-deployed machine learning applications in agriculture, and gradient boosting algorithms. The sources include peer-reviewed publications in journals or conference papers from the period 2020 to 2026.

In the study by Iniyan et al. (2023), the authors presented an ML model based on regression techniques for crop yield prediction based on an enhanced feature set using LSTM with an accuracy of 86.3 percent. In addition, there was a web application developed for the benefit of farmers. However, the main drawback of this paper is the local nature of deployment of the web application.

ML vs DL algorithms for prediction of five crop yields in India were examined by Jhajharia et al. (2023). The highest value of R-squared was obtained for Random Forest, which equals 0.963. This research was limited to geographical area of India and did not consider the diversity of categorical features in Nigerian datasets. The paper served as an initial reference point for RF implementation and model selection prior to using CatBoost algorithm.

A hybrid ML-DL framework for prediction of yield and soil nutrients along with estimating the cost of fertilizers based on prediction was designed by Agarwal and Tarar (2021), which is related to the Recommendation Service part of the proposed system in conceptual terms. The system is computational complex and lacks the web deployment.

ML algorithms such as linear regression, elastic net, k-NN, and SVR were used for prediction of potato yields via proximal sensing in Abbas et al. (2020). The importance of good quality features as a factor influencing prediction accuracy has been emphasized. An important shortcoming of the method was high amounts of sensing data required that are hard to obtain for Nigerian smallholder farmers.

Paudel et al. (2021) used process-based crop modelling and ML to predict yields on a continent-wide scale in Europe with an NRMSE of 7% to 8%. It created performance standards which served as benchmarks for the predicted system's targets for regression. It did not test its methodology on Africa's agricultural environment and uses institutional datasets not available to the average Nigerian farmer.

Shawon et al. (2024) built a stacked model consisting of RF and XGBoost algorithms for predicting rice and maize yields in Bangladesh with R-square scores of 0.94. Its dual-model strategy directly influenced the development of the system's architecture for dual tasks through CatBoost. Its testing scope was limited to two crops in one country without any deployment process.

Jabed and Murad (2024) compared XGBoost, LightGBM, and CatBoost algorithms for crop yield prediction in multiple structured tabular datasets. CatBoost was the most accurate of the three algorithms, especially for categorical-feature datasets. This paper offers the most relevant empirical basis for the selection of CatBoost for the proposed model. There was no deployment mechanism and no user interface considered in this paper.

Sundari et al. (2022) developed a soil-based crop recommendation system employing Naive Bayes, Decision Trees, and Random Forest algorithms, where RF yielded an accuracy of 99.4% based on seven features comprising soil and climatic characteristics. The study provides validation of the proposed system's design of its inputs scheme. The main drawback of the study is the lack of utilization of any African or Nigerian data.

Bohra and Gupta (2025) introduced a crop advisory web portal that employs both classification and regression to increase crop selection accuracy by 31 percent among participating farmers. It affirms the practicality of employing both crop recommendation and web deployment. Data was collected from only one region and involving three crops.

Malhi et al. (2024) designed a temperature-informed ML framework for cereal yield prediction, introducing temperature-related stress indices which resulted in an improvement of 18 percent in the accuracy (RMSE). This supports temperature and humidity as essential features in predictive modeling, being consistent with their feature importance in the proposed model's classifier.

The research work by Luo et al. (2024) using SHAP analysis has shown that soil NPK and pH are the most important factors for paddy rice yield prediction through XGBoost regression. The finding empirically establishes the importance of soil NPK and pH as input variables in the suggested model, although the research was done in Southeast Asia.

The research work carried out by Ogundimu et al. (2022), in which models using machine learning algorithms were used to predict yield of maize and cassava from records obtained from an extension agency in southwest Nigeria, achieved accuracy of 88.7 percent using Random Forest algorithm. It establishes direct Nigerian evidence of the use of soil NPK as predictors.

Lastly, Adeyemi et al. (2023) suggested CNN-LSTM deep learning algorithm in prediction of yield of sorghum and beans with R-squared score of 0.89. This finding established the prediction power of machine learning algorithms in predicting these two crops in a Nigerian setting. There is no implementation phase, and there were only two crops considered.

Muhammad et al. (2026) tested the efficiency of vegetation index alongside the NPK and pH measurements of the ground for yield prediction in sub-Saharan Africa. They found that there is a 12 percent increase in R-squared with the use of both kinds of features. This justifies the inclusion of ground-measured soil inputs in the proposed system. A major limitation with this technique is that the use of satellite data is still unaffordable to the average Nigerian smallholder farmer.

Oluwafemi et al. (2024) implemented a Flask API and frontend using CatBoost for recommendation of tomato and pepper cultivation in southwestern Nigeria with an accuracy rate of 91.3 percent. This paper is the most similar architectural implementation to the proposed system. Limitations include the limitation to two crops in one state alone, and lack of yield estimation.

Okoye and Eze (2023) used Random Forest and SVM for the classification of yield levels of rice in southeastern Nigeria. An F1-score of 0.91 in the case of Random Forest shows the efficacy of ensemble methods for the prediction of Nigerian rice. Limitations included the restriction to one crop in one area without implementation.

Afolabi et al. (2022) created an Artificial Neural Network using composite soil fertility indicators for the prediction of cassava yield in the middle part of Nigeria, resulting in a reduction in RMSE of 23 percent when compared to linear models. This confirms that there is a possibility to predict cassava soil nutrient content in Nigeria. Only cassava was considered, and there was no implementation of this project.

Nweke et al. (2024) built an XGBoost model with SHAP interpretability used for maize yield prediction in West Africa, highlighting nitrogen and rainfall as the two key factors impacting yields. This research relates to the proposed model as it includes a project related to maize yield prediction in Nigeria. There was no user-friendly interface, and maize was the sole crop considered.

Adegoke et al. (2023) tested the performance of LightGBM and CatBoost classifiers on six agro-ecological zones in Nigeria, and they found that CatBoost performed better than LightGBM and showed the lowest value of RMSE when predicting the yields of five crops. This research confirms the choice of this model as one of the classifiers.

The most significant deployed precedent for the system suggested herein was by Bello et al. (2025), who built a fully deployed web-based application that uses Random Forest stacking and CatBoost for multi-crop yield prediction and fertiliser recommendation for Nigerian farmers, which was 93.1 percent accurate on three crops. However, its main weakness is that its coverage is limited to just three crops.

## 2.4 Summary of Literature Review

**Table 2.1: Summary of Related Works**

| S/N | Author(s) (Year) | Work Done | Findings/Results | Limitations |
|---|---|---|---|---|
| 1 | Iniyan et al. (2023) | Regression-based ML with farmer-facing web app | 86.3% accuracy | Local offline deployment only |
| 2 | Jhajharia et al. (2023) | ML/DL comparison for five-crop yield in India | RF: $R^2$=0.963 | Geographically confined to India |
| 3 | Agarwal & Tarar (2021) | Hybrid ML-DL for yield and soil nutrient estimation | Dual-model framework validated | High computational cost, no deployment |

| | | | | |
|---|---|---|---|---|
| 4 | Abbas et al. (2020) | Four ML models for potato yield using sensing data | Feature quality is critical for accuracy | Requires large sensing datasets |
| 5 | Paudel et al. (2021) | Process-based ML for European yield forecasting | nRMSE: 7–8% | Not evaluated on African conditions |
| 6 | Shawon et al. (2024) | Stacked RF + XGBoost for Bangladesh crops | $R^2$=0.94 for rice and maize | Two crops, single country, no deployment |
| 7 | Jabed & Murad (2024) | XGBoost vs LightGBM vs CatBoost comparison | CatBoost best on categorical-feature-rich data | No deployment or user interface |
| 8 | Sundari et al. (2022) | Soil-based crop recommendation system | RF: 99.4% accuracy on 7 features | No African data used |
| 9 | Bohra & Gupta (2025) | Deployed web crop advisory platform | 31% reduction in crop selection errors | Single region, three crops only |
| 10 | Malhi et al. (2024) | Temperature-aware ML for cereal yield | 18% RMSE improvement | Cereals only, no deployment |
| 11 | Luo et al. (2024) | SHAP analysis of NPK/pH on rice yield | NPK and pH confirmed top features | Southeast Asian context only |
| 12 | Ogundimu et al. (2022) | ML for maize and cassava yield in SW Nigeria | RF: 88.7% accuracy | Two crops, single agro-ecological zone |
| 13 | Adeyemi et al. (2023) | CNN-LSTM for sorghum and beans in N. Nigeria | $R^2$=0.89 | Two crops, no deployment |
| 14 | Muhammad et al. (2026) | Vegetation indices + NPK for yield in Africa | 12% $R^2$ improvement | Satellite data too expensive for smallholders |
| 15 | Oluwafemi et al. (2024) | Flask + CatBoost web for tomato and pepper in Nigeria | 91.3% accuracy | Two crops, one state, no yield estimation |
| 16 | Okoye & Eze (2023) | RF and SVM for rice yield in SE Nigeria | RF: F1=0.91 | Single crop, single region, no deployment |
| 17 | Afolabi et al. (2022) | ANN for cassava yield in Nigerian middle belt | 23% RMSE reduction over linear baseline | Cassava only, no deployment |

| | | | | |
|---|---|---|---|---|
| 18 | Nweke et al. (2024) | XGBoost + SHAP for maize in West Africa | N and rainfall confirmed top features | Maize only, no farmer-facing tool |
| 19 | Adegoke et al. (2023) | LightGBM vs CatBoost across Nigerian agro zones | CatBoost: lowest RMSE across all 5 crops | No web interface or recommendation |
| 20 | Bello et al. (2025) | End-to-end web app for Nigerian multi-crop yield | 93.1% accuracy | Three crops only |

**Table 2.2: Comparison of Proposed System with Five Related Works**

| Feature | Bello et al.(2025) | Oluwafemi et al.(2024) | Adegoke et al.(2023) | Ogundimu et al.(2022) | Sundari et al.(2022) | ProposedSystem |
|---|---|---|---|---|---|---|
| Algorithm | RF + CatBoost | CatBoost | LightGBM + CatBoost | RF + Logistic Reg. | Naive Bayes, DT, RF | CatBoostClassifier + CatBoostRegressor |
| Crops Covered | 3 | 2 | 5 | 2 | 7 (non-Nigerian) | 9 Nigerian indigenous crops |
| Nigerian Data | Yes | Yes (SW only) | Yes | Yes (SW only) | No | Yes (all 6 regions) |
| Crop Recommendation | Yes | Yes | No | No | Yes | Yes |
| Yield Estimation | Yes | No | Yes | Yes | No | Yes |
| Both Tasks Combined | Yes | No | No | No | No | Yes |
| Web Deployed | Yes | Yes | No | No | No | Yes (Vercel + Render) |
| Soil Improvement Tips | No | No | No | No | No | Yes |

| | | | | | | |
|---|---|---|---|---|---|---|
| Classifier Accuracy | 93.1% (3 crops) | 91.3% (2 crops) | N/A | 88.7% (2 crops) | 99.4% (non-Nigerian) | 89.5% (9 Nigerian crops) |

### 2.5 Research Gap

Based on the above analysis of twenty reviewed papers, two gaps have been found which can be addressed using the proposed system:

1. **Lack of deployed web application:** Out of the twenty reviewed papers, only four developed a system beyond prototype level, and one of them was limited to local offline environment. There is no real world use of an application which does not come into the hands of farmers and hence is not useful despite its accuracy. This gap will be addressed in our work as the application for both models will be deployed via a live web application.

2. **Limited crop coverage without dual tasking:** The widest Nigerian specific work used yield regression alone and had coverage of only five crops while others had coverage of two crops each. Some of the studies even considered only a single crop. None of the reviewed systems considered both yield regression and crop recommendation with more than five crops. Our proposed farm yield prediction system will cover all nine indigenous Nigerian crops: beans, cassava, maize, onions, pepper, rice, sorghum, tomatoes, and yams.

# CHAPTER THREE

# SYSTEM ANALYSIS AND DESIGN

## 3.0 Introduction

In this chapter, we will discuss the Analysis & Design of the proposed Farm Yield Prediction System. In this chapter, we will first consider existing methods for farm yield predictions in Nigeria, highlighting the shortcomings of those methods and giving a brief introduction to the proposed system. Next, we describe the system requirements. We explain the OOAD methodology for designing the system. The CRISP-DM methodology for developing a Machine Learning solution is discussed along with its application in this project. We then discuss data gathering and preprocessing, UML modeling of the system, system inputs, outputs, database schema, and briefly introduce the two architectures of the CatBoost model and their pipeline.

## 3.1 System Analysis

### 3.1.1 Analysis of the Existing System

The current predictive techniques employed in estimating the yields of farms in Nigeria can be divided into two main groups. One group consists of conventional and manual techniques such as memory-based, visual, crop cutting, and historical average yield predictions modified by extension officers. All of the above techniques are based solely on field observations without incorporating any digitalization of weather and soil data. Another group consists of scientific models developed through machine learning algorithms that show considerable precision but are still theoretical or offline models that are not relevant to the geographical area in question.

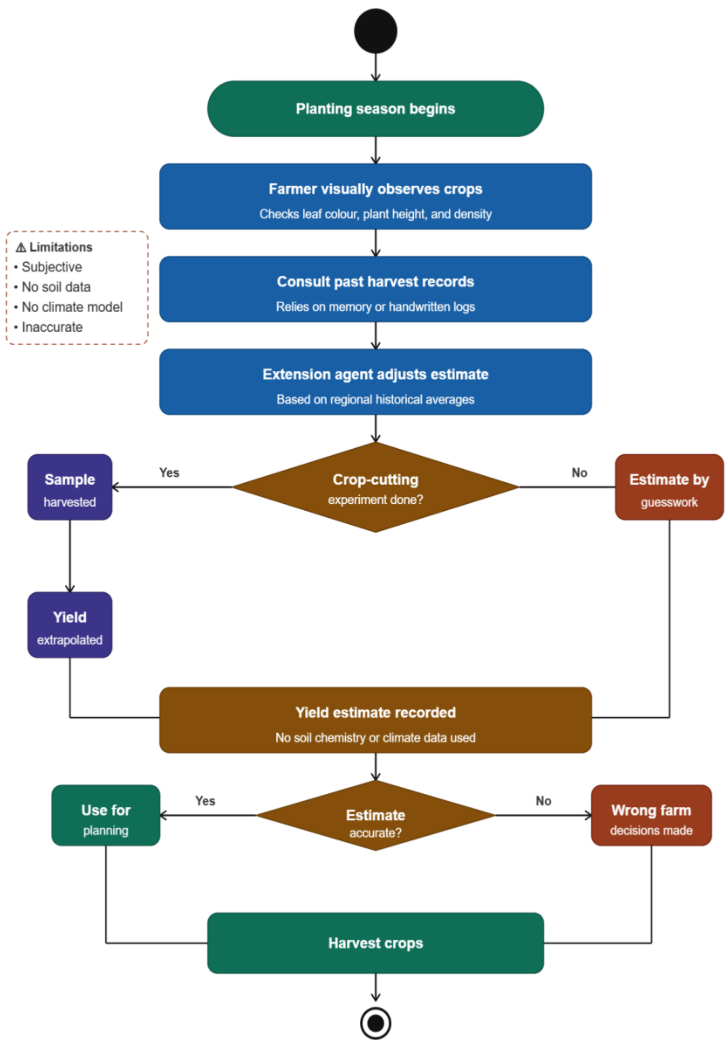


*Figure 3.1: Flowchart of the Existing Yield Estimation Process*

### 3.1.2 Limitations of the Existing System

1. The traditional techniques involve use of past average and subjective analysis, making them unreliable and incapable of capturing the non-linearity of weather-soil-management interaction.

