People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Mohamed Khider University - Biskra
Faculty of Exact Sciences
Computer Science Department

Order Number: ....

THESIS

In Candidacy for the Degree of
DOCTOR 3rd CYCLE IN COMPUTER SCIENCE
Option: Artificial intelligence and its applications

TITLE

Smart Predictive Agriculture Based on Data Science

Presented by **M'hamed Mancer**

Defended on: 4 December 2025

In front of the jury composed of:

Mr. Khaled Rezeg	Professor at University of Biskra	President
Mr. Labib Sadek Terrissa	Professor at University of Biskra	Supervisor
Mr. Soheyb Ayad	Associate Professor at University of Biskra	Co-Supervisor
Mr. Abdelhak Merizig	Associate Professor at University of Biskra	Examiner
Mr. Abdelouahab Belazoui	Associate Professor at University of Batna 2	Examiner

Academic year : **2024 – 2025**

Abstract

Smart agriculture integrates digital technologies, sensors, the Internet of Things, big data, and artificial intelligence to transform traditional farming into precision-oriented and data-driven systems. These systems aim to improve productivity while making better use of resources. At the beginning of each growing season, farmers must make decisions that guide the success of the entire production cycle. The most important of these choices is deciding which crops to plant and how to divide land among them. This choice influences all later activities, such as planning the planting schedule, preparing the soil, and organizing the use of inputs. Because of its importance, crop selection is often described as the first step in farm planning. The first contribution of this thesis responds to this problem by introducing an interpretable crop selection system. The system integrates SHAP-based explanations to show how soil properties and climate conditions affect each recommendation. It combines strong predictive ability with clear explanations, offering a practical tool that farmers and advisors can use with greater trust.

After the crop has been chosen, the next important question is *"how much to expect."* Accurate yield forecasting allows farmers to organize inputs, schedule labor, manage uncertainty, and prepare for market activities. The second contribution of this thesis addresses this by designing a stacked ensemble learning framework, developed with greenhouse tomato production as a case study. It delivers accurate daily yield forecasts and achieves better results than standard regression methods, providing a reliable decision-support tool for greenhouse management.

Since both crop selection and yield forecasting depend on the quality of agricultural data, the third contribution focuses on how this data can be kept secure, reliable, and trustworthy. To achieve this, a blockchain-based approach is proposed that integrates encryption, distributed file storage, and smart contracts. The approach ensures data traceability, confidentiality, and tamper-resistance.

Keywords: Smart Predictive Agriculture; Crop Selection; Interpretable Machine Learning; SHAP; Tomato Yield Prediction; Ensemble learning, Blockchain; Data Integrity; Decision Support Systems.

Résumé

L'agriculture intelligente intègre les technologies numériques, les capteurs, l'Internet des objets, le big data et l'intelligence artificielle afin de transformer l'agriculture traditionnelle en systèmes de production de précision guidés par les données. Ces systèmes visent à accroître la productivité tout en optimisant l'utilisation des ressources. Au début de chaque saison culturale, les agriculteurs doivent prendre des décisions déterminantes pour la réussite de l'ensemble du cycle de production. La plus importante concerne le choix des cultures à planter, décision qui influence toutes les étapes ultérieures, telles que la planification du calendrier de semis, la préparation du sol et l'organisation des intrants.

La première contribution de cette thèse propose un système interprétable de sélection des cultures. Ce système intègre des explications basées sur SHAP pour montrer comment les propriétés du sol et les conditions climatiques influencent chaque recommandation. Il associe une forte capacité de prédiction à des explications claires, offrant ainsi un outil pratique que les agriculteurs et les conseillers peuvent utiliser en toute confiance.

Une fois la culture choisie, la question suivante est *« combien espérer »*. La deuxième contribution de cette thèse traite cette problématique en concevant un cadre d'apprentissage ensembliste empilé, appliqué à la production de tomates en serre comme étude de cas. Ce modèle fournit des prévisions quotidiennes fiables du rendement et surpasse les méthodes de régression classiques, constituant ainsi un outil efficace d'aide à la décision pour la gestion des serres.

Étant donné que la sélection des cultures et la prévision du rendement reposent toutes deux sur la qualité des données agricoles, la troisième contribution examine comment garantir la sécurité, la fiabilité et la confiance dans ces données. Pour répondre à cet enjeu, une approche basée sur la blockchain est proposée, intégrant le chiffrement, le stockage distribué et les contrats intelligents. Cette approche assure la traçabilité, la confidentialité et la résistance à la falsification des données.

Mots-clés : Agriculture Prédictive Intelligente ; Sélection des Cultures ; Apprentissage Automatique Interprétable ; SHAP ; Prédiction du Rendement de la Tomate ; Apprentissage d'ensemble ; Blockchain ; Intégrité des Données ; Systèmes d'Aide à la Décision.

الملخص

الزراعة الذكية تدمج التقنيات الرقمية وأجهزة الاستشعار وإنترنت الأشياء والبيانات الضخمة والذكاء الاصطناعي لتحويل الزراعة التقليدية إلى أنظمة دقيقة قائمة على البيانات. تهدف هذه الأنظمة إلى تحسين الإنتاجية مع الاستخدام الأمثل للموارد. في بداية كل موسم زراعي، يتعين على المزارعين اتخاذ قرارات تحدد نجاح دورة الإنتاج بأكملها. وأهم هذه القرارات هو اختيار المحاصيل التي سيتم زراعتها وكيفية توزيع الأراضي بينها. هذا القرار يؤثر على جميع الأنشطة اللاحقة مثل تخطيط برنامج الزراعة، تحضير التربة، وتنظيم استخدام المدخلات. وبسبب أهميته، يُعتبر اختيار المحاصيل غالبًا الخطوة الأولى في تخطيط المزرعة. إلا أن العديد من الأدوات الرقمية التي تدعم هذا القرار تعتمد على نماذج الذكاء الاصطناعي و التي تعتبر كصناديق سوداء, تفتقر إلى التفسير وقد تقلل من ثقة المزارعين. المساهمة الأولى في هذه الأطروحة تعالج هذه المشكلة من خلال تقديم نظام قابل للتفسير لاختيار المحاصيل. حيث يدمج هذا النظام شروحات مبنية على تقنية SHAP لبيان كيفية تأثير خصائص التربة والظروف المناخية على كل توصية. كما يجمع بين قدرة تنبؤية قوية وتفسيرات واضحة، مما يوفر أداة عملية يمكن للمزارعين والمستشارين استخدامها بثقة أكبر.