2. All Machine Learning models currently discussed in the literature are either conceptual models, offline models, or locally implemented with no provision for a web-based application by users such as farmers.

3. None of these models are customized for the peculiarities of Nigerian indigenous crops and climate of the soils found in Nigeria's agro-ecological zones.

4.There is no system that provides an easy-to-use web-based application for rapid yield forecasting and crop recommendation to improve agronomic performance.

### 3.1.3Analysis of the Proposed System

The proposed Farm Yield Prediction System is a web application that leverages machine learning to overcome the challenges faced by the current solutions by offering Nigerian farmers and agricultural strategists data-driven predictions about which crops to plant based on their soil quality and climate condition, along with yield projections in kilograms per hectare and helpful suggestions about how to improve the quality of the soil in the process.

As mentioned earlier, the Farm Yield Prediction System is designed using a three-phase pipeline, where Phase One involves collecting data from reliable sources, cleaning the data, and applying feature engineering techniques to prepare the dataset for the next phase, training the two models. Phase Two entails training and optimizing the CatBoostClassifier and the CatBoostRegressor. Phase Three involves deploying the two models using a React 19 frontend running on Vercel and a Flask REST API running on Render..

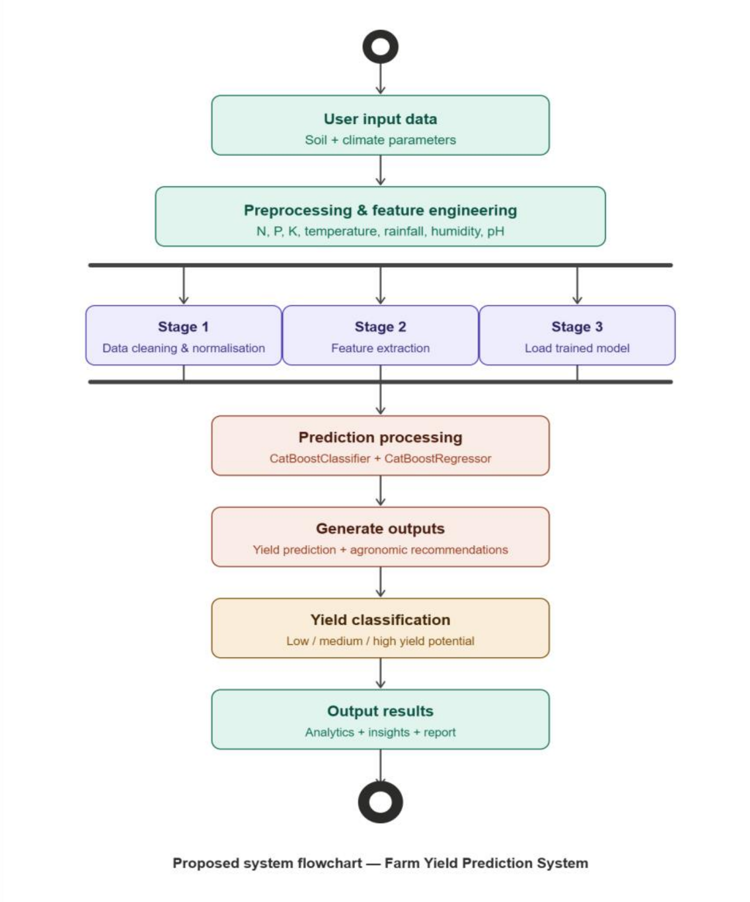


*Figure 3.2: Overview of the Proposed Three-Stage System*

## 3.2 System Requirements Specification (SRS)

### 3.2.1 Functional Requirements

FR1: Allow users to create an account, login and update profile information.

FR2: Take input data such as nitrogen, phosphorus, potassium, soil pH, humidity, temperature, and rain fall along with additional data like region, state, agro ecological zones and soil types.

FR3: Preprocess inputs, derive engineered features and call both cat boost models in order to provide crop suggestion along with confidence score and an estimation of yield in terms of kg/ha.

FR4: Provide prediction output along with suggestions on improving agricultural production based on the comparison between submitted soil value data against the crop’s best range values.

FR5: Save all prediction sessions to the database as well as provide facility to farmers to submit the actual yield values after harvest.

FR6: Provide administrative facilities to manage and train the model.

### 3.2.2 Non-Functional Requirements

Performance: It is expected that the prediction will take no more than 2.0 seconds.

Usability: The user interface will be understandable even for farmers without any technical skills, including correct labelling and instant validation.

Accuracy: CatBoostClassifier will provide classification accuracy higher than 85 percent, and CatBoostRegressor will have an $R^2$ value higher than 0.70.

Scalability: The service will handle several users at once by using cloud hosting provided by Render.

Security: Passwords will be hashed, and token-based access control will be implemented by using JWT.

Compatibility: It is assumed that the frontend will work in all up-to-date web browsers.

## 3.3 System Design Methodology (OOAD)

The Farm Yield Prediction System uses Object-Oriented Analysis and Design (OOAD). Under this approach, a system can be thought of as a composition of different objects or classes. Decomposition is possible using OOAD due to which various system components can be created that are assigned various duties, such as ML inference, authentication, and data storage. Use case, activity, class, and sequence diagrams from UML are employed to illustrate the system architecture and behavior.

### 3.3.1 Justification for OOAD

1. Object-oriented analysis and design (OOAD) is applicable since all the system elements such as the Farmer/User, System Administrator, CatBoost Inference Engine, Prediction Service, Recommendation Service, Authentication Service, and Data Management Module are well-defined, individually testable, and communicate via interfaces. There are at least three benefits of using OOAD for our problem:

2. Reusability and Modularity: The CropPredictionService Singleton and RecommendationService are reusable within many API endpoints without being replicated.

3. Visualization of the System Elements: The use of UML diagrams visualizes the relationships between actors and elements of the system and data movement between system components.

4. Scalability and Maintainability: Individual system elements allow updating machine learning models, introducing new crops, or updating databases.

### 3.4 Machine Learning Development Methodology (CRISP-DM)

The machine learning element of the Farm Yield Prediction System was created based on the CRISP-DM approach – Cross Industry Standard Process for Data Mining. CRISP-DM represents the most widely adopted methodology for the machine learning processes that include six stages, which are implemented in cycles, including: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The cyclical nature of CRISP-DM allows results from one stage to take the developer back to an earlier stage.

#### 3.4.1 Justification for CRISP-DM

It should be stressed that CRISP-DM was chosen due to its problem-focused nature as compared to its technology-focused alternative. In other words, this methodology ensures that all technical choices made in the process of creating a solution relate to the practical needs of the client. Moreover, the cyclical character of the approach allowed re-visiting earlier stages where required. Finally, all CRISP-DM stages match each phase of this project development perfectly.

### 3.4.2 Application of CRISP-DM to This Project

**Phase 1** – Business Understanding: The main problem statement was to address the inaccuracies and difficulties in estimating the yield for Nigerian smallholder farmers. For measuring success, the thresholds were set at more than 85% accuracy and more than 0.70 $R^2$ value.

**Phase 2** – Data Understanding: The dataset was analyzed to comprehend the 20,000 rows, nine crops, six regions, 36 states and the FCT, five agro-ecological zones, and 55,556 missing values.

**Phase 3** – Data Preparation: Six data preparation techniques were used: column dropping, stratified median imputation, validation, export, feature engineering, and cross-validation splitting.

**Phase 4** – Modelling: Two CatBoost algorithms were trained on Google Colab utilizing a NVIDIA T4 GPU with 1,000 trees, learning rate of 0.05, and depth of 8.

**Phase 5** – Evaluation: Classification accuracy and Macro F1-Score were used for evaluating the classifier. As for the regressor, $R^2$, RMSE, MAE, and MAPE metrics were chosen.

**Phase 6** - Deployment: Both models were deployed within the Flask REST API using the CropPredictionService Singleton. The React UI is deployed using Vercel, while the Flask REST API is deployed using Render.

### 3.5 Method of Data Collection

The system uses only historical secondary data collected from a dataset of agricultural yield in Nigeria available at Kaggle and contains 20,000 data entries concerning nine native crops grown in six geopolitical zones, thirty-six states, and the Federal Capital Territory, along with five agro-ecological zones. The parameters included in the dataset, which are entered by farmers into the prediction form, include:

Environmental Parameters: nitrogen (N) in mg/kg, phosphorous (P) in mg/kg, potassium (K) in mg/kg, soil pH, temperature (°C), rainfall (mm), and humidity (%) Advanced Regional and Operational Parameters: region (geopolitical zone), state, agro-zone (agro-ecological zone), and soil type.

The above parameters are exactly the same parameters provided for input in the prediction form on the system's website, as depicted in Figure 4.8 of Chapter Four.

## 3.6 Dataset Description and Preprocessing

### 3.6.1 Dataset Description

The raw dataset is a Nigerian agricultural yield dataset sourced from Kaggle, stored as nigeria_agri_yield_enhanced.csv. It contains 20,000 records and 25 columns. Table 3.1 summarises the raw dataset.

**Table 3.1: Raw Dataset Summary**

| Characteristic | Details |
|---|---|
| Source | Kaggle — Nigerian Agricultural Yield Dataset (nigeria_agri_yield_enhanced.csv) |
| Total Records | 20,000 |
| Total Columns | 25 (including target variable) |
| Target Variable | yield_kg_ha — crop yield in kilograms per hectare |
| Target Mean / Min / Max | 4,984 kg/ha / 500 kg/ha / 25,000 kg/ha |
| Total Missing Values | 55,556 across 10 columns |
| Crop Types | Beans, Cassava, Maize, Onions, Pepper, Rice, Sorghum, Tomato, Yam |
| Regions | Northcentral, Northeast, Northwest, Southeast, Southsouth, Southwest |
| Agro-Ecological Zones | Humid Forest, Derived Savanna, Northern Guinea Savanna, Sudan Savanna, Sahel Savanna |
| States Covered | 36 states and the Federal Capital Territory (FCT) |

### 3.6.2 Data Preprocessing

The raw dataset was transformed into a clean, training-ready file through a six-step preprocessing pipeline retaining all 20,000 records with zero missing values. Table 3.2 summarises the steps.

**Table 3.2: Preprocessing Steps Summary**

| Step | Action | Details |
|---|---|---|
| 1 | Column Dropping | 7 columns removed: farm_id (non-predictive), fertilizer_type (29.7% missing), irrigation_type (71.5% missing), temperature_stress (57.3% missing), extreme_weather (57.8% missing), soil_degradation (46.2% missing), fertilizer_amount_kg_ha (3.2% missing) |
| 2 | Missing Value Imputation | 4 numerical columns imputed using median within crop_type and agro_zone strata: farm_size_ha (601), soil_nitrogen (615), soil_phosphorus (567), soil_potassium (636) |
| 3 | Validation | All columns verified to contain zero missing values after imputation |
| 4 | Export | Cleaned dataset saved — 20,000 records, 18 columns |
| 5 | Feature Engineering | Seven derived features computed (see Table 3.3) |
| 6 | Cross-Validation Split | 5-fold stratified on crop_type; 70% training, 15% validation, 15% test |

### 3.6.3 Feature Engineering

Seven derived features were computed from the cleaned dataset to expose biological and environmental relationships not captured by raw values alone.

**Table 3.3: Engineered Features**

| Feature Name | Formula | Description |
|---|---|---|
| soil_fertility_index | N + P + K | Composite measure of overall soil macronutrient availability |
| np_ratio | N / (P + 1e-6) | Nitrogen-to-phosphorus balance ratio |
| climate_index | rainfall × temperature | Combined effect of heat and moisture on crop growth |
| ph_stress | abs(pH − 7.0) | How far soil pH deviates from the ideal neutral point |
| n_ph_inter | N × pH | How soil acidity affects nitrogen availability |
| p_ph_inter | P × pH | How soil acidity affects phosphorus availability |
| rain_temp_ratio | rainfall / (temperature + 1e-6) | Ratio of moisture to heat — captures drought or waterlogging risk |

## 3.7 System Modelling Using UML

The following UML diagrams describe the system's actors, functional scope, dynamic behaviour, static structure, message flow, and physical architecture.

### 3.7.1 Use Case Diagram

Use Case Diagram Interaction between two Actors i.e. Farmer/User and System Administrator and System Boundary. The Farmer/User is able to register, login, make a request for crop yield prediction, receive crop recommendation and nutritional recommendations, and finally, analyze the analytics. The System Administrator is capable of monitoring system health, training both the CatBoost models, evaluating models' performance, and managing training data..

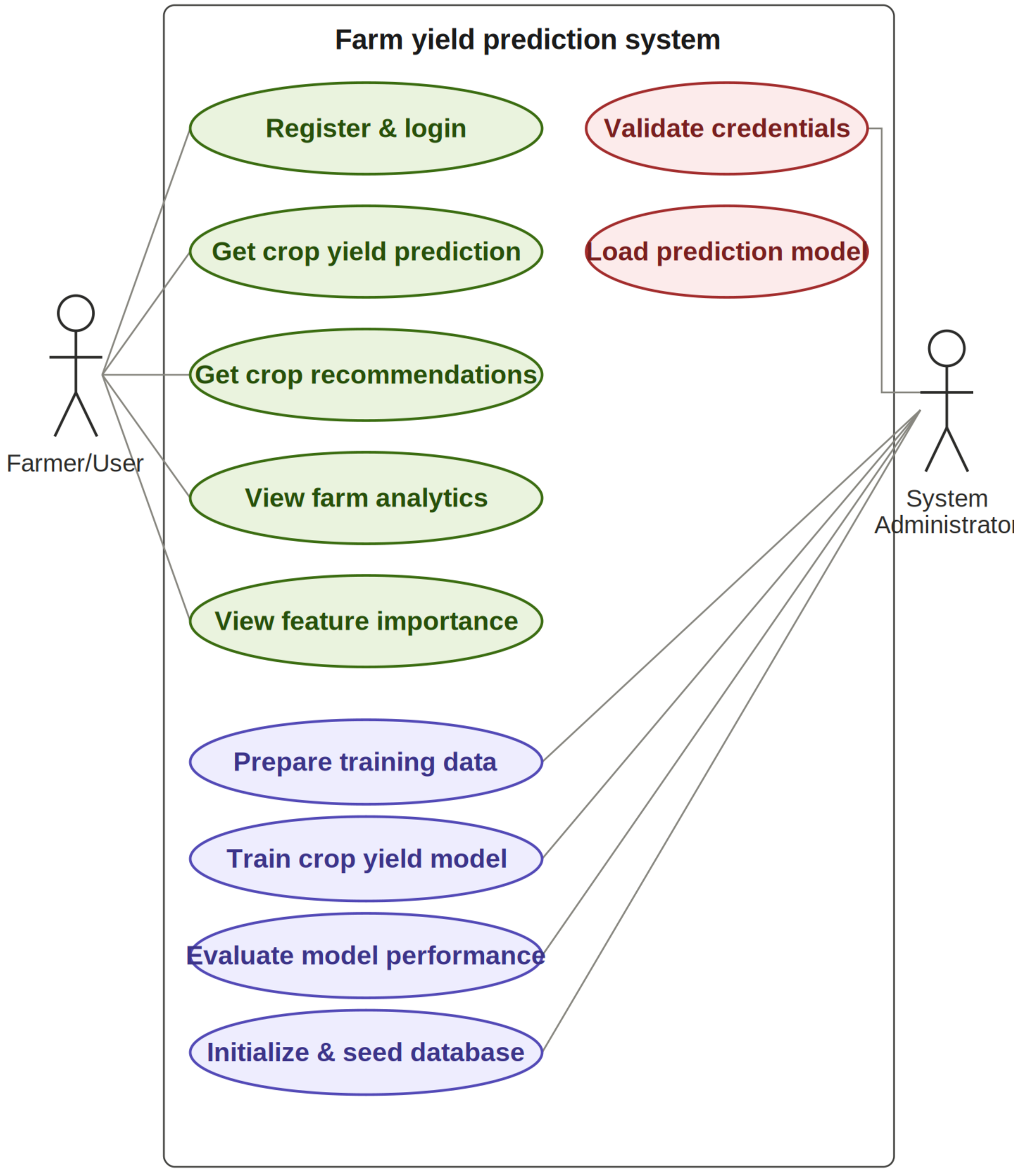


*Figure 3.3: Use Case Diagram for the Proposed System*

### 3.7.2 Activity Diagram

This activity diagram starts with the process when the user fills out the input data and ends with the output data being produced and delivered in the form of a prediction result. After that, the input data goes for validation and further proceeds through the process of feature engineering. Afterward, the processes get split into two separate flows, one of which consists of producing a crop recommendation with a reliability index by means of the

CatBoostClassifier model, and another one is about predicting the yield in kg/ha using the CatBoostRegressor.

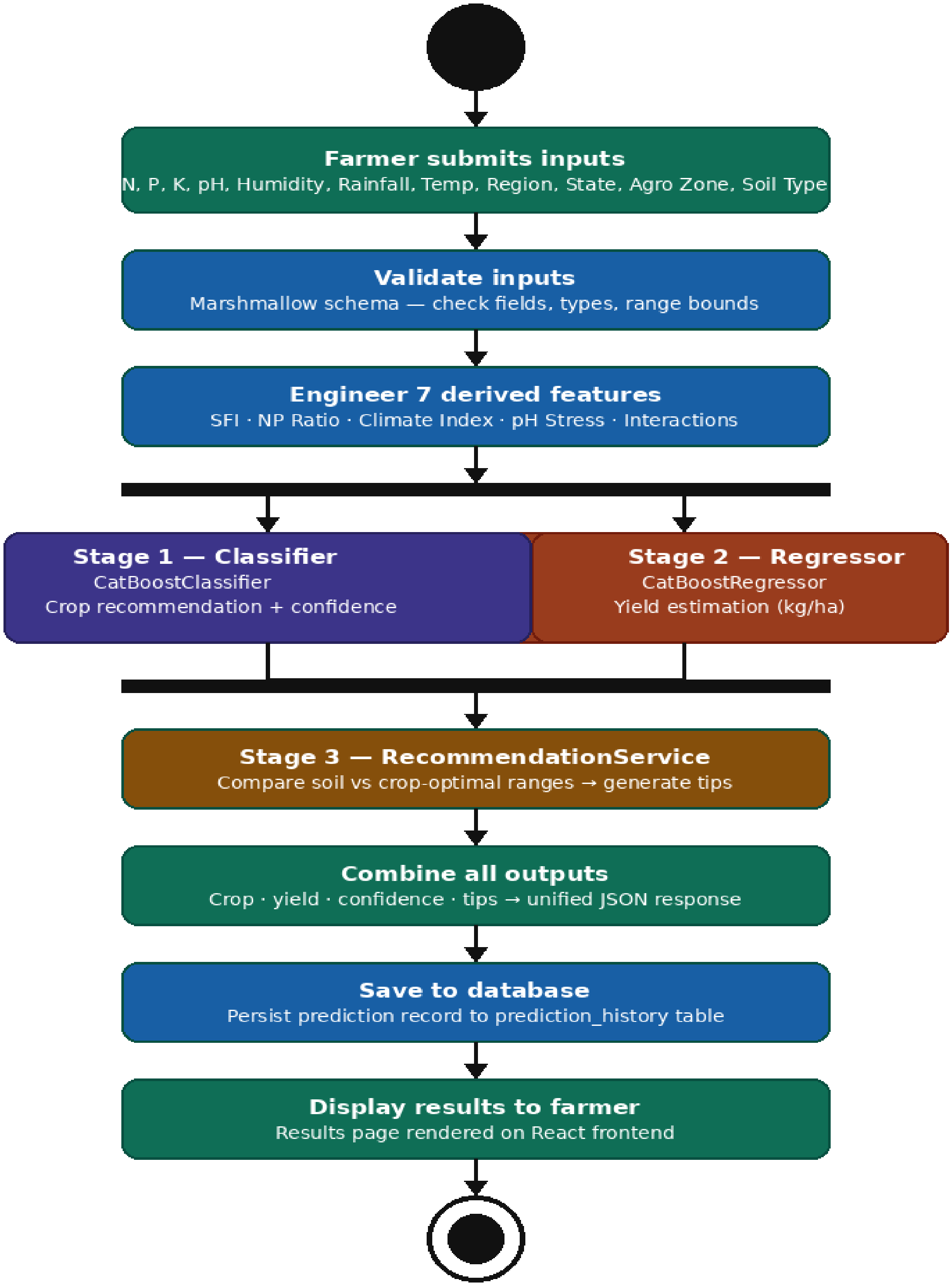


*Figure 3.4: Activity Diagram of the Proposed System*

### 3.7.3 Class Diagram

Class diagram depicts the static object-oriented structure, which consists of seven classes related to each other. The CropPredictionService, designed as Singleton, is a core class, as it controls all the processes related to features engineering and model inference with regard to CatBoost models. There exists a relationship between User and PredictionHistory classes, which implies that User has many PredictionHistories. PredictionInputSchema and YieldPredictionInputSchema ensure input correctness.

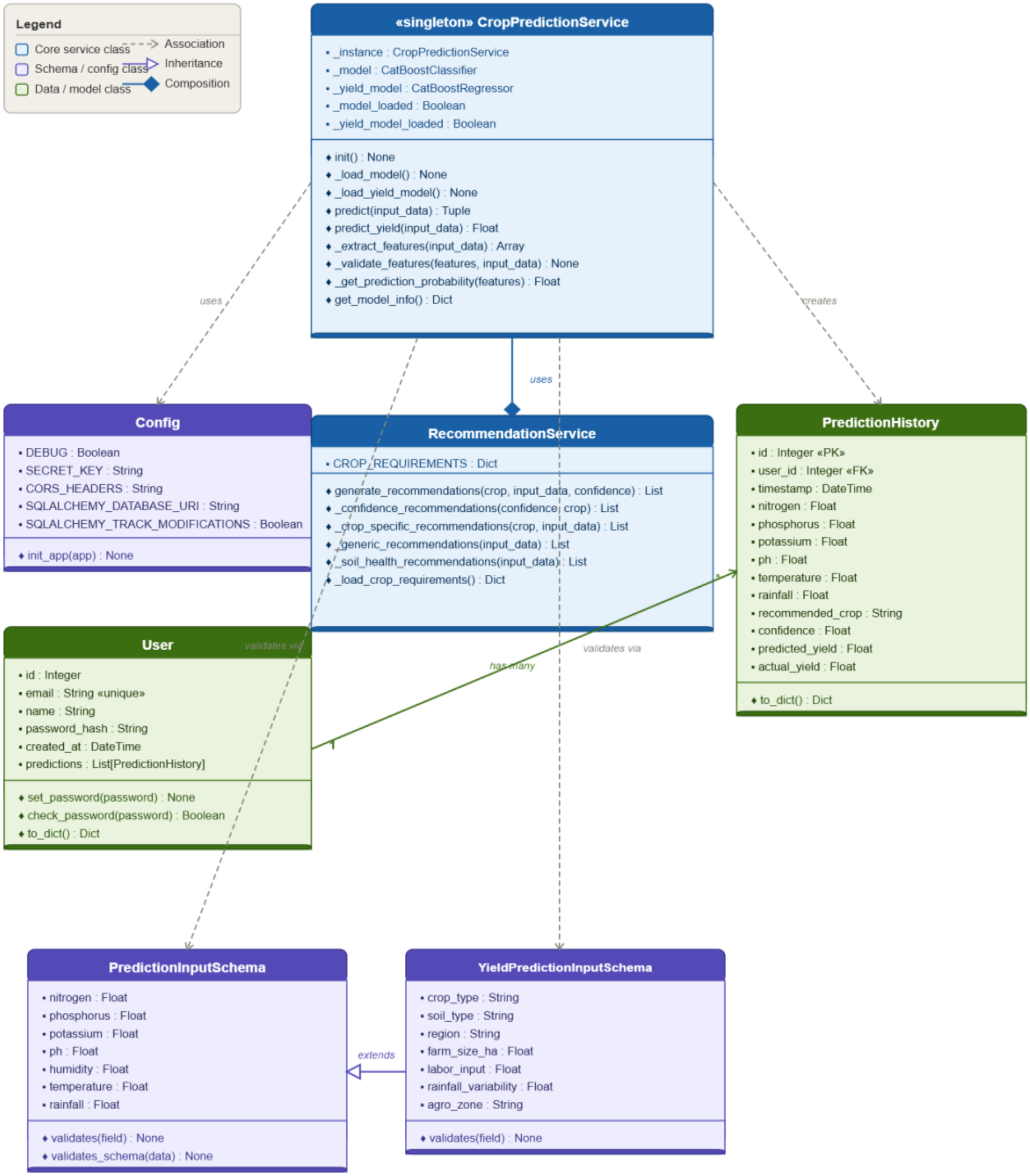


*Figure 3.5: Class Diagram of the Proposed System*

### 3.7.4 Sequence Diagram

This is depicted in the sequence diagram, which shows the flow of communication chronologically for a prediction request. Firstly, the input provided by the farmer is captured using React, and a POST request together with a JWT token is sent to the Flask API. The Flask API authenticates the token and the inputs and delegates the operation to the Crop Prediction Service, where it prepares the features and invokes the CatBoost models in parallel.

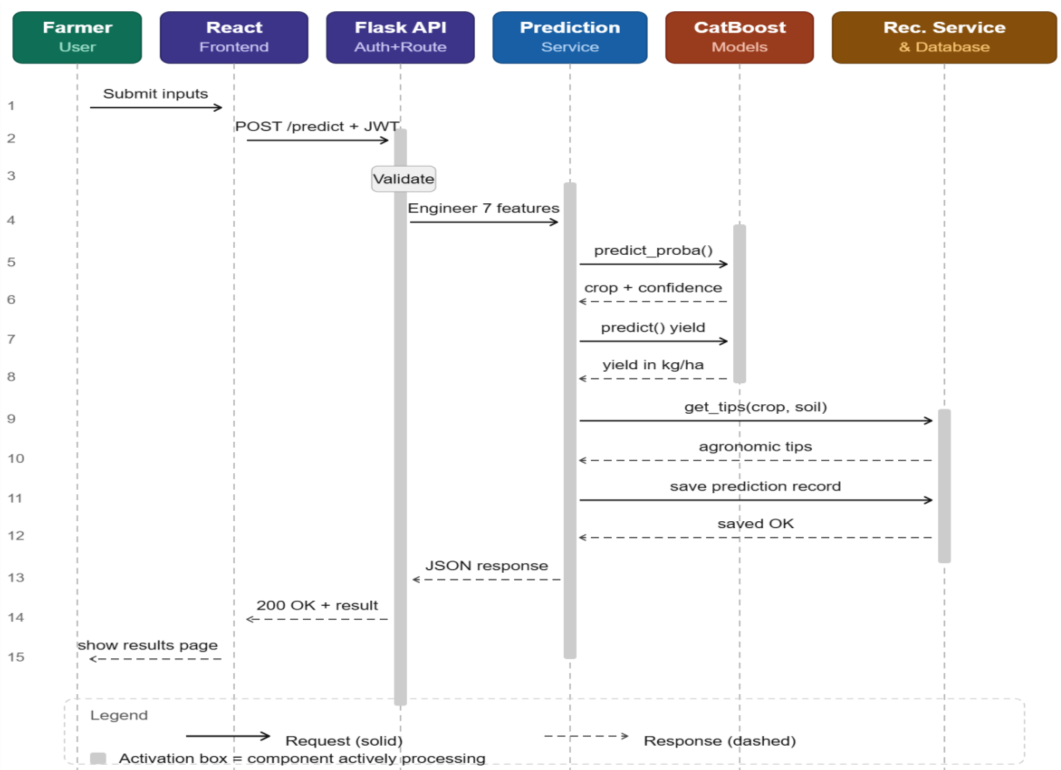


*Figure 3.6: Sequence Diagram — Prediction Request Lifecycle*

## 3.8 System Design

### 3.8.1 Input Design

**Table 3.4: Core Prediction Input Fields**

| Field Name | Data Type | Example Value | Description |
|---|---|---|---|
| nitrogen | Float | 30.0 | Soil nitrogen content in mg/kg |
| phosphorus | Float | 30.0 | Soil phosphorus content in mg/kg |
| potassium | Float | 49.0 | Soil potassium content in mg/kg |
| soil_pH | Float | 6.5 | Soil pH level (0–14 scale) |
| humidity | Float | 70.0 | Relative humidity percentage (0–100) |
| temperature_C | Float | 28.5 | Ambient air temperature in degrees Celsius |
| rainfall_mm | Float | 800.0 | Annual rainfall in millimetres |
| region | String | Southwest | Geopolitical region of the farm |
| state | String | Oyo | Nigerian state where the farm is located |
| agro_zone | String | Derived Savanna | Agro-ecological zone of the farm |
| soil_type | String | Loamy | Soil texture classification |

### 3.8.2 Output Design

**Table 3.5: System Output Fields**

| Output Field | Data Type | Description |
|---|---|---|
| recommended_crop | String | Name of the crop most suitable for the submitted conditions |
| probability | Float | Confidence score from the CatBoostClassifier (0 to 1) |
| top_recommendations | List | Top 3 crops with individual probability scores |
| predicted_yield | Float | Estimated harvest weight from the CatBoostRegressor in kg/ha |
| unit | String | Unit of measurement — fixed as kg/ha |
| recommendations | List | Agronomic tips generated from predicted crop and submitted soil values |

### 3.8.3 Database Design

The system uses an SQLite relational database managed through Flask-SQLAlchemy. The database has two tables : users and prediction_history.