بعد اختيار المحصول، يبرز السؤال المهم التالي وهو: *"كم سيكون العائد المتوقع؟"* إذ تسمح التوقعات الدقيقة للإنتاج للمزارعين بتنظيم المدخلات والاستعداد للأنشطة التسويقية. المساهمة الثانية في هذه الأطروحة تعالج هذا الجانب من خلال تصميم إطار تعلم آلي تجميعي مكدس، طُوّر باستخدام إنتاج الطماطم في البيوت المحمية كدراسة حالة. وقد وفر هذا الإطار تنبؤات يومية دقيقة بالمردود، وتفوق على طرق الانحدار التقليدية، مما يجعله أداة موثوقة لدعم القرار في إدارة البيوت المحمية.

ونظرًا لأن اختيار المحاصيل والتنبؤ بالمردود يعتمدان معًا على جودة البيانات الزراعية، فإن المساهمة الثالثة تركز على كيفية الحفاظ على هذه البيانات آمنة وموثوقة. ولتحقيق ذلك، تم اقتراح نهج قائم على تقنية البلوكشين يدمج التشفير، والتخزين الموزع للملفات، والعقود الذكية. هذا النهج يضمن تتبع البيانات، وسريتها، ومقاومتها للتلاعب. كما يتيح تبادلًا شفافًا وقابلًا للتدقيق للمعلومات بين المزارع والمؤسسات، مما يعزز الثقة في الأنظمة الزراعية المعتمدة على البيانات.

الكلمات المفتاحية: الزراعة التنبؤية الذكية؛ اختيار المحاصيل؛ التعلّم الآلي القابل للتفسير؛ التنبؤ بغلة الطماطم؛ التعلم الجماعي؛ البلوك تشين؛ سلامة البيانات؛ نظم دعم القرار.

Acknowledgements

I would like to express my sincere thanks to my supervisor, **Pr. Labib Sadek Terrissa**, for his guidance, support, and encouragement throughout this research.

My deep gratitude also goes to my co-supervisor, **Dr. Soheyb Ayad**, for his help, advice, and continuous assistance.

I would like to thank the members of the jury, **Pr. Khaled Rezeg**, **Dr. Abdelhak Merizig**, and **Dr. Abdelouahab Belazoui**, for taking the time to evaluate my work and for their valuable comments.

Finally, I thank everyone who supported me in any way during the completion of this thesis.

Contents

List of Tables

List of Figures

List of Abbreviations

AGC Autonomous Greenhouse Challenge

AES Advanced Encryption Standard

AI Artificial Intelligence

ANOVA Analysis of Variance

AUC-ROC Area Under the Receiver Operating Characteristic Curve

AEs Autoencoders

ARIMA AutoRegressive Integrated Moving Average

AS Agricultural Site

CID Content Identifier

CN Consensus Node

CO$_2$ Carbon Dioxide

CNN Convolutional Neural Network

Cum_irr Cumulative Irrigation

CV Cross-Validation

DT Decision Tree

EDA Exploratory Data Analysis

EC Electrical Conductivity

ETH Ethereum (Blockchain Platform)

FAO Food and Agriculture Organization

F1 F1 Score (harmonic mean of precision and recall)

FN False Negative

FP False Positive

GAN Generative Adversarial Network

GDP Gross Domestic Product

GHI Global Hunger Index

GI Government Institution

HPS High-Pressure Sodium (lamp)

ICE Individual Conditional Expectation

IoT Internet of Things

IQR Interquartile Range

IPFS InterPlanetary File System

KNN K-Nearest Neighbor

LGB LightGBM

LIME Local Interpretable Model-agnostic Explanations

LSTM Long Short-Term Memory

LED Light-Emitting Diode

MAE Mean Absolute Error

MCC Matthews Correlation Coefficient

MI Mutual Information

Min–Max Minimum–Maximum Normalization

ML Machine Learning

MLP Multi-Layer Perceptron

MSE Mean Squared Error

MTL Multitask Learning

NB Naïve Bayes

NSE Nash–Sutcliffe Efficiency

nRMSE Normalized Root Mean Squared Error

PCA Principal Component Analysis

PDP Partial Dependence Plot

PAR Photosynthetically Active Radiation

Prod Production

RF Random Forest

RL Reinforcement Learning

RNN Recurrent Neural Network

RMSE Root Mean Squared Error

ROI Return on Investment

ROC Receiver Operating Characteristic

SHAP SHapley Additive exPlanations

SD Standard Deviation

SMOTE Synthetic Minority Over-sampling Technique

SVR Support Vector Regression

SVM Support Vector Machine

SHA256 Secure Hash Algorithm 256-bit

Std Standard Deviation

TFP Total Factor Productivity

TN True Negative

TP True Positive

Tair Greenhouse Air Temperature

Rhair Relative Humidity in Greenhouse Air

XAI Explainable Artificial Intelligence

XGB XGBoost

WLSTM Weight-based Long Short-Term Memory

Z-score Standard Score

R^2 Coefficient of Determination

Chapter 1

General introduction

1.1 Context

Agriculture plays a central role in ensuring global food security, supporting economic development, and maintaining environmental balance. It provides livelihoods for billions of people and remains a key source of employment and income, especially in developing regions. However, the sector faces a range of challenges that threaten its ability to remain productive and resilient. The world population is projected to approach 10 billion by 2050, driving a sharp increase in food demand and creating pressure for higher productivity and more efficient resource use [6] . At the same time, climate variability, resource constraints, and environmental degradation place heavy demands on farming systems, limiting the capacity of traditional practices to adapt to changing environmental and economic conditions [7].