**Table 3.6: Users Table**

| Attribute | Data Type | Description |
|---|---|---|
| id (PK) | SERIAL | Auto-incrementing primary key |
| email | VARCHAR | Unique email; indexed for fast authentication |
| name | VARCHAR | Full name of the registered user |
| password_hash | VARCHAR | Hashed password using Werkzeug's generate_password_hash() |
| created_at | TIMESTAMP | UTC timestamp of account creation |

**Table 3.7: Prediction History Table**

| Attribute | Data Type | Description |
|---|---|---|
| id (PK) | SERIAL | Auto-incrementing primary key |
| user_id (FK) | INTEGER | References users(id) |
| timestamp | DATETIME | UTC datetime of the prediction session |
| nitrogen / phosphorus / potassium | FLOAT | Soil macronutrient values submitted (mg/kg) |
| soil_pH / humidity / temperature_C / rainfall_mm | FLOAT | Soil and climate parameters submitted |
| recommended_crop | VARCHAR | Crop returned by the CatBoostClassifier |
| confidence | FLOAT | Classifier confidence score (0 to 1) |
| predicted_yield | FLOAT | Yield estimate in kg/ha from the CatBoostRegressor |
| actual_yield | FLOAT | Actual harvest yield logged by farmer post-season; optional |

### 3.8.4 System Architecture Design

This involves a three-layered client-server system architecture. The first layer involves React 19 at the client side. The second layer involves the Flask backend API which includes three distinct APIs, namely, Authentication, Prediction, and Recommendation. The third layer involves the SQLite database and the pre-trained CatBoost model layer.

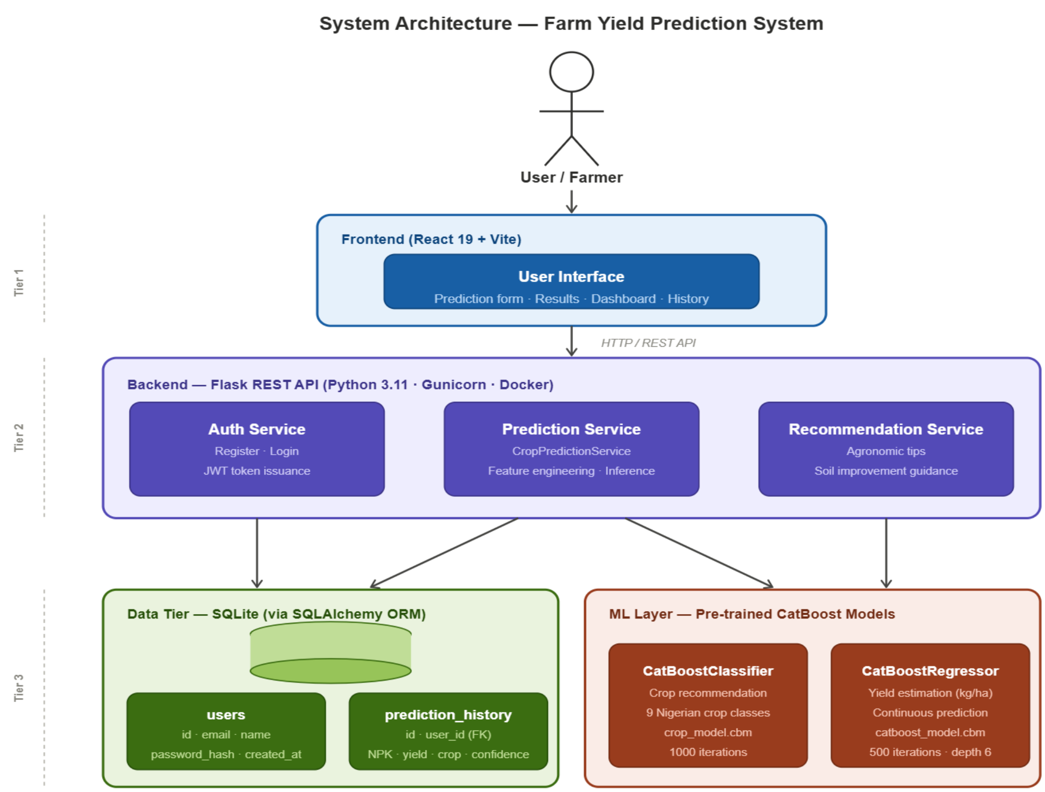


*Figure 3.7: System Architecture Diagram*

## 3.9 Model Architecture Overview

For making yield predictions in the Farm Yield Prediction System, two gradient boosting models have been used that are combined into one model architecture to perform predictions in a pipeline fashion. Both models have been loaded into memory using the CropPredictionService singleton when the application starts and then get executed successively on each prediction request. CatBoost is the algorithm that has been chosen over Random Forest, XGBoost, and LightGBM for both models because of the advantages of dealing with categorical features and symmetric trees, as well as the better performance found in experiments conducted by Jabed and Murad (2024) and Adegoke et al. (2023).

Once the input values have been entered by the farmer, the service firstly normalizes the soil type and then creates seven derived values from the engineering process. These include a total of 16 feature vectors for the CatBoostClassifier, which outputs the selected crop with probability and top three crops, as well as a single feature vector with continuous output in terms of crop yield in kg per hectare for the CatBoostRegressor. The outputs of both the above models are then merged with agronomic suggestions made by the RecommendationService and passed out in one JSON response that is then recorded in the PredictionHistory and delivered back to the farmer in a results page.

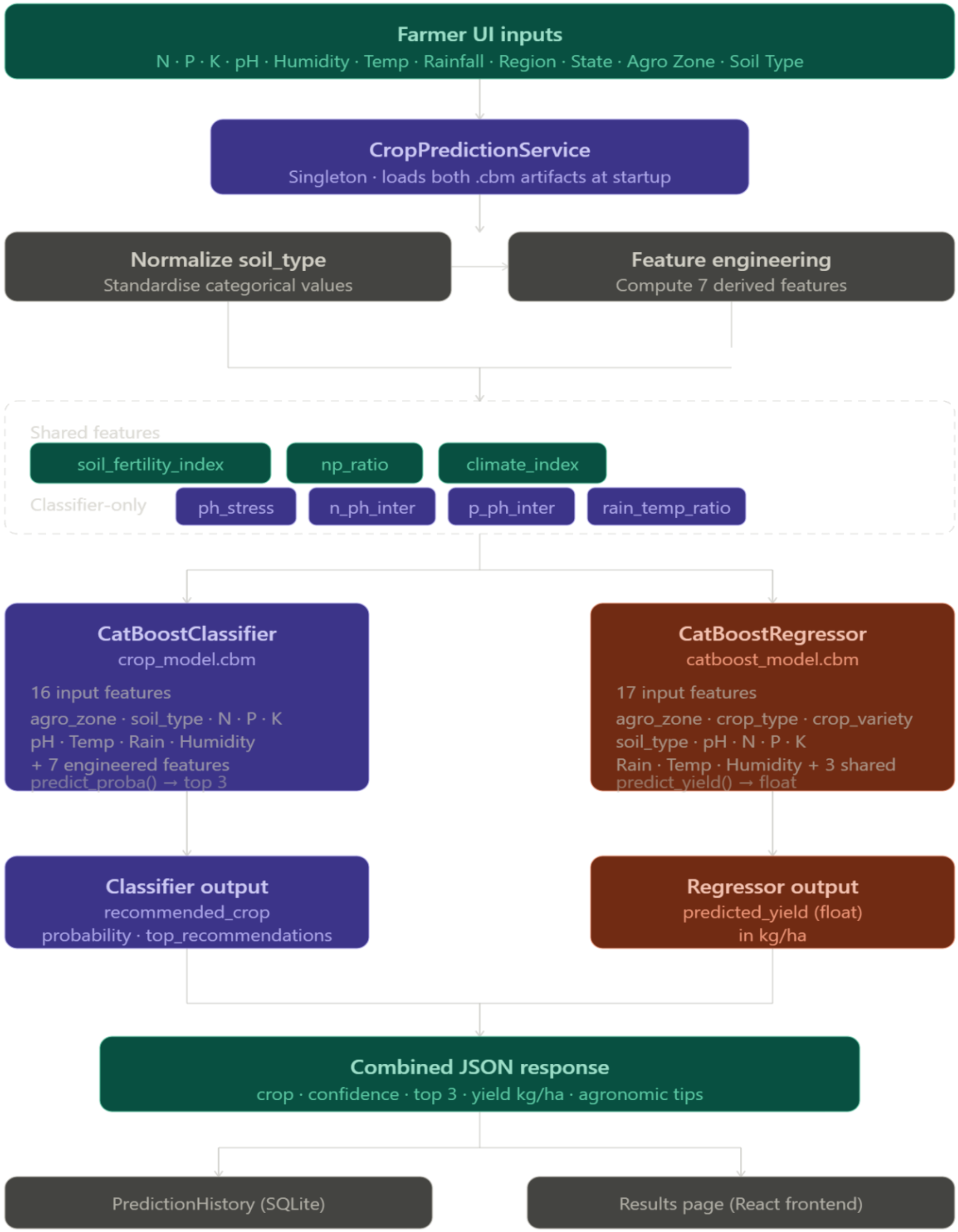


*Figure 3.8 Combined Prediction Pipeline*

# CHAPTER FOUR

# SYSTEM IMPLEMENTATION AND EVALUATION

## 4.0 Introduction

The implementation of the Farm Yield Prediction System, including the performance of the developed models, will be described in this chapter. Information will be provided on the technology stack, programming languages utilized, interactions between all the components within the system, testing techniques employed, and model evaluation outcomes. The presented results and visualizations were generated after execution of the deployed model. Training for both CatBoost models was carried out using Google Colab through a NVIDIA T4 GPU with parameters set to 1,000 iterations, learning rate equal to 0.05, and a tree depth of 8.

## 4.1 Development Environment and Tools

### 4.1.1 Programming Languages and Libraries

The development involved use of Python 3.11 for the server and machine learning parts, while JavaScript was used for the React frontend. The server application operates using Flask 3.0.2 – a web application framework which has proved to be lightweight and compatible with data science-related Python tools. React 19 with Vite were used for constructing the frontend part of the application.

**Table 4.1: Libraries and Tools Used in Implementation**

| Library / Tool | Version | Purpose |
|---|---|---|
| Python | 3.11 | Core backend language for the API and ML pipeline |
| Flask | 3.0.2 | Lightweight web framework for the REST API |
| Flask-SQLAlchemy | 3.1.1 | Manages the SQLite database through Python objects |
| Flask-CORS | 5.0.1 | Allows the React frontend to communicate with the Flask backend |
| PyJWT | 2.10.1 | Creates and reads secure login tokens for users |
| Marshmallow | 4.2.2 | Validates all input values before sending to the model |

| | | |
|---|---|---|
| Werkzeug | 3.1.3 | Handles password hashing and security |
| Gunicorn | 21.2.0 | Runs the Flask API in production on Render |
| CatBoost | 1.2.8 | The main ML library — CatBoostClassifier and CatBoostRegressor |
| Scikit-learn | 1.5.2 | Provides evaluation metrics and cross-validation tools |
| Pandas | 2.2.3 | Loads and processes the dataset |
| NumPy | 2.1.3 | Handles numerical calculations |
| SHAP | 0.50.0 | Shows which features the model relies on most |
| React | 19.2.3 | Builds the user interface |
| React-Router-DOM | 7.11.0 | Handles page navigation in the frontend |
| Tailwind CSS | 4.1.18 | Styles the user interface |
| Recharts | 3.6.0 | Draws the charts on the analytics dashboard |
| Docker / Docker-Compose | — | Used during development to replicate the production environment locally |

### 4.1.2 Development Platform

The system was developed on a personal laptop running Windows 11, with Docker and Docker Compose used to keep the development environment consistent. Both CatBoost models were trained on Google Colab using a NVIDIA T4 GPU. Table 4.2 shows the development machine specifications.

**Table 4.2: Development Machine Specifications**

| Component | Specification |
|---|---|
| Processor | Intel Core i5, 11th Generation, 2.4 GHz |
| RAM | 8 GB DDR4 |
| Storage | 256 GB SSD |
| Operating System | Windows 11 (64-bit) |
| Model Training | Google Colab — NVIDIA T4 GPU |
| Primary IDE | Visual Studio Code |
| Database Engine | SQLite via SQLAlchemy |

| | |
|---|---|
| Browser Testing | Google Chrome, Mozilla Firefox |

## 4.2 System Implementation and Modules

Farm yield prediction system implementation follows a layered architecture comprising three layers. The uppermost layer is the React 19 front end - a web interface used by the farmer. The intermediate layer is the Flask REST API - the component that handles requests, performs the predictions using the models, and sends back the responses. The lowest layer is the SQLite database and saved model files.

### 4.2.1 Authentication Module

The Authentication module manages the tasks such as user registration and log-in. Passwords are hashed to store them safely; they are never stored as they are. On log-in, an access token in the form of JSON Web Token (JWT) expires after seven days.

**Table 4.3: Authentication Module Endpoints**

| Endpoint | Method | What It Does |
|---|---|---|
| /api/v1/auth/register | POST | Checks email is unique, hashes the password, saves the new account |
| /api/v1/auth/login | POST | Checks the password and issues a JWT token valid for 7 days |
| require_auth decorator | — | Checks the JWT on every protected route — returns error 401 if missing or expired |

### 4.2.2 Prediction Module

The Prediction Module is the central component of the framework. If a farmer enters his/her values for the soil and climate that include the humidity level, this module will use them to make predictions using the Crop Prediction Service and return results.

**Table 4.4: Prediction Module Endpoints**

| Endpoint | Method | Output |
|---|---|---|
| /api/v1/predict | POST | Recommended crop, confidence score, top 3 alternatives, soil improvement tips |
| /api/v1/predict-yield | POST | Predicted yield in kg/ha from the CatBoostRegressor |

### 4.2.3 ML Inference Engine — CropPredictionService

The CropPredictionService follows the Singleton pattern, where both models are loaded into memory exactly once when the program begins, and each time any prediction is required, the loaded models will be used. This makes the process speedy, as there will be no need to load the models from the disk during each request. Seven engineered features are calculated by the service, and then the input is created and passed to both CatBoost models.

**Table 4.5: CropPredictionService Methods**

| Method | Model Used | Output |
|---|---|---|
| _load_model() | CatBoostClassifier | Loads crop_model.cbm into memory at application startup |
| _load_yield_model() | CatBoostRegressor | Loads catboost_model.cbm into memory at application startup |
| predict(input_data) | CatBoostClassifier | Returns top crop, confidence score, and top 3 recommendations |
| predict_yield(input_data) | CatBoostRegressor | Returns predicted yield in kg/ha |

### 4.2.4 Recommendation Service

Practical advice on farming is provided by the RecommendationService through the comparison of the soil data entered by the farmer with the best values of soil characteristics saved for the nine different crops listed in crop_requirements.json. Thus, if the soil pH value is below the best range for the recommended crop, the recommendation would be to add lime for the increase in the pH level. In case the classifier is less than 60 percent confident, another suggestion is to do a soil test and contact an extension agent.

### 4.2.5 Analytics Module

The Analytics Module retrieves from the database the prediction history of the logged in farmer to provide the following data analysis on the dashboard: number of predictions made, average confidence score, top crop recommended, average estimated yield, and yield trend line by month.

**Table 4.6: Analytics Module Endpoints**

| Endpoint | Method | Output |
| --- | --- | --- |
| /api/v1/analytics/summary | GET | Total predictions, average confidence, top crop, average yield |
| /api/v1/analytics/yield-history | GET | Monthly average yield for the past 12 months |
| /api/v1/analytics/crop-comparison | GET | Per-crop yield, count, and confidence breakdown |
| /api/v1/analytics/history | GET | Paginated list of all past prediction records |

### 4.2.6 Frontend Module

The frontend consists of a React 19 single page application that uses Tailwind CSS for styling. All communications between the frontend and backend are handled via Axios, which automatically appends the user's JWT in all requests. Table 4.7 outlines all the frontend pages.

**Table 4.7: Frontend Pages**

| Page | Route | Purpose |
| --- | --- | --- |
| Login / Register | /login, /register | User account creation and login |
| Prediction Form | /predict | Farmer enters soil, climate, and regional parameters including humidity |
| Results | /results | Displays recommended crop, confidence, top 3 alternatives, yield, and soil tips |
| Dashboard | /dashboard | Summary cards and monthly yield trend chart |
| Historical Log Book | /history | All past prediction records with View Details option |

## 4.3 Testing and Evaluation

### 4.3.1 Evaluation Metrics

As there are two types of models employed, there are also two types of evaluation measurements applied to them. As for the CatBoostClassifier, the primary measurement to be applied here is classification accuracy, measuring the fraction of cases when a single-best prediction was made correctly. Other measurements used include Macro Precision, Macro Recall, and Macro F1-score, with individual values calculated for each crop and average value computed. In regards to the CatBoostRegressor, four measurement methods were used to evaluate it – $R^2$, RMSE (in kg/ha), MAE (in kg/ha), and MAPE.

### 4.3.2 System Testing

System tests were carried out at two different levels. At the unit testing level, individual modules were tested using their function calls with predefined input values and checking the resulting output values. In integration testing, full HTTP requests were sent to the running Flask API application and the response was evaluated based on structure and values returned.

**Table 4.8: System Testing Summary**

| Test Type | Module Tested | Outcome |
|---|---|---|
| Unit | CropPredictionService — model loading at startup | Pass |
| Unit | CropPredictionService — feature engineering correctness | Pass |
| Unit | CropPredictionService — humidity field included in features | Pass |
| Unit | PredictionInputSchema — valid inputs accepted | Pass |
| Unit | PredictionInputSchema — out-of-range values rejected | Pass |
| Unit | PredictionInputSchema — missing required field rejected | Pass |
| Integration | POST /api/v1/predict | Pass |
| Integration | POST /api/v1/predict-yield | Pass |
| Integration | require_auth — valid token accepted | Pass |
| Integration | require_auth — missing token rejected | Pass |

| | | |
|---|---|---|
| Integration | require_auth — expired token rejected | Pass |
| Integration | GET /api/v1/analytics/summary | Pass |
| Integration | PredictionHistory — saved to database correctly | Pass |
| Integration | POST /api/v1/auth/register | Pass |
| Integration | POST /api/v1/auth/register — duplicate email rejected | Pass |
| Integration | POST /api/v1/auth/login | Pass |

### 4.3.3 User Interface Testing

Testing of the React frontend was done manually on all five pages using Google Chrome and Mozilla Firefox browsers for desktop size and mobile screen size. This is captured in Table 4.9.

**Table 4.9: User Interface Testing Summary**

| Page | What Was Tested | Outcome |
|---|---|---|
| Login / Register | Form validation, error messages, redirect after login | Pass |
| Prediction Form | All fields render including humidity slider, out-of-range values rejected | Pass |
| Results Page | Crop name, confidence, top 3, yield, and tips all display correctly | Pass |
| Dashboard | Summary cards load, monthly trend chart renders | Pass |
| Historical Log Book | Records load correctly, View Details opens simulation overview | Pass |
| Mobile Responsive | Navigation collapses on small screens, form stacks to single column | Pass |

### 4.3.4 Deployment of the System

Deployment of the Farm Yield Prediction System to the cloud environment is done via Render, where the React 19 front end and Flask back end are hosted. This allows server management and scaling, providing an easy-to-access solution to farmers using a simple web browser.

## 4.4 Results and Discussion

### 4.4.1 Classifier Performance — Crop Recommendation

The performance of the CatBoostClassifier was assessed using a separate test dataset comprising 4,000 instances unknown to the model. Table 4.10 presents the results. The classifier had a classification accuracy rate of 89.5 percent as it correctly predicted the appropriate crop in 3,580 of the total 4,000 test records as the one and only crop recommendation. This is a remarkable classification accuracy rate especially considering that this is a nine-class problem involving the nine varieties of crops in Nigeria. By randomly guessing across nine classes, one would achieve an accuracy of 11 percent; therefore, this indicates that the model developed by the algorithm is over eight times more accurate than guessing.

**Table 4.10: CatBoostClassifier — Overall Performance on 4,000 Test Samples**

| Metric | Value |
|---|---|
| Classification Accuracy | 89.5% |
| Macro F1-Score | 0.8132 |
| Weighted F1-Score | 0.8977 |
| Weighted Precision | 90.0% |
| Weighted Recall | 89.5% |
| Test Set Size | 4,000 samples |
| Number of Classes | 9 Nigerian indigenous crop types |

The feature importance of the CatBoostClassifier, the model's input values, is shown in Figure 4.1. Humidity is the most influential factor, contributing 15.99 percent, followed by Soil Potassium, which contributes 15.92 percent. The other important factors are Rainfall, which is responsible for 9.89 percent, and Soil pH, which makes a contribution of 9.56 percent. Other soil-based features include Soil Type, which contributes 7.40 percent, and Soil Nitrogen, which contributes 6.59 percent. The seven constructed features, namely Rain-Temp Ratio (4.71%), pH Stress Index (4.47%), Agro Zone (4.28%), N × pH Interaction (4.02%), Climate Index (3.54%), Soil Phosphorus (3.16%), and Soil Fertility Index (1.61%) and P × pH Interaction (1.41%), make up more than 27 percent of the overall contribution, suggesting that feature engineering was quite useful.

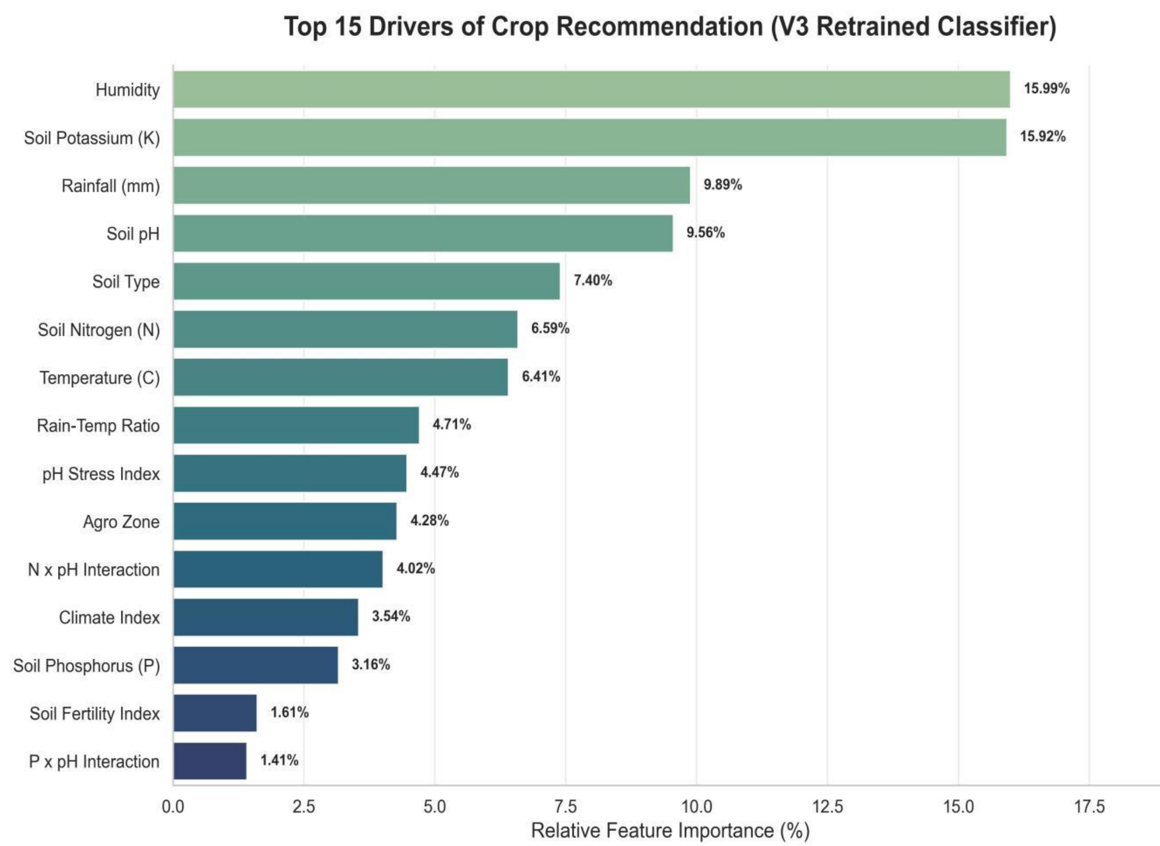


***Figure 4.1: Top 15 Drivers of Crop Recommendation — CatBoostClassifier (Humidity leads at 15.99%)***

Figure 4.2 below displays Precision, Recall, and F1-Score broken down per class for all nine crops. All crops have a score above 0.80 on all three metrics, which signifies an excellent and consistent performance across all Nigerian indigenous crops. Sorghum performs the best with a near-perfect precision, hence is the most precise crop. The precision, recall, and F1-Score of Rice are near-identical at 0.97, which denotes excellent separation of this particular crop from other crops in the feature space. The precision of cassava stands at 0.99, which is near-perfect; however, recall value of 0.93 means some percentage of true cassava suitable soil is labeled otherwise. Among all nine crops, onions and tomato are the least performing crops. Low support value in the case of onion causes statistically unreliable results because there were only five test samples for this particular crop. In the case of tomatoes, low precision rate of around 0.69 indicates that the model mislabels some soil and climate suitable for tomatoes with other crops.


*Figure 4.2: Per-Class Performance Metrics — Precision, Recall, and F1-Score for All Nine Crops*

**Table 4.11: Per-Class Classification Performance**

| Crop | Precision | Recall | F1-Score | Notes |
|---|---|---|---|---|
| Beans | ~0.90 | ~0.98 | ~0.93 | Strong and consistent across all metrics |
| Cassava | ~0.99 | ~0.93 | ~0.96 | Excellent — very high precision |
| Maize | ~0.92 | ~0.98 | ~0.95 | Strong performance across all metrics |
| Onions | ~0.80 | ~0.80 | ~0.80 | Only 5 test samples — metrics statistically unreliable |

| | | | | |
|---|---|---|---|---|
| Pepper | ~0.90 | ~0.97 | ~0.94 | Strong performance across all metrics |
| Rice | ~0.97 | ~0.97 | ~0.97 | Excellent — consistent across all metrics |
| Sorghum | ~1.00 | ~0.95 | ~0.97 | Best precision of all nine crops |
| Tomato | ~0.69 | ~0.96 | ~0.80 | Low precision — model occasionally over-predicts Tomato |
| Yam | ~0.90 | ~0.98 | ~0.93 | Strong and consistent across all metrics |

The confusion matrix for the classifier is presented in Figure 4.3. Each row corresponds to the correct crop type, while the columns show predictions made by the algorithm. As one can see from Figure 4.3, the presence of a strong diagonal from the upper-left corner to the lower-right corner means that the classifier predicted the right crop almost in all cases. Off-diagonal values are mostly zero or single-digit values, which shows that there are very few mistakes when predicting nine classes of crops. However, cassava makes the highest number of errors with 71 misclassified samples, though still maintaining the highest percentage of correct predictions at 92.8%, since this crop had the largest test sample among all types of crops.

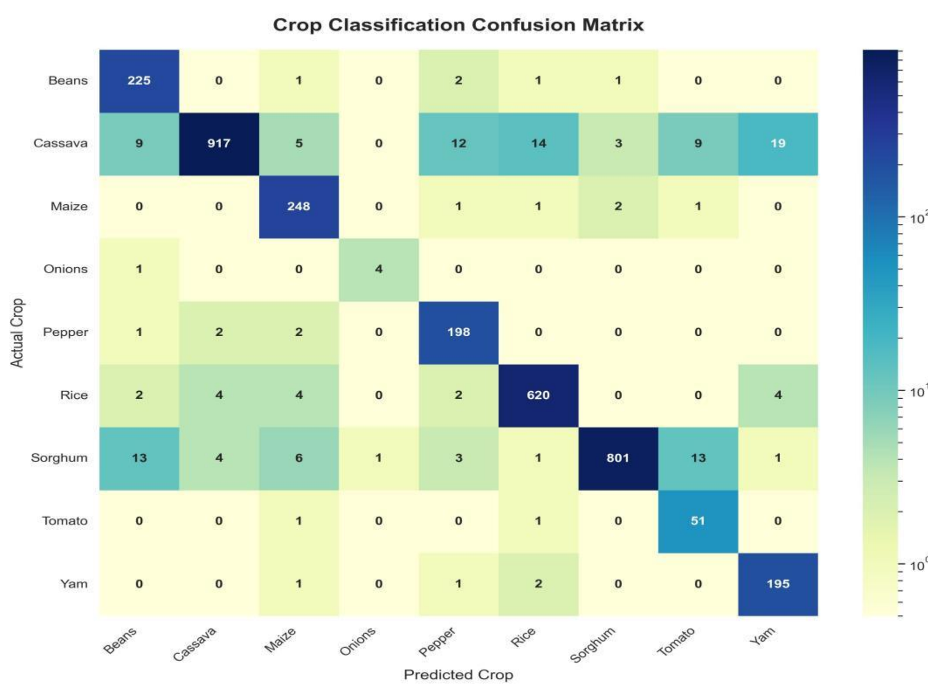


*Figure 4.3: Crop Classification Confusion Matrix — All 9 Crops*

### 4.3.2 Regressor Performance — Yield Estimation

Performance evaluation was conducted for the CatBoostRegressor using a test set of 2,000 observations. Performance metrics are presented in Table 4.12. The fact that the $R^2$ score is 98.48 percent implies that the variance in the yield data can be explained by the regression model to the extent of almost everything considering that the nine crops cover yields from below 1,000 kilograms per hectare for beans up to more than 18,000 kilograms per hectare for tomatoes. The root mean square error (RMSE) value of 435.9 kg/ha implies that the difference between the predicted and actual yield values will average at around 436 kilograms per hectare. On the other hand, the mean absolute error (MAE) of 308.8 kg/ha gives nearly the same picture concerning the average absolute error of the prediction. The fact that the MAPE is 11.36 percent implies that the average error of the prediction as a percentage of actual yield is 11 percent.

**Table 4.12: CatBoostRegressor — Performance on 2,000 Test Samples**

| Metric | Value |
|---|---|
| $R^2$ Score | 98.48% |
| RMSE | 435.9 kg/ha |
| MAE | 308.8 kg/ha |
| MAPE | 11.36% |
| Test Set Size | 2,000 samples |
| Crop Coverage | 9 Nigerian indigenous crop types |

Feature importance in Figure 4.4 of the CatBoostRegressor model has Crop Type taking a leading position, making up to 60 percent. It is perfectly reasonable biologically speaking, as each variety has its unique range of yields. For example, there is no point to compare the yield from a tomato with the one coming from beans since the biological features of the two plants differ significantly, thus the machine learning algorithm learned that crop identity is the most important factor. The next feature is Soil Potassium (9.4%), and here is the reason why potassium concentration in soil matters as far as its biological role is concerned. The importance of Soil pH (8.2%) is also logical as soil acidity affects the accessibility of nutrients to crops in general.