This situation has encouraged the development of smart agriculture, supported by advanced technologies such as data science, artificial intelligence (AI), machine learning (ML), and blockchain. Smart agriculture relies on combining information from many sources, including soil sensors, satellite images, climate forecasts, and market data, to guide accurate and data-driven decisions. This approach, often described as smart predictive agriculture, helps farmers and decision-makers improve management practices, use resources more efficiently, and build farming systems that can adapt to changing environmental and economic conditions [8] [9].

Data science supports this transformation by offering analytical tools and methods that extract useful knowledge from complex datasets. Among these tools, machine learning enables predictive modeling for tasks such as crop selection and yield forecasting, providing guidance for agricultural planning and resource management [10]. Despite significant progress, important gaps remain. One key limitation is the limited interpretability of many predictive agricultural systems, which often function as “black boxes.” This absence of explainability reduces trust and slows adoption among farmers and practitioners, who require clear and understandable recommendations to make confident decisions in real-world conditions.

Furthermore, As data-driven approaches gain importance for improving agricultural productivity, accurate yield forecasting in controlled environments such as greenhouses remains a complex challenge. The interactions among environmental variables are often nonlinear and interdependent, making precise prediction difficult and increasing economic uncertainty. At the same time, the rapid expansion of digital farming technologies requires strong safeguards to protect agricultural data, ensuring its integrity and maintaining reliable and traceable data flows that are essential for collaborative farming practices.

1.2 Problem Statement

In the rapidly changing landscape of agriculture, farmers and stakeholders are increasingly faced with the challenge of making decisions that balance productivity, sustainability, and resilience. The transition from traditional, experience-based practices to data-driven agriculture is reshaping how these decisions are made, but also introducing new complexities and demands [11].

At the beginning of each growing season, farmers face a series of planning tasks that determine the success of the entire production cycle. The first and most influential of these tasks is deciding which crops to grow and how to allocate land among them. This early decision serves as the foundation for all subsequent actions, including planting schedules, soil preparation, and resource management. Agricultural research and extension services consistently identify crop and land-use planning as the starting point of seasonal planning

[12, 13]. Whish et al.[14] describe the question of "what to plant, when, and where" as a complex challenge encountered by every farmer. Similarly, the University of Minnesota Extension guide explicitly lists "Step 1: Decide what to grow" as the opening action in farm planning[15]. Practical guides to crop rotations also begin with crop selection, recognizing that decisions on crop type and rotation shape every later stage of management. A Kentucky Extension note likewise emphasizes land-use planning as the first decision, asking, "Should this land be cropped? If so, with what crop or crop rotation?"[16].

In practice, farmers consider a range of factors before finalizing this critical choice, including field conditions, crop rotations, soil health, and local climate. Once the crop is determined, subsequent tasks such as selecting planting dates and preparing fertilizers or soil amendments follow in a logical sequence. The decision of what to plant initiates the entire seasonal workflow, making it the key point where accurate, data-supported recommendations can have the greatest impact on farm productivity and resource use [17].

After deciding what to plant, the next important question is "how much to expect?" Reliable yield forecasts are essential for organizing labor, planning storage, arranging marketing activities, and managing farm finances. Recent progress in artificial intelligence has advanced yield prediction by combining historical production data, satellite observations, and real-time measurements collected from the field. Machine learning methods are able to detect complex patterns among weather conditions, soil characteristics, and crop growth, allowing farmers to refine management practices and marketing plans as new information becomes available during the season. These developments point to the need for practical, context-aware forecasting tools that can deliver accurate and timely predictions for decision-making.

The reliability of agricultural decision-making depends on the security and integrity of the foundational data. The rapid growth of digital technologies in farming, including sensors, automated equipment, and shared data platforms, offers new opportunities for data-driven management but also introduces significant risks. This raises an important question: how can the data that supports these decisions be kept secure, accurate, and trustworthy?

Conventional data management systems often fall short in preventing tampering, unauthorized access, or the loss of data origin, which can weaken confidence among farmers and other stakeholders and slow the adoption of smart agricultural practices. Protecting data integrity and ensuring transparent data flows have therefore become essential for developing collaborative farming systems where choices about “what to plant” and “how much to expect” can be made with greater confidence.

1.3 Contributions

This thesis addresses the gaps and limitations identified in contemporary agricultural practices by making three key contributions, each aimed at enhancing the effectiveness and reliability of smart predictive agriculture:

Contribution 1: Interpretable Crop Selection System for Optimized Farming Decisions

The first contribution of this thesis is the design and implementation of an interpretable and high-accuracy crop selection system, addressing the need for both predictive reliability and model transparency in smart agriculture. The main stages of this contribution are outlined below:

- **Dataset Construction and Characterization:** A balanced dataset of 2,200 records covering 22 crop types was used. Each crop is described through key agronomic features, including soil nitrogen (N), phosphorus (P), potassium (K), pH, temperature, humidity, and rainfall.