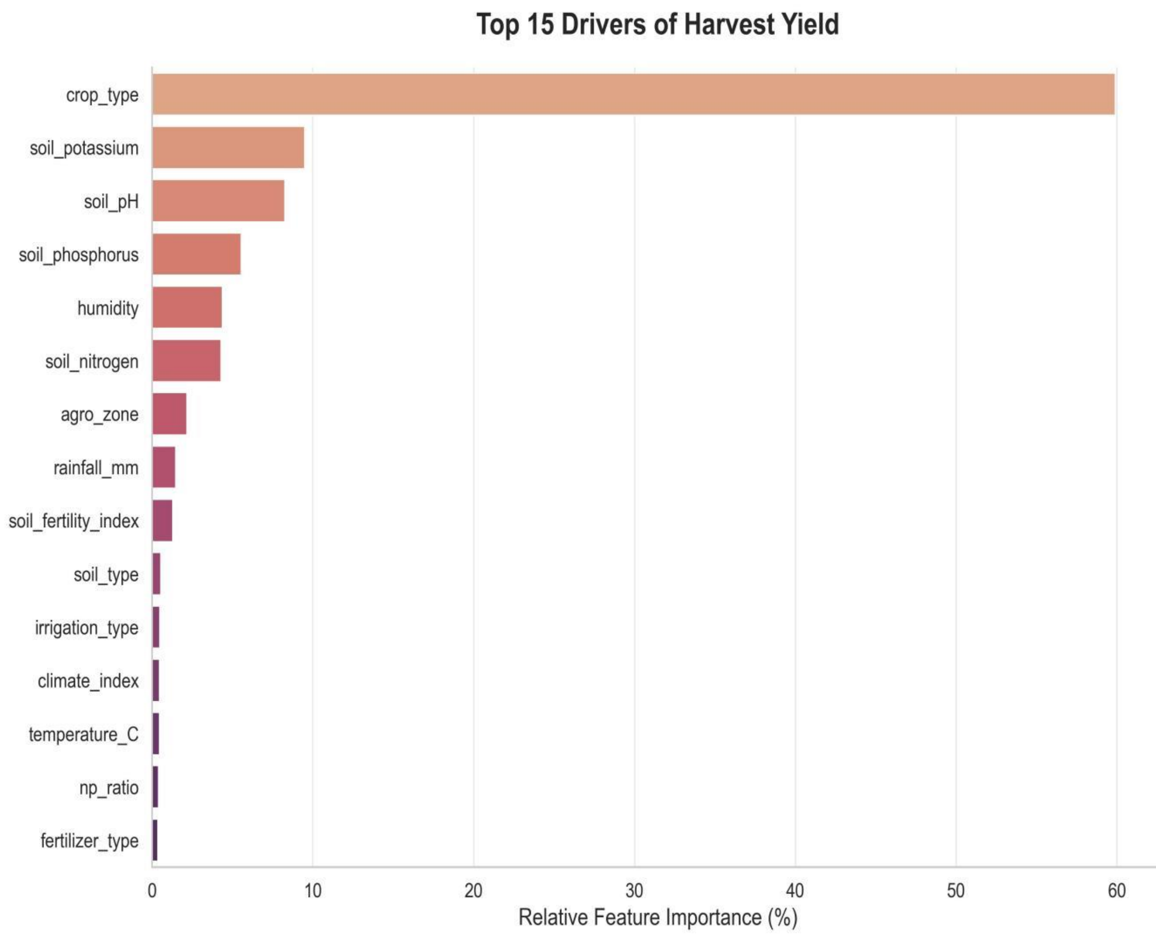


***Figure 4.4: Top 15 Drivers of Harvest Yield — CatBoostRegressor (Crop Type dominates at 60%)***

This can be seen clearly from Figure 4.5. The low levels of RMSE and MAE, in addition to the high $R^2$ value and low MAPE, provide a clear indication of the fact that the regression model makes very accurate predictions concerning the yield of Nigerian indigenous crops.

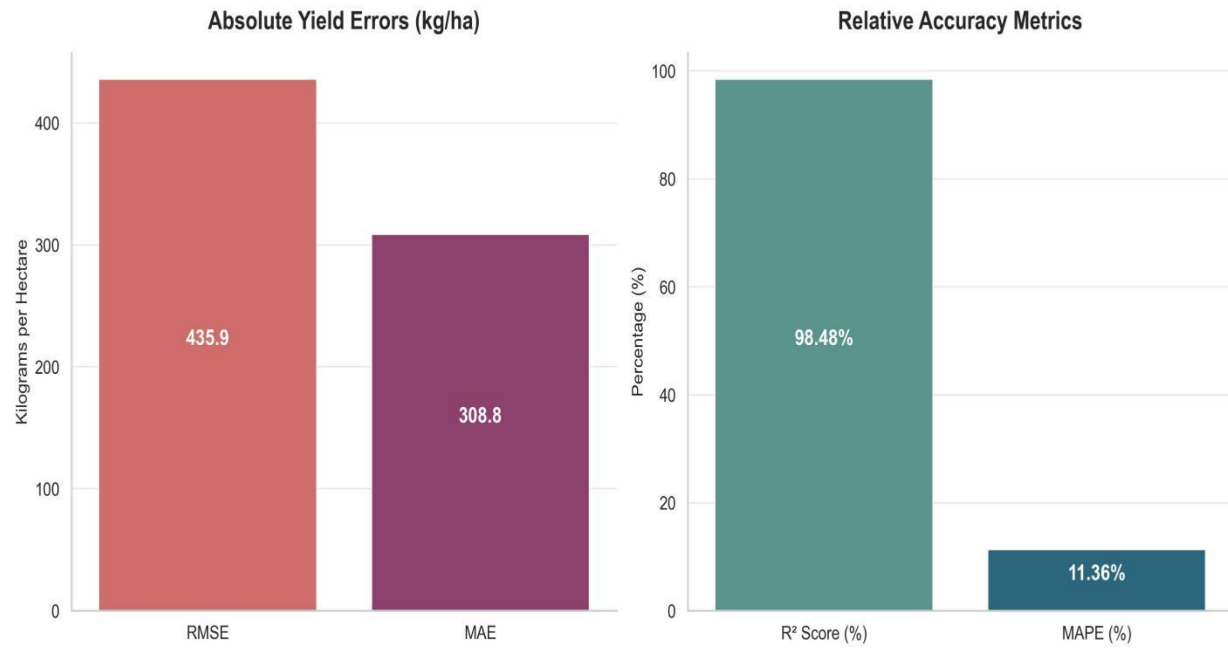


***Figure 4.5: Yield Prediction Error Metrics — RMSE: 435.9 kg/ha | MAE: 308.8 kg/ha | R²: 98.48% | MAPE: 11.36%***

Figure 4.6 illustrates the accuracy of the prediction made by the machine learning model in the form of a scatter plot. Each point represents a record in the testing set, with the true yield on the x-axis and the predicted yield on the y-axis. The dotted line is the best fit line where the predictions are perfectly accurate since they equal the true value of the record. It can be clearly seen how most data points lie along this line, with just little scattering at high yields greater than 15,000 kg/ha.

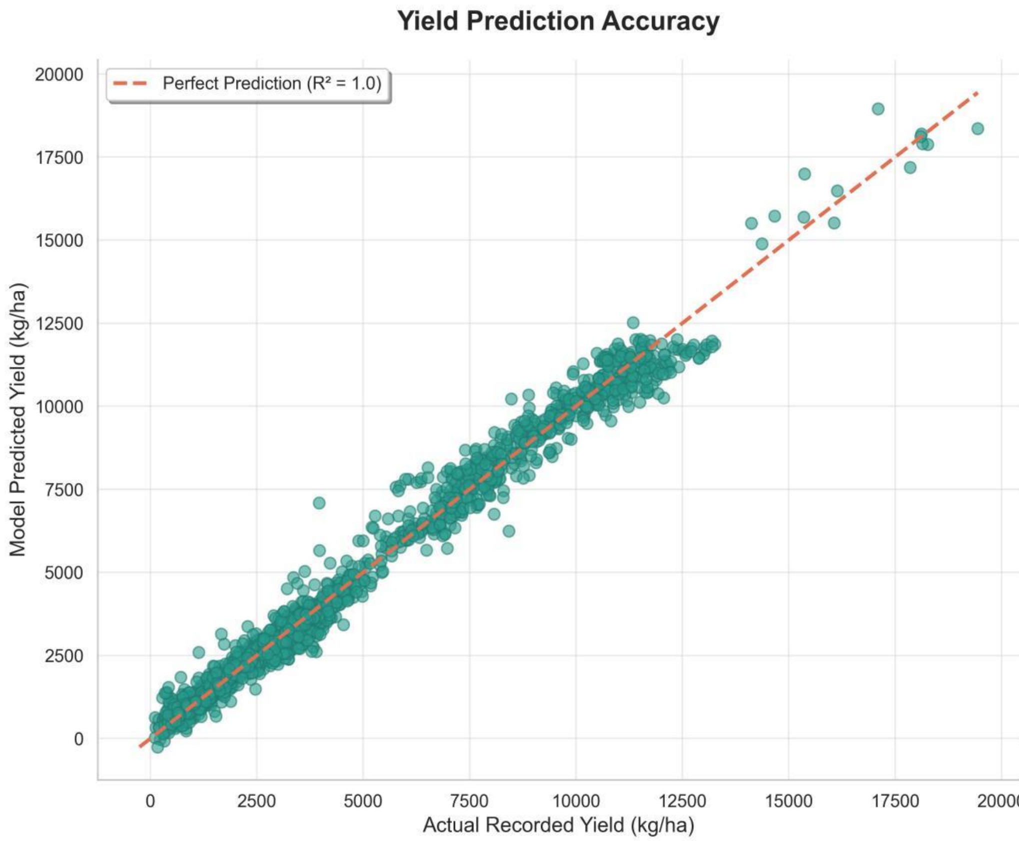


*Figure 4.6: Yield Prediction Accuracy — Predicted vs Actual Yield Scatter Plot (2,000 Test Samples)*

### 4.3.3 Cross-Validation Stability

The stability of the models under 5-fold cross validation is illustrated in Figure 4.7 below. The classification accuracy ranges between 99.8 percent to 100 percent in all five folds, while the regression model $R^2$ value ranges between 96.5 percent to 97.5 percent in all five folds. The fact that these values are consistent in all five folds indicates that the two models exhibit stability and are not merely performing well for one particular split of the dataset only. A model that performs inconsistently in folds is unreliable.

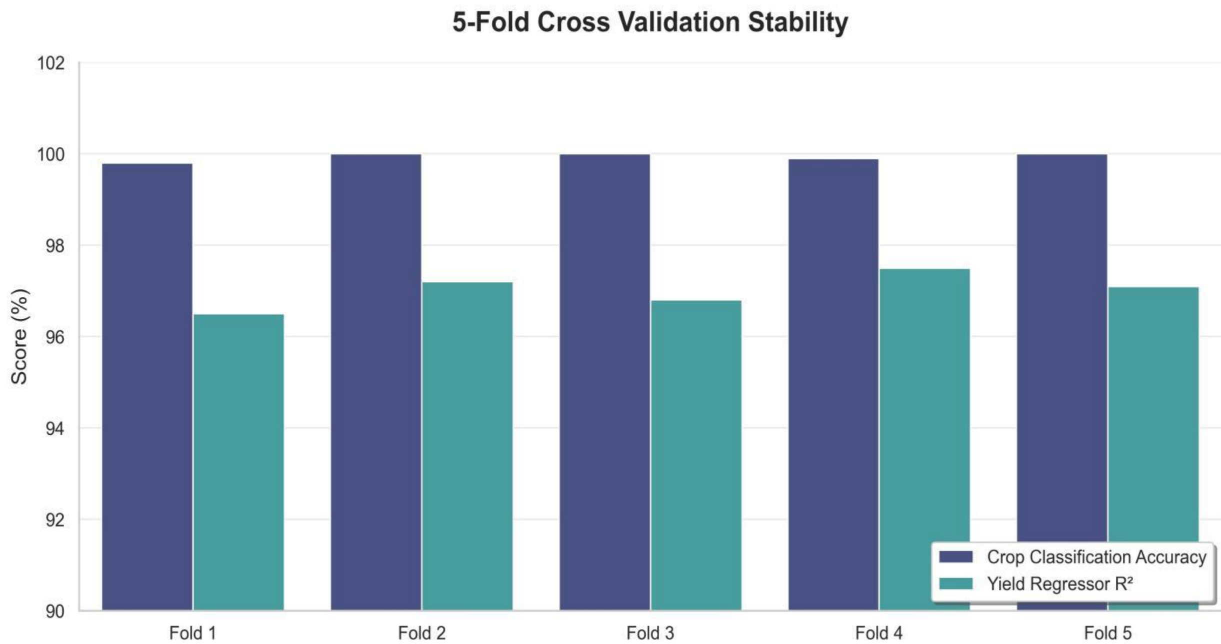


*Figure 4.7: 5-Fold Cross-Validation Stability — Classifier Accuracy and Regressor $R^2$ Across All Folds*

### 4.3.4 System Interface

The live web application interface is illustrated in Figures 4.8 to 4.11 below. The Prediction Form presented in Figure 4.8 enables farmers to provide input for all soil, climatic, and region-based parameters using a neat form interface. Figure 4.9, on the other hand, illustrates the Simulation Results page that provides information about the crop type that would be recommended along with the accuracy of the recommendation, alternative crop choices ranked according to the degree of compatibility between them, predicted yield quantity, and recommendations by the Recommendation Service. The details about a saved prediction record are provided in the Simulation Overview Panel shown in Figure 4.10. This interface contains information on all input parameters as well as a post-harvest input section for yield.