- **Exploratory Data Analysis (EDA):** A detailed statistical and visual analysis was conducted to assess feature distributions, identify influential predictors, and examine relationships between input variables and crop classes. This stage provided essential insights that guided the selection of suitable preprocessing techniques.

- **Data Preprocessing:** The preprocessing pipeline included outlier detection and imputation, feature scaling, categorical label encoding, and data augmentation to expand each crop class. These steps ensured data quality and improved compatibility with machine learning models.

- **Proposed CS-AdaRF-SHAP System:** The system leverages an adaptive boosting strategy that trains a sequence of Random Forest classifiers while iteratively reweighting misclassified instances. This enhances the model's ability to distinguish between crops with similar feature patterns and increases robustness across diverse agricultural conditions. Hyperparameters were carefully tuned to achieve optimal predictive performance. The model integrates SHapley Additive exPlanations (SHAP) to provide both global and local interpretability. SHAP values quantify the contribution of each feature to the final recommendation, offering clear and agronomically meaningful explanations suitable for stakeholders.

  The proposed CS-AdaRF-SHAP system achieved high test accuracy along with strong precision, recall, and F1-score values. Most errors occurred between agronomically similar crop classes. The system consistently outperformed baseline and ensemble models, demonstrating its suitability for real-world deployment in smart agriculture.

Contribution 2: Data-Driven Crop Yield Prediction

The second contribution of this thesis is the development of a data-driven system for predicting crop yield, using tomato production in greenhouse conditions as a case study. This work addresses the practical question of "how much to expect," which is essential for planning inputs, scheduling labor, and organizing marketing activities. The proposed approach employs a stacked ensemble learning framework that combines the predictive outputs of several models to improve the accuracy of daily yield estimation.

A careful preprocessing procedure was designed to secure data quality and reliability. The workflow included systematic cleaning, temporal alignment, normalization, data augmentation, and the selection of key features that capture relevant environmental and crop growth dynamics. The model was trained and evaluated on real multivariate greenhouse data and achieved higher predictive accuracy than standard regression techniques.

Contribution 3: Blockchain-Based Approach to Securing Data in Smart Agriculture

The third contribution of this thesis is the implementation of a blockchain-based approach to ensure the security, integrity, and reliable sharing of agricultural data in IoT-enabled greenhouse environments. Addressing the essential challenge of "How can the data that supports these decisions remain secure, reliable, and trustworthy?" the proposed approach integrates blockchain technology, smart contracts, edge computing, and distributed file storage (IPFS) into a unified framework.

The system enables all registered agricultural sites to collect, encrypt, and transmit data to a distributed platform under the supervision of a central government institution. The workflow incorporates cryptographic hashing (SHA256) for integrity verification, AES encryption for data confidentiality, and IPFS for tamper-evident, decentralized storage. Transactional metadata, including data ownership, access rights, and file hashes, is securely recorded on the blockchain via custom smart contracts, ensuring immutability, transparency, and auditable access.

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2: Preliminaries and Basic Concepts

This chapter introduces the fundamental theories and background necessary for understanding the remainder of the thesis.It begins with an overview of data science, its lifecycle, and the role of feature engineering in building predictive models. Core principles of machine learning and deep learning are then introduced. The chapter also discusses interpretable and explainable AI. Finally, it presents the fundamentals of blockchain technology, outlining its potential for ensuring data security and trust in smart agriculture.

Chapter 3: Smart Agriculture: State of the Art

This chapter reviews the evolution of agriculture from traditional practices to modern, AI-driven systems. It discusses the main challenges and limitations of conventional approaches, examines recent advances in smart agriculture, including AI-based crop selection, yield prediction, and blockchain-enabled data security, and identifies the key research gaps that motivate and shape the contributions of this thesis.

Chapter 4: Contribution 1: Interpretable Crop Selection for Optimized Farming Decisions

This chapter presents the first contribution of the thesis, which focuses on the design of an interpretable crop selection system. It describes the architecture and methodology of the proposed CS-AdaRF-SHAP system, reports the experimental results demonstrating both its performance and interpretability, and concludes with a discussion of its practical implications for real-world agricultural decision-making.

Chapter 5: Contribution 2: Data-Driven Crop Yield Prediction

This chapter presents the second contribution of the thesis, which addresses the challenge of predicting crop yield (with a focus on tomato) in greenhouse environments. It introduces the proposed stacked ensemble learning framework, details the dataset and data preprocessing methods, and provides a thorough evaluation of predictive performance in comparison with baseline models.

Chapter 6: Contribution 3: Blockchain-Based Approach to Securing Data in Smart Agriculture

This chapter presents the third contribution of the thesis, a secure approach for managing agricultural data using blockchain and IPFS. It describes the system architecture and key implementation steps, including smart contract deployment and data encryption. The chapter also demonstrates the advantages of the approach in ensuring data integrity, privacy, and reliable data sharing among agricultural stakeholders.

Chapter 7: General Conclusion and Perspectives

The final chapter integrates the main outcomes of the thesis, reviewing the challenges addressed and the proposed solutions. It discusses the scientific and practical contributions to the field of smart agriculture, evaluates the results achieved, and concludes with perspectives for future research and development.

Chapter 2

Preliminaries and Basic Concepts

2.1 Introduction

This chapter presents the foundational concepts central to our thesis and provides a comprehensive overview of the main domains that will be explored in the subsequent chapters. Section 1.2 presents the fundamentals of Data Science, covering its lifecycle, preprocessing techniques, feature engineering, and exploratory data analysis methods. Section 1.3 introduces Machine Learning (ML) and Deep Learning (DL), outlining their core concepts, methodologies, and evaluation metrics. Section 1.4 focuses on Interpretable and Explainable AI (XAI). Finally, Section 1.5 reviews Blockchain technology with attention to the mechanisms that ensure data security, integrity, and traceability.