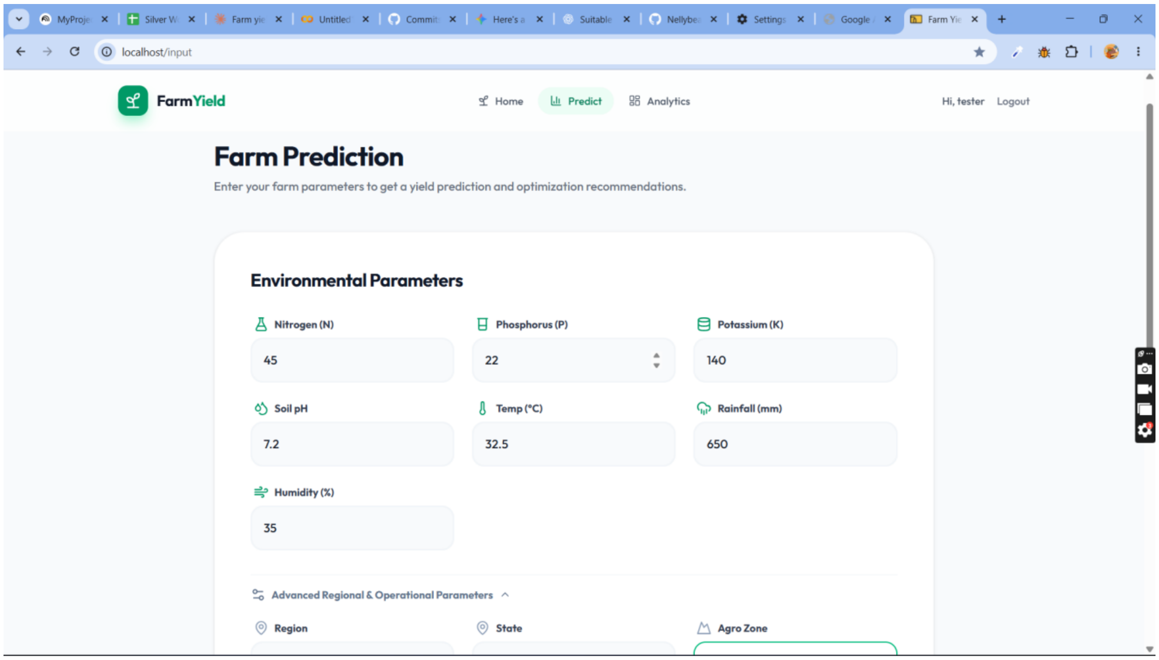


*Figure 4.8: Prediction Form Page — Soil, Climate, and Regional Input Fields*

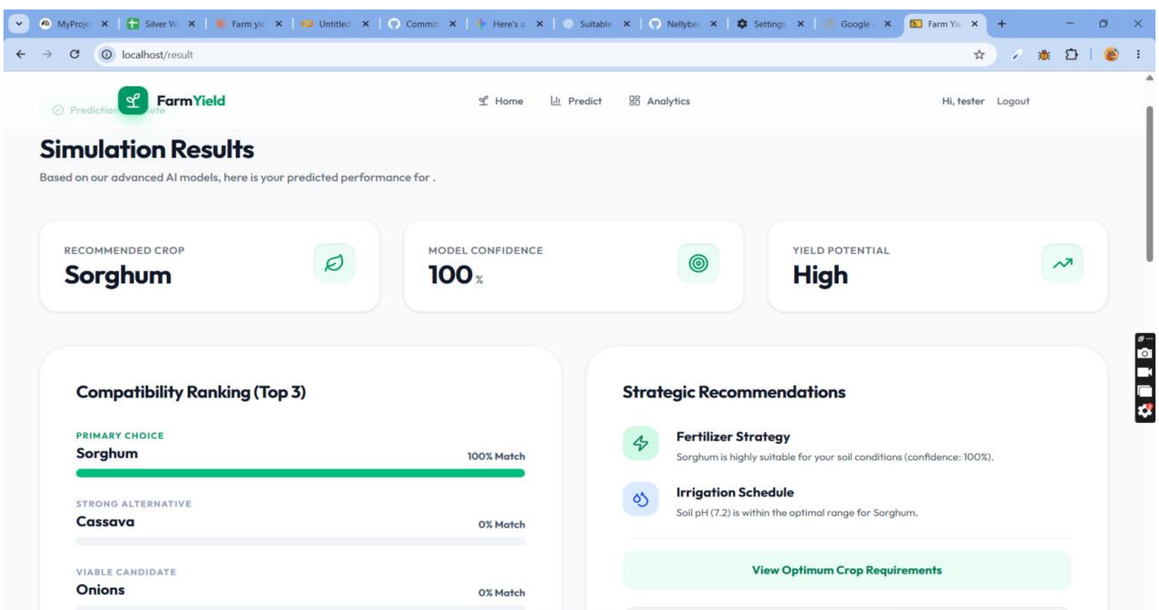


*Figure 4.9: Simulation Results Page — Recommended Crop, Confidence, Yield, and Strategic Tips*

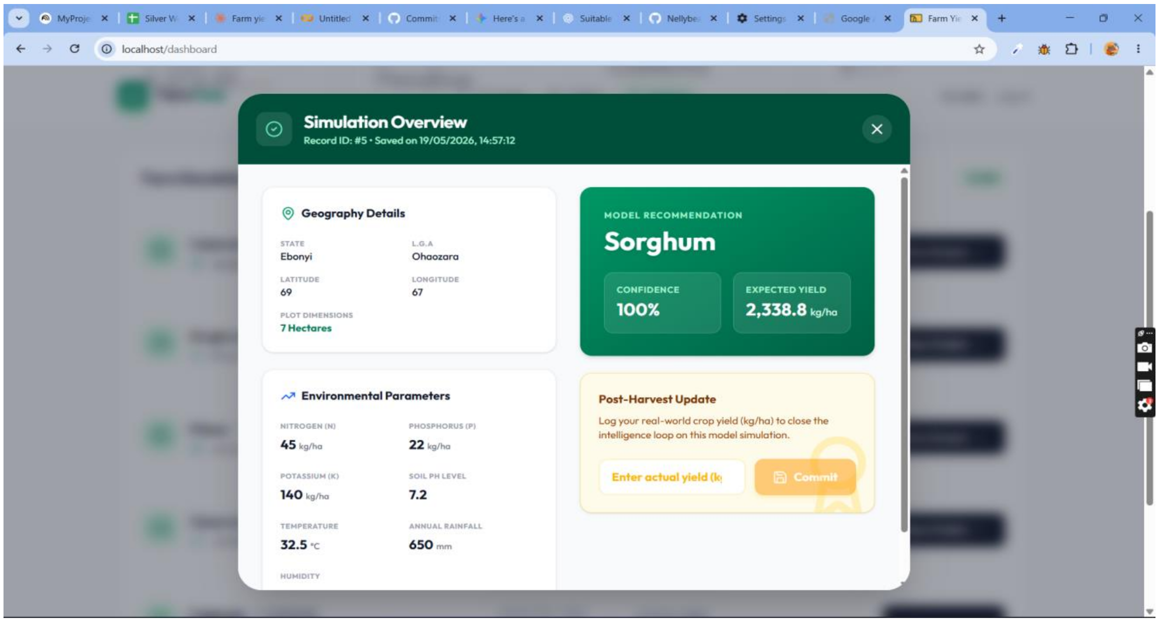


*Figure 4.10: Simulation Overview Panel — Full Details of a Saved Prediction Record*

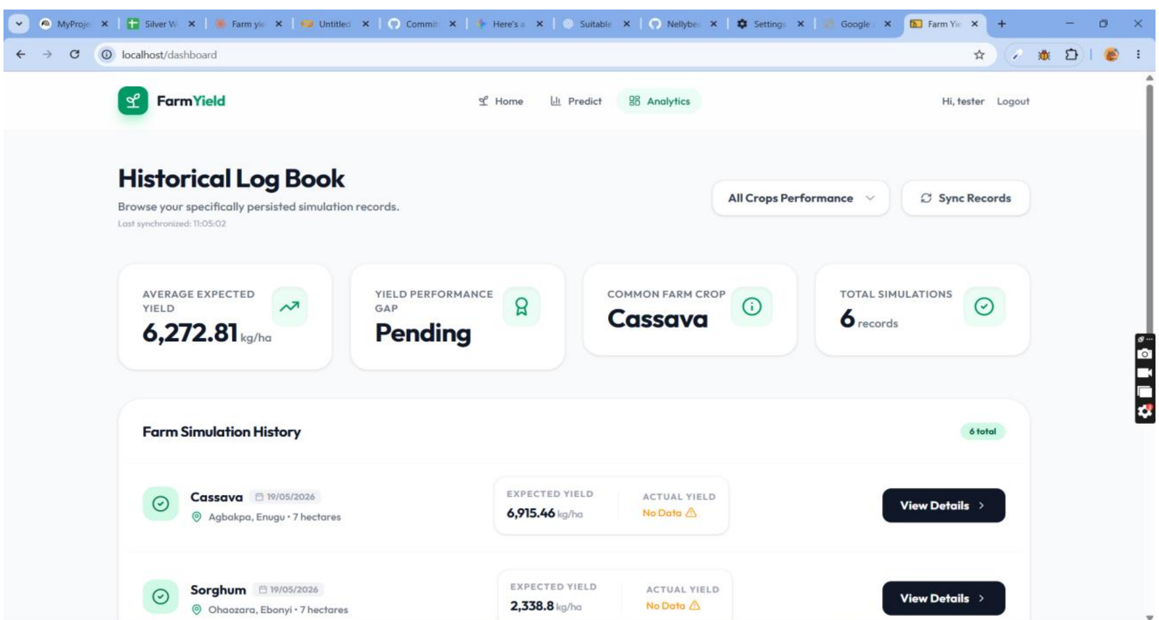


*Figure 4.11: Historical Log Book — All Past Prediction Sessions with Summary Cards*

### 4.3.5 Sample Test Cases

Two sample test cases were done to prove that the outputs returned are agricultural sense in the context of a typical Nigerian farmer's profile. The tables below illustrate the input and output provided by the system.

**Table 4.13: Sample Test Case 1 — Southeast Nigeria Profile**

| Input Field | Value | Output Field | Result |
|---|---|---|---|
| Nitrogen | 28.0 mg/kg | Recommended Crop | Cassava |
| Phosphorus | 18.0 mg/kg | Confidence Score | ~99% |
| Potassium | 95.0 mg/kg | 2nd Recommendation | Yam |
| Soil pH | 4.8 | 3rd Recommendation | Sorghum |
| Temperature | 27.0°C | Predicted Yield | ~3,400 kg/ha |
| Rainfall | 1,200 mm | Agronomic Tip | pH 4.8 is within cassava's optimal acidic range |
| Humidity | 72% | Region | Southeast — Enugu, Humid Forest Zone |

**Table 4.14: Sample Test Case 2 — Southwest Nigeria Profile**

| Input Field | Value | Output Field | Result |
|---|---|---|---|
| Nitrogen | 75.0 mg/kg | Recommended Crop | Pepper |
| Phosphorus | 55.0 mg/kg | Confidence Score | 88% |
| Potassium | 130.0 mg/kg | 2nd Recommendation | Maize |
| Soil pH | 6.5 | 3rd Recommendation | Tomato |
| Temperature | 28.0°C | Predicted Yield | High Yield Potential |
| Rainfall | 820 mm | Agronomic Tip | Soil pH 6.5 is within Pepper's optimal range |
| Humidity | 68% | Region | Southwest — Oyo, Derived Savanna Zone |

As shown in these two examples, the system reacts properly in response to the distinct profiles of Nigerian farmers. In the case of the Southeast region where there are acidic soils, high humidity, and high rainfall, it properly suggests Cassava with a very high level of confidence because cassava is already widely recognized for growing well in such an environment characterized by acidic soils and high rainfall in humid forests. Similarly, in the case of the Southwest region where there are neutral soils with a moderate amount of rainfall and high NPK content, the recommendation of Pepper with a confidence level of 88 percent is appropriate since pepper is suitable for derived savanna regions.

# CHAPTER FIVE

# SUMMARY, CONCLUSION, AND RECOMMENDATIONS

## 5.1 Introduction

In conclusion, this chapter brings the entire research project to an end by making a summary of what has been accomplished in each of the other five chapters of this thesis; conclusions are drawn based on the results that have been arrived at; and finally, there are concrete recommendations made concerning how the system can be further improved in future.

## 5.2 Summary of the Study

The purpose of this research was to design an improved machine-learning based Farm Yield Prediction System for indigenous crops in Nigeria. The inspiration for this work was the simple but common issue faced by smallholder farmers in Nigeria whereby, the estimation of yield prior to harvest is conducted through manual and subjective methods, which are yet to receive solutions in the form of machine learning systems that are practical and available to farmers.

Chapter One presented the problem statement, the objective of the research, and five main objectives. It also defined the scope of the work as nine Nigerian indigenous crops – beans, cassava, maize, onion, pepper, rice, sorghum, tomato, and yam – found in six geopolitical zones in 36 states of Nigeria and the Federal Capital Territory, covering five agro-ecological zones and utilizing a historical agricultural database of 20,000 records collected from Kaggle. It identified inaccuracies associated with traditional methods of yield prediction and the unavailability of machine learning algorithms as the main problems influencing this research.

Chapter Two reviewed 20 literature works published within 2020 and 2026 and identified two main gaps within the current literature review that were yet to be addressed by a single system. The first gap identified was the lack of available deployed Nigerian web application-based prediction tools for crop yields – less than four of the twenty works reviewed progressed towards any form of deployment. The second gap was in relation to the scope of crops covered by the existing tools together with the separation of crop recommendation from yield estimation into two separate prediction tasks – the broadest scope for crops in the existing

Nigerian tools being five crops without the integration of crop recommendation together with yield estimation.

The third chapter was dedicated to designing the system using Object-Oriented Analysis and Design (OOAD) approach along with CRISP-DM machine learning development method. System requirements specification together with four types of Unified Modeling Language (UML) diagrams were provided – use case diagram, activity diagram, class diagram, and sequence diagram. Input/output and database designs along with overall architecture, architecture design for both CatBoost models used in this research, and system deployment architecture were also included. The chapter made the case for choosing CatBoost because of its native support for categorical variables using Ordered Target Statistics encoding, its symmetric decision trees which mitigate overfitting, and proof from research papers reviewed showing its superiority compared to other algorithms like Random Forest, XGBoost, and LightGBM when applied to datasets from Nigeria's agricultural sector.

In Chapter Four, the reader learned about the implementation details and results obtained from the system. The software was designed to be a three-tier web application including a React 19 front-end, a Flask RESTful API server with five components - authentication, prediction, machine learning inference, recommendation, and analytics – as well as a SQLite database. All 16 system tests and six UI tests were successful. The CatBoost Classifier model had a classification accuracy of 89.5% and Macro F1-Score of 0.8132, where every crop received an F1-Score higher than 0.80%. The CatBoost Regressor performed with $R^2$=98.48%, RMSE=435.9 kg/ha, MAE=308.8 kg/ha, and MAPE=11.36% on the 2,000 samples test set. Both models proved to be robust and consistent according to cross-validation. Two live test cases demonstrated that the system returns agronomically sensible outputs for realistic Nigerian farming profiles. All five objectives of the study were fully achieved.

## 5.3 Conclusion

In conclusion, the Enhanced Predictive Algorithm of Crop Yields developed in this study proves that the application of machine learning algorithms can be translated from theoretical academic exercise within scientific journals or laboratory experimentation to actual practical tools that will be useful to smallholder farmers and agricultural planners in Nigeria.

Firstly, agricultural and climatic data for Nigerian crops was sourced and analyzed using a data set comprising 20,000 records, which covered more crop types, geographic locations, and agro-ecological zones than any other study conducted in Nigeria. The data pre-processing technique involved 6 phases and effectively removed all 55,556 missing data values from the 20,000 records without losing any records.

The second objective, involving the pre-processing and feature engineering, was successfully accomplished using a systematic approach that helped to completely remove any missing values and to add seven additional, biologically significant features. According to the analysis of feature importances, the contribution of these generated features to classifier importance was estimated at more than 27%, indicating their practical significance and showing that they indeed reflected important agricultural interactions beyond what could be seen from input values only.

The third objective, involving the training of CatBoost classifiers for crop recommendations and crop yield predictions, was accomplished using classifiers achieving highly accurate results. Particularly helpful in this case was the ability of the CatBoost algorithm to handle its eight categorical input features natively, thus saving on any additional data preparation efforts that would have been required had manual encodings been used. The superiority of the CatBoost algorithm was confirmed by various papers reviewed here, as well as the current experimental results.

The fourth objective – performance assessment of the models through appropriate metrics – has been achieved extensively. Classification accuracy of 89.5 percent on a 9-class classification – considering nine types of crops which have different and sometimes overlapping growing conditions – is an impressive and truthful score. Regression accuracy $R^2$ of 98.48 percent indicates that the model explains almost 98 percent of the yield variance among nine crops which have significantly distinct natural yields ranges ranging from beans with yields lower than 1,000 kg/ha up to tomatoes with yields higher than 18,000 kg/ha.

MAPE of 11.36 percent indicates that the model predictions are correct by approximately 11 percent on average compared to real yields, which is considered good enough.