2.2 Data Science Fundamentals

2.2.1 Definition and Scope

Data Science is an interdisciplinary field that applies scientific methods, algorithms, and computational systems to extract knowledge from both structured and unstructured data. As illustrated in Figure 2.1, it integrates principles from statistics, computer science, mathematics, and domain-specific expertise to analyze complex datasets and support informed decision-making [18].

Each of these components plays an important role:

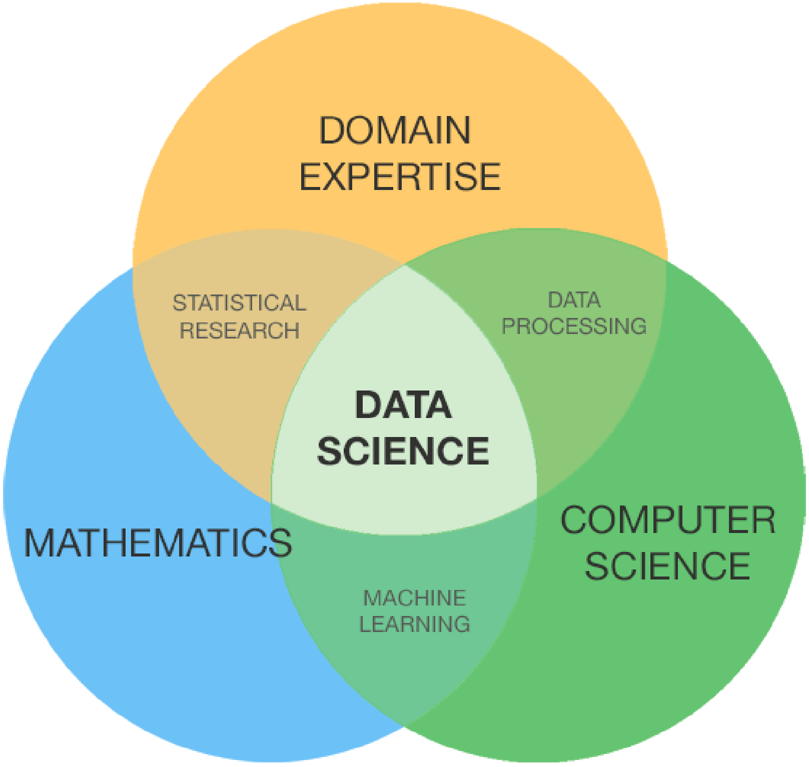


Figure 2.1: Core components of Data Science.

- **Statistics and Mathematics** provide the theoretical foundations for data modeling, hypothesis testing, and quantitative analysis.
- **Computer Science** supports scalable data processing, algorithm design, and the implementation of machine learning methods.
- **Domain Expertise** ensures that analytical approaches and data-driven solutions remain relevant, interpretable, and actionable within a specific context.

The integration of these components enables Data Science to:

- Detect patterns, trends, and anomalies within complex datasets.
- Develop predictive and prescriptive models that inform and optimize decision-making.
- Support automation, real-time analytics, and adaptive systems across diverse domains.

Data Science has become essential across a wide range of disciplines, including:

- **Finance**: It is applied in credit risk modeling, fraud detection, and algorithmic trading to enhance decision-making and risk management.
- **Healthcare**: It supports predictive diagnostics, genomics, and personalized treatment strategies, enabling more accurate and patient-centered care.
- **Marketing**: It facilitates customer segmentation, recommendation systems, and campaign optimization, which improves customer engagement and business outcomes.
- **Agriculture**: It contributes to crop yield prediction, soil and climate analytics, precision irrigation, and sustainable resource management, promoting efficiency and resilience in food production [19].

Relevance to Agriculture

In agriculture, Data Science enables stakeholders, including farmers, agronomists, and policymakers, to make informed and evidence-based decisions. By integrating historical records, sensor measurements, weather forecasts, and remote sensing imagery, its applications include:

- **Crop recommendation systems**: Identifying suitable crops for site-specific soil and climate conditions.
- **Predictive modeling**: Developing early warning systems for disease outbreaks, pest invasions, and yield variability.
- **Resource optimization**: Improving the efficiency of water and fertilizer use through data-driven strategies [19].

2.2.2 Data Science Lifecycle

The Data Science lifecycle consists of a structured sequence of phases that transform raw data into actionable information and predictive models. This iterative process ensures methodological robustness and adaptability across diverse domains, including smart agriculture. Although several models have been proposed, a comprehensive review by [20] identifies six core phases that are common to most Data Science process frameworks:

1. **Problem Definition:** Defining the research question or business objective, which establishes the foundation for the entire Data Science project.

2. **Data Acquisition:** Collecting relevant data from diverse sources such as sensors, databases, and external repositories, while ensuring data quality and relevance.

3. **Data Preparation:** Cleaning and transforming the data to address issues such as missing values, outliers, and inconsistencies, which enables effective analysis.

4. **Modeling:** Applying statistical and machine learning algorithms to discover patterns, generate predictions, or perform classification, depending on the specific problem.

5. **Evaluation:** Measuring model performance with appropriate metrics (e.g., accuracy, precision, recall) to validate reliability and robustness.

6. **Deployment:** Implementing the model in operational environments to support real-time decision-making and facilitate continuous monitoring for performance improvement.

2.2.3 Data Preprocessing & Feature Engineering

Data preprocessing comprises a range of systematic operations designed to enhance data quality and consistency [22, 23]:

- **Handling Missing Values:** Incomplete data may arise from sensor malfunctions, recording errors, or limitations in data collection protocols. Several strategies are commonly employed:

  - *Deletion:* Removing records or attributes with a small proportion of missing entries when the loss of information is minimal.
  - *Simple Imputation:* Substituting missing values with statistical measures such as the mean, median, or mode of the corresponding attribute.

  - *Advanced Imputation:* Applying more sophisticated techniques, including K-nearest neighbors or regression-based approaches, to estimate missing values based on observed patterns in the data.

- **Outlier Detection and Treatment:** The presence of Outliers can strongly affect statistical results and reduce the accuracy of models. To identify such anomalies, a variety of techniques are employed:

  - *Univariate Methods:* Approaches such as Z-scores, interquartile range (IQR) analysis, and visual inspection through boxplots.
  - *Multivariate Methods:* Techniques including Mahalanobis distance or isolation forests, which account for relationships across multiple variables.

  Once detected, outliers may be addressed through removal, capping extreme values, or applying suitable transformations to reduce their impact on downstream analyses.

- **Feature Scaling and Normalization:** Because many machine learning algorithms are sensitive to differences in feature magnitudes, scaling is often an essential step to ensure balanced contributions of all variables. Common approaches include:

  - *Standardization:* Transforming features so that they have a mean of zero and a standard deviation of one.
  - *Min–Max Normalization:* Rescaling features to fall within a fixed interval, typically [0,1], which preserves relative relationships while constraining absolute ranges.
  - *Robust Scaling:* Applying transformations based on the median and interquartile range (IQR), thereby reducing sensitivity to extreme values or outliers.

- **Encoding Categorical Variables:** Since many machine learning algorithms require numerical input, categorical attributes must be transformed into suitable numerical representations. Common strategies include:

  - *One-Hot Encoding:* Generating a set of binary indicator columns, each corresponding to a distinct category, thereby avoiding any assumption of order.
  - *Ordinal Encoding:* Assigning integers to categories that possess a meaningful order or ranking, preserving their relative structure.
  - *Target Encoding:* Substituting categorical levels with the mean value of the target variable, a method that can be effective but requires careful application to reduce the risk of data leakage.

- **Class Balancing:** Imbalanced datasets can lead to biased models that favor majority classes, reducing overall predictive performance. To address this issue, several techniques are commonly applied:

  - *Random Oversampling/Undersampling:* Modifying class distributions by either duplicating minority class samples or removing instances from the majority class.
  - *SMOTE (Synthetic Minority Over-sampling Technique):* Creating synthetic examples for underrepresented classes by interpolating between existing minority samples, thereby improving class representation without simple duplication.

- **Data Partitioning:** To evaluate model performance reliably and prevent overfitting, datasets are typically divided into distinct subsets for training, validation, and testing. Common approaches include:

  - *Hold-out Validation:* Splitting the dataset into independent subsets, where one portion is used for training and another for testing model performance.
  - *Stratified Sampling:* Creating partitions that preserve the original distribution of classes, which is particularly important in imbalanced datasets.
  - *Cross-Validation:* Repeatedly partitioning the data into multiple folds to assess model stability and robustness across different training–testing splits.

Feature Engineering

Feature engineering refers to the process of constructing new input variables from existing data with the goal of enhancing model accuracy and interpretability [24]. Within agricultural applications, this process often includes:

- **Temporal Features:** Deriving time-related variables, such as growing degree days or the number of days since planting, to capture seasonal and developmental patterns in crops.
- **Spectral Indices:** Computing vegetation metrics, for example the Normalized Difference Vegetation Index (NDVI), from multispectral or hyperspectral imagery to quantify plant health and vigor.
- **Soil–Weather Interactions:** Integrating soil moisture measurements with temperature records to generate indicators of drought stress or other environmental constraints.
- **Dimensionality Reduction:** Employing statistical methods such as Principal Component Analysis (PCA) to condense high-dimensional datasets into a smaller set of informative features while minimizing redundancy.

Successful feature engineering typically requires a combination of domain knowledge and iterative experimentation, as the most informative features are often context-specific and depend on both the crop system and the modeling objective.

2.2.4 Exploratory Data Analysis & Visualization

Purpose and Significance

Exploratory Data Analysis (EDA) represents a critical stage in the data science workflow, particularly in the context of smart agriculture. It involves the systematic examination and summarization of key dataset characteristics, frequently supported by visual techniques. Through EDA, researchers can reveal underlying patterns, identify anomalies, evaluate assumptions, and conduct preliminary hypothesis testing using a combination of statistical measures and graphical representations [25].

In agricultural applications, EDA serves several important functions:

- Revealing patterns and relationships among key variables, such as soil characteristics, weather conditions, and crop yields, which provide understanding of fundamental agronomic processes.
- Detecting outliers or unusual observations that may reflect measurement errors, sensor malfunctions, or exceptional environmental events.
- Evaluating overall data quality and completeness to ensure that subsequent analyses are based on reliable and representative information.
- Guiding the choice of suitable modeling approaches and informing feature engineering strategies by prioritizing the most relevant attributes within the dataset.

Statistical Techniques

A range of statistical methods are commonly applied during EDA to describe and interpret the characteristics of agricultural datasets:

- **Descriptive Statistics:** Computing summary measures such as the mean, median, standard deviation, skewness, and kurtosis to characterize central tendency, variability, and distributional shape.
- **Correlation Analysis:** Assessing the strength and direction of relationships between variables, often through Pearson correlation for linear associations or Spearman rank correlation for non-linear monotonic patterns.
- **Hypothesis Testing:** Employing inferential procedures such as analysis of variance (ANOVA) or chi-square tests to examine group differences or evaluate associations among categorical variables [26].