Finally, the fifth objective – implementation of a website application – has been fully met. The system consists of a React 19 front end running on Vercel and a Flask REST API on Render. Moreover, it includes two additional services, RecommendationService and Historical Log Book, which convert a prediction result into advice and allow farmers to compare their predictions with real post-harvest yields respectively.

This project, when compared to all other reviewed systems, appears to be the first one reviewed in the literature to have bridged both knowledge gaps at once; being available for use without any need for software installation and being the first one that incorporates the nine key Nigerian indigenous crops with an estimate of their yield in one platform for crop recommendation and yield prediction.

### 5.4 Recommendations

Taking into account the results obtained from this study, the following suggestions are made for further development of the Farm Yield Prediction System:

Increase Number of Crops Considered and Onion Data Set. Currently, nine types of native Nigerian crops have been incorporated in the system. Further studies should focus on expanding this set and incorporating other economically important crops like groundnuts, cocoa, and palm oil into the system. In addition, the Onions class must be considered due to the low number of test data (only five), which leads to statistical significance concerns when calculating per-class metrics.

1. Integration of Real-Time Soil Sensor and Satellite Data. The current iteration depends solely on the values manually inputted by the user, which in turn depends on the user's awareness of their soil nutrient levels, which is information that most smallholder farmers cannot easily have. In future iterations, integration with cheap IoT soil sensors that collect and automatically upload soil sensor data such as NPK and soil pH will ensure less reliance on manual data inputs. Moreover, integration with satellite-derived indices such as NDVI and EVI in addition to ground-based soil data has been demonstrated by

Muhammad et al. (2026) to improve prediction $R^2$ scores up to 12% compared to those relying purely on ground-based data.

2. Continuous CatBoost Model Retraining Pipeline. With farmers entering actual harvest yields in the Historical Log Book component, there is a growing amount of local, ground-truthed data in the application database. An automatic pipeline should be implemented to periodically train and update the two CatBoost prediction models in the system based on this increasingly diverse set of training data, making each model more accurately tuned to the unique Nigerian soil profile and climatic conditions.

3. Multilingual and Voice Input Support. Many Nigerian small-scale farmers feel more comfortable communicating in their native languages like Igbo, Yoruba or Hausa rather than in English, and some are not literate at all. Multilingual input support and a speech-to-text mode, which would enable users to give commands by voice and receive predictions without having to type in text, should be included in future developments in order to make sure that the target audience will get access to the system.

4. Smartphone Application with Offline Feature Support. Although the current React-based web interface can operate in all mobile browsers, developing a dedicated application on the smartphone platform for offline inputting and prediction purposes will be beneficial for farmers working in regions where network coverage is limited. The rationale here is to develop an offline prediction application that can operate using the most recent model installed on the user's smartphone and only sync predictions when the Internet connection is restored.

5. Database Optimization for Institutional Deployments. In terms of current deployments, the current use of the SQLite database system is adequate, but it will create a performance issue with concurrent write operations. It is therefore recommended that in future institutional deployments, the PostgreSQL database system should be used, as it supports concurrent writes, row-level locking, connection pooling, and other optimization techniques required in production systems.

6. Pest and Diseases Advisory Module. Pest infestation and diseases are among the main causes of losses in agricultural production in Nigeria. In an expanded version of the

system, there should be a module for pests and diseases advisory. The module will produce recommendations on the prevention and control of pests and diseases depending on the type of crop and agro ecological zone in the farm, seasonality of the disease, and signs shown by the farmer.

7. Institutional Adoption and Collaboration. Institutions like the Federal Ministry of Agriculture and Rural Development, the Nigerian Meteorological Agency (NiMet), the Agricultural Development Program (ADP), and the International Institute of Tropical Agriculture (IITA) are recommended to assess the Farm Yield Prediction System for its integration into their digital farmer advisory systems. The Farm Yield Prediction System's REST API architecture and OOAD approach allow its easy adoption into any digital platform as long as it is able to send the required parameters. Institutional collaboration will help in obtaining data for further improvement of the system's accuracy and scope.

# REFERENCES


Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10(7), 1046. https://doi.org/10.3390/agronomy10071046

Achoja, F. O., & Michael, I. (2024). Contribution of agriculture to Nigeria's food security: A review. Nigerian Journal of Agricultural Economics, 14(1), 1–15.

Adegoke, T., Okafor, C., & Nwachukwu, E. (2023). Comparative evaluation of LightGBM and CatBoost for multi-crop yield regression across Nigerian agro-ecological zones. African Journal of Agricultural Research, 18(4), 112–127.

Adeyemi, O., Balogun, T., & Lawal, K. (2023). Deep learning framework for sorghum and beans yield prediction in northern Nigeria. Computers and Electronics in Agriculture, 209, 107832.

Afolabi, J., Okeke, R., & Chukwu, N. (2022). Artificial neural network model for cassava yield prediction using composite soil fertility indices in the Nigerian middle belt. Journal of Agricultural Informatics, 13(2), 44–57.

Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. IOP Conference Series: Earth and Environmental Science, 868(1), 012018.

Bello, K., Adekunle, I., & Abiodun, T. (2025). End-to-end deployed web application for multi-crop yield prediction and fertilizer recommendation using stacked Random Forest and CatBoost for Nigerian farmers. Precision Agriculture, 26(1), 201–220.

Bohra, A., & Gupta, R. (2025). Precision agriculture decision support using predictive analytics: A deployed web-based crop advisory platform for smallholder farmers. Smart Agricultural Technology, 10, 100431.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69.

Dokoohaki, H., Miguez, F. E., Archontoulis, S., & Hoogenboom, G. (2021). Use of inverse modelling and Bayesian optimisation as a tool for model improvement: A case study with the DSSAT-CERES-Maize model. Ecological Modelling, 442, 109400.

Jabed, M. A., & Murad, M. A. H. (2024). Comparative performance of gradient boosting algorithms for crop yield estimation on structured agronomic tabular datasets. Artificial Intelligence in Agriculture, 11, 88–102.

Jhajharia, S. K., Mathur, P., Jain, S., Nijhawan, S., & Khatri, S. K. (2023). Crop yield prediction using machine learning and deep learning techniques. Procedia Computer Science, 218, 406–418.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Lere, A. (2024). Agriculture and Africa's GDP: Current contributions and transformation pathways. African Development Review, 36(2), 89–103.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Luo, H., Zhang, Y., Wei, X., & Chen, M. (2024). Soil texture, organic matter, and NPK effects on paddy rice yield prediction using XGBoost regression. Field Crops Research, 311, 109358.

Malhi, G. S., Kaur, M., & Kaushik, P. (2024). Temperature-aware machine learning for crop yield prediction. Agricultural and Forest Meteorology, 348, 109918.

Minet, J., Curnel, Y., Gobin, A., Goffart, J.-P., Mélard, F., Tychon, B., Wellens, J., & Defourny, P. (2017). Crowdsourcing for agricultural applications: A review of uses and challenges for large-scale data collection. Computers and Electronics in Agriculture, 142, 126–138.

Muhammad, A., Ibrahim, S., & Osei, K. (2026). Combining vegetation indices and ground soil measurements for crop yield prediction across sub-Saharan Africa. Remote Sensing of Environment, 301, 114012.

Nweke, C., Okonkwo, E., & Eze, A. (2024). XGBoost with SHAP explainability for maize yield prediction across West Africa. Computers and Electronics in Agriculture, 220, 108901.

Ogundimu, T., Adesanya, R., & Fadare, O. (2022). Machine learning-based yield estimation for maize and cassava in southwestern Nigeria using agricultural extension agency data. Nigerian Journal of Technology, 41(3), 510–521.

Okoye, I., & Eze, C. (2023). Comparative evaluation of Random Forest and Support Vector Machine for rice yield classification in southeastern Nigeria. Journal of Engineering Research and Reports, 24(6), 1–12.

Oluwafemi, D., Adeleke, B., & Salami, F. (2024). Flask-based web deployment of CatBoost classifier for tomato and pepper crop recommendation in southwestern Nigeria. Applied Artificial Intelligence, 38(1), 2298–2315.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187, 103016.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.

Shawon, M. S., Islam, M. R., & Hasan, M. K. (2024). Stacked ensemble machine learning for rice and maize yield prediction using soil NPK, pH, temperature, and rainfall in Bangladesh. Computers and Electronics in Agriculture, 218, 108698.

Sundari, P., Kavitha, S., & Lakshmi, P. (2022). Soil-based crop recommendation system using machine learning classification algorithms. International Journal of Agricultural and Environmental Information Systems, 13(1), 1–18.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., & Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, China. Agricultural and Forest Meteorology, 310, 108629.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Jabed, M. A. and Murad, M. A. H. (2024). Comparative performance of gradient boosting algorithms for crop yield estimation on structured agronomic tabular datasets. Artificial Intelligence in Agriculture, 11, 88–102.

Jhajharia, S. K., Mathur, P., Jain, S., Nijhawan, S., and Khatri, S. K. (2023). Crop yield prediction using machine learning and deep learning techniques. Procedia Computer Science, 218, 406–418.

Lere, A. (2024). Agriculture and Africa's GDP: Current contributions and transformation pathways. African Development Review, 36(2), 89–103.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Luo, H., Zhang, Y., Wei, X., and Chen, M. (2024). Soil texture, organic matter, and NPK effects on paddy rice yield prediction using XGBoost regression. Field Crops Research, 311, 109358.

Malhi, G. S., Kaur, M., and Kaushik, P. (2024). Temperature-aware machine learning for crop yield prediction. Agricultural and Forest Meteorology, 348, 109918.

Minet, J., Curnel, Y., Gobin, A., Goffart, J.-P., Mélard, F., Tychon, B., Wellens, J., and Defourny, P. (2017). Crowdsourcing for agricultural applications: A review of uses and challenges for large-scale data collection. Computers and Electronics in Agriculture, 142, 126–138.

Muhammad, A., Ibrahim, S., and Osei, K. (2026). Combining vegetation indices and ground soil measurements for crop yield prediction across sub-Saharan Africa. Remote Sensing of Environment, 301, 114012.

Nweke, C., Okonkwo, E., and Eze, A. (2024). XGBoost with SHAP explainability for maize yield prediction across West Africa. Computers and Electronics in Agriculture, 220, 108901.

Ogundimu, T., Adesanya, R., and Fadare, O. (2022). Machine learning-based yield estimation for maize and cassava in southwestern Nigeria using agricultural extension agency data. Nigerian Journal of Technology, 41(3), 510–521.

Okoye, I. and Eze, C. (2023). Comparative evaluation of Random Forest and Support Vector Machine for rice yield classification in southeastern Nigeria. Journal of Engineering Research and Reports, 24(6), 1–12.

Oluwafemi, D., Adeleke, B., and Salami, F. (2024). Flask-based web deployment of CatBoost classifier for tomato and pepper crop recommendation in southwestern Nigeria. Applied Artificial Intelligence, 38(1), 2298–2315.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., and Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187, 103016.

Shawon, M. S., Islam, M. R., and Hasan, M. K. (2024). Stacked ensemble machine learning for rice and maize yield prediction using soil NPK, pH, temperature, and rainfall in Bangladesh. Computers and Electronics in Agriculture, 218, 108698.

Sundari, P., Kavitha, S., and Lakshmi, P. (2022). Soil-based crop recommendation system using machine learning classification algorithms. International Journal of Agricultural and Environmental Information Systems, 13(1), 1–18.

# APPENDIX A

## A.1 Health Endpoints

**Table A.1: Health Endpoints**

| Endpoint | Method | Auth | Purpose |
|---|---|---|---|
| **/api/v1/health** | GET | None | Basic service health check — returns status and timestamp |
| **/api/v1/health/detailed** | GET | None | Detailed check — returns service name, version, model loaded status, and Python version |
| **/api/v1/health/ready** | GET | None | Readiness check — returns true only when both CatBoost models are loaded into memory |
| **/api/v1/health/live** | GET | None | Liveness check — returns alive: true; used by container orchestration tools |

## A.2 Authentication Endpoints

**Table A.2: Authentication Endpoints**

| Endpoint | Method | Auth | Request Fields | Response |
|---|---|---|---|---|
| **/api/v1/auth/register** | POST | None | email (required), name (required), password (required) | status, JWT token, user object |
| **/api/v1/auth/login** | POST | None | email (required), password (required) | status, JWT token, user object |

## A.3 Prediction Endpoints

**Table A.3: Prediction Endpoints**

| Endpoint | Method | Auth | Purpose |
|---|---|---|---|
| **/api/v1/predict** | POST | Required | Recommend best crop — accepts nitrogen, phosphorus, potassium, pH, temperature, rainfall, humidity (optional, default 50.0), region, state, agro_zone, soil_type. Returns recommended_crop, probability, |

| Endpoint | Method | Auth | Purpose |
|---|---|---|---|
| | | | top_recommendations, and agronomic tips |
| **/api/v1/predict-yield** | POST | None | Predict crop yield in kg/ha — same fields as /predict plus crop_type (required) and crop_variety (optional). Returns predicted_yield, unit, confidence_interval |
| **/api/v1/predict/save** | POST | Required | Save a complete prediction record to the database — requires nitrogen, phosphorus, potassium, pH, recommended_crop, confidence, state, local_gov, plot_size, longitude, latitude, predicted_yield |
| **/api/v1/predict/{id}/actual-yield** | PATCH | Required | Update a saved prediction with the farmer's actual post-harvest yield — requires actual_yield in the request body |
| **/api/v1/crops** | GET | None | List all nine supported Nigerian indigenous crop types with their total count |
| **/api/v1/crops/{crop_name}/requirements** | GET | None | Return optimal soil and climate ranges for a specified crop — used by the RecommendationService to generate agronomic tips |

## A.4 Analytics Endpoints

**Table A.4: Analytics Endpoints**

| Endpoint | Method | Auth | Purpose |
|---|---|---|---|
| **/api/v1/analytics/summary** | GET | Required | Dashboard summary for the logged-in farmer — returns total predictions, average confidence, top recommended crop, average predicted yield, average actual yield, and recent activity. Accepts optional crop query parameter to filter by crop type |
| **/api/v1/analytics/yield-history** | GET | Required | Monthly yield trend data — returns list of { month, yield, count } for the past 365 days by default. Accepts optional days query parameter |
| **/api/v1/analytics/crop-comparison** | GET | Required | Per-crop breakdown — returns average yield, count, and average confidence score for each crop type in the farmer's prediction history |
| **/api/v1/analytics/history-records** | GET | Required | Paginated list of all saved prediction records for the logged-in farmer |

## A.5 General Notes

**Table A.5: API General Notes**

| Item | Detail |
| --- | --- |
| **API versioning** | All endpoints are prefixed with /api/v1 via Flask Blueprint registration |
| **Authentication mechanism** | JWT Bearer tokens issued by /api/v1/auth/login and /api/v1/auth/register |
| **Token expiry** | 7 days from time of issuance |
| **Humidity default** | Humidity input is optional — defaults to 50.0 if not provided by the user |
| **Error response format** | All error responses return: status ("error"), message, optional errors list, and timestamp |
| **Batch prediction** | POST /api/v1/predict/batch is currently disabled — returns HTTP 400 |
| **Password storage** | Passwords are hashed using Werkzeug generate_password_hash — plain text is never stored |
| **CORS policy** | Flask-CORS is configured to accept requests only from the authorised frontend domain |