Visualization Techniques

Visualization is an essential component of EDA, offering an accessible means of interpreting complex datasets and uncovering patterns that may not be apparent through numerical analyses alone. Commonly employed visualization methods include [27]:

- **Histograms and Density Plots:** Depict the distribution of individual variables, providing understanding into central tendency, spread, and overall shape.
- **Box Plots:** Display distributions by focusing on medians, quartiles, and variability, while also allowing the detection of potential outliers.
- **Scatter Plots:** Display the relationship between two continuous variables, making it possible to observe correlations, clusters, or emerging trends.
- **Heatmaps:** Represent correlation matrices or spatially referenced data in a compact visual form, facilitating the recognition of systematic patterns and clusters.
- **Time Series Plots:** Track the evolution of variables across time, which is particularly valuable for monitoring crop growth dynamics, weather conditions, or seasonal effects.
- **Geospatial Maps:** Illustrate the spatial distribution of agricultural variables, supporting site-specific management practices and precision farming decisions.

Integration with Data Pipeline

The findings derived from EDA play a crucial role in shaping and refining the broader data preprocessing pipeline [26]. By systematically examining the data, EDA provides evidence-based guidance for several subsequent steps, such as:

- **Data Cleaning:** Detecting missing values, inconsistencies, or anomalies that require imputation, correction, or removal to ensure data reliability.
- **Feature Selection:** Identifying variables that hold the greatest relevance for predictive modeling, while discarding redundant or uninformative attributes.
- **Model Selection:** Informing the choice of algorithms by revealing structural characteristics of the data, such as linearity, dimensionality, or class imbalance.

2.3 Machine Learning and Deep Learning

2.3.1 Machine Learning Basics

Machine Learning (ML) is an interdisciplinary domain concerned with the design of algorithms that can learn patterns from data and generate predictions or decisions without relying on explicit rule-based programming. Drawing upon concepts from computer science, statistics, and applied mathematics, ML constitutes a foundational element of artificial intelligence (AI) and has become a key driver of data-driven decision-making across diverse fields, including agriculture [28].

In agricultural applications, ML supports the analysis of complex datasets originating from diverse sources, including in-field sensors, satellite imagery, and historical farm records. By leveraging these data streams, ML techniques can be used to predict crop yields, detect the onset of diseases, optimize the allocation of resources, and improve the efficiency and sustainability of farm management practices [29].

Prominent perspectives on ML can be framed as follows:

- **Algorithmic Optimization Perspective:** ML is viewed as the process of designing computer programs that improve their performance on specific tasks by optimizing objective functions through experience with data [30].
- **Predictive Pattern Recognition:** From this perspective, ML emphasizes the development of methods that autonomously detect patterns within datasets and use them to forecast future outcomes or events [31].
- **Actionable Regularity Discovery:** Here, ML is understood as the automated identification of regularities or structures in data through computational algorithms, with the goal of transforming these findings into practical, decision-oriented outputs [32].

2.3.2 Methodologies

Machine learning (ML) methodologies are commonly classified according to the type of input data and the corresponding learning objectives. The principal paradigms include supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, multitask learning, and transfer learning [33]. Each of these approaches provides distinct advantages and is chosen with respect to the problem setting, data availability, and desired outcomes.

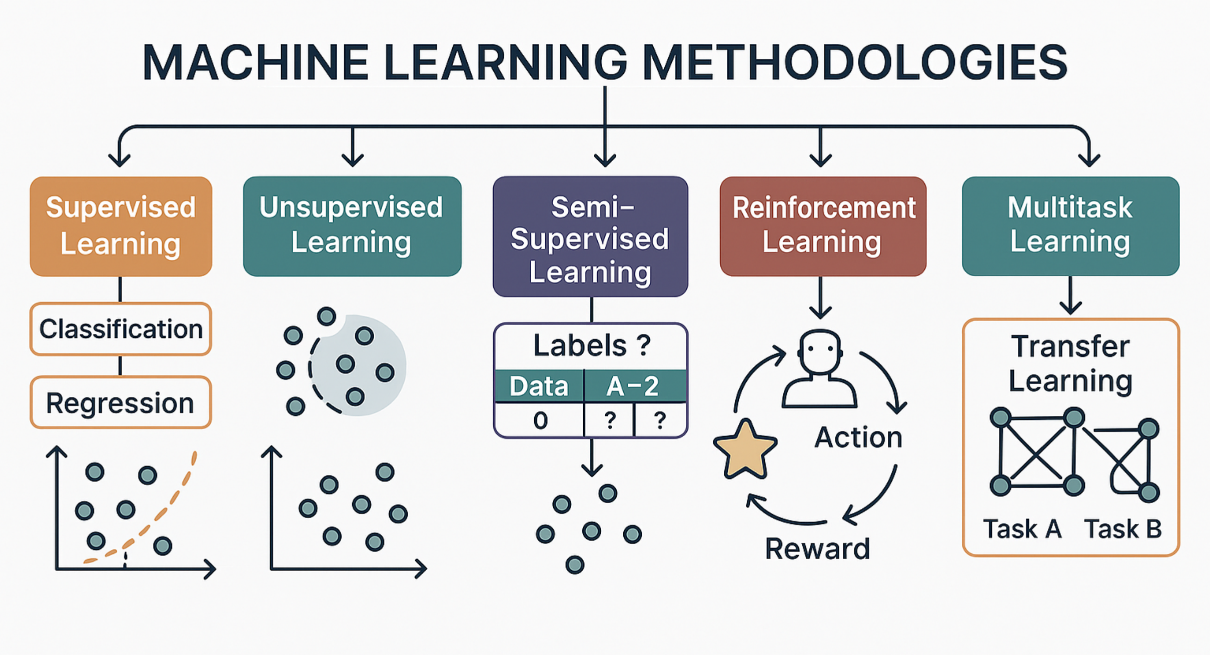


Figure 2.2: Machine Learning Methodologies.

Supervised Learning

Supervised learning refers to the process of training models on labeled datasets, in which each input vector $\mathbf{x}_i$ is associated with a corresponding output label y_i. The primary objective is to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}$ that predicts outputs from inputs with high accuracy by minimizing a predefined loss function over the training set:

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i), y_i), \tag{2.1}$$

where $\mathcal{L}$ denotes the chosen loss function and $\mathcal{F}$ represents the hypothesis space of candidate functions. In practice, supervised learning underpins a wide range of tasks, most notably classification and regression, making it one of the most widely applied paradigms in machine learning [28].

Unsupervised Learning

Unsupervised learning addresses the analysis of datasets that lack predefined labels, with the primary goal of discovering hidden patterns, groupings, or inherent structures within the data. Techniques commonly employed in this paradigm include clustering methods, dimensionality reduction approaches, and anomaly detection algorithms. [29].

Semi-Supervised Learning

Semi-supervised learning integrates a limited set of labeled examples with a substantially larger pool of unlabeled data during model training. This paradigm is especially advantageous in situations where the process of generating high-quality labels is costly, labor-intensive, or otherwise impractical. To make effective use of the available unlabeled data, a range of strategies can be applied, including self-training, co-training, and graph-based approaches, each of which seeks to enhance predictive performance by exploiting the underlying structure of the data [34].

Reinforcement Learning

Reinforcement learning (RL) is a paradigm in which an agent interacts dynamically with an environment, gradually learning to select actions that maximize cumulative rewards while minimizing penalties. The learning process is inherently iterative, relying on trial-and-error exploration combined with feedback signals that shape the agent's decision-making policy over time. Within agricultural systems, RL shows considerable promise for applications such as the coordination of autonomous farming machinery, optimization of irrigation schedules, and the development of adaptive pest management strategies [28].

Multitask Learning

Multitask learning (MTL) is an approach designed to enhance generalization by training models on several related tasks at the same time, thereby enabling the sharing of underlying representations across them. This strategy is particularly effective when the tasks are interdependent or draw upon overlapping sources of information, as the joint learning process allows the model to exploit shared structure and reduce overfitting to any single task. Within agricultural applications, MTL can be employed to predict multiple crop traits or environmental variables simultaneously, offering a more comprehensive understanding of complex agroecosystems [35].

Transfer Learning

Transfer learning focuses on transferring knowledge from a source task to improve learning in a target task, especially when the target task has limited data. Pretrained models on large datasets can be fine-tuned for specific agricultural tasks, such as disease detection or yield estimation, enhancing performance with minimal labeled data [36].

2.3.3 Model Evaluation Metrics

Evaluating the performance of machine learning (ML) models is a critical step in the development and deployment of data-driven solutions. The choice of appropriate evaluation metrics determines how effectively a model's predictions can be assessed and whether it is suitable for practical applications. Well-defined metrics provide an objective basis for comparing different models, guiding model selection, and ensuring robustness across diverse problem settings. This section outlines key evaluation metrics commonly used in classification and regression tasks, presenting their mathematical definitions and discussing their comparative advantages and limitations.

Classification Metrics

In classification problems, the goal of a model is to assign inputs to discrete categories. The quality of these predictions is commonly evaluated through a confusion matrix, which provides a structured summary of the model's performance. The matrix is composed of the following elements:

- **True Positives (TP)**: Instances correctly identified as belonging to the positive class.
- **True Negatives (TN)**: Instances correctly identified as belonging to the negative class.
- **False Positives (FP)**: Negative instances that are incorrectly classified as positive.

- **False Negatives (FN)**: Positive instances that are incorrectly classified as negative.

From the confusion matrix, a number of widely used performance metrics can be derived, each illustrating different aspects of model behavior:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Represents the proportion of correctly classified instances relative to the total number of cases. While useful as a general indicator, accuracy can give a distorted picture when datasets are highly imbalanced, as it may overlook minority classes.

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.3}$$

Quantifies the reliability of positive predictions by indicating the fraction of predicted positives that are truly positive. High precision reflects a model that makes few false positive errors.

Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.4}$$

Assesses the model's ability to identify all relevant positive instances. A high recall value means that most of the actual positives are successfully detected, even if this comes at the expense of more false positives.

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.5}$$

Provides a single measure that balances precision and recall by calculating their harmonic mean. It is especially useful when one seeks to account for both types of classification error simultaneously.

Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.6}$$

Offers a more comprehensive evaluation by incorporating all four elements of the confusion matrix. Unlike accuracy, MCC remains informative even in the presence of strong class imbalance, making it a robust alternative [37].

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) Reflects the model's capacity to discriminate between classes over a range of decision thresholds. A higher AUC value indicates stronger overall separability between positive and negative classes.

Regression Metrics

Regression problems concern the prediction of continuous variables, and their evaluation relies on metrics that quantify the accuracy and reliability of model outputs. Commonly employed measures include:

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.7}$$

Reflects the average absolute deviation between predictions and observed values, providing an intuitive measure of overall error magnitude without accounting for direction.

Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.8}$$

Gives greater weight to larger errors by squaring the residuals, making it particularly sensitive to outliers.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.9}$$

Expresses the average prediction error in the same units as the target variable, thereby facilitating direct interpretability.

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.10}$$

Represents the proportion of variance in the dependent variable that is explained by the model, with values closer to 1 indicating stronger explanatory power.

Nash–Sutcliffe Efficiency (NSE)

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.11}$$

Frequently applied in hydrological and environmental modeling, where it serves as a measure of predictive skill relative to the mean of observed data [38].

2.3.4 Deep Learning

Deep Learning (DL), a specialized branch within the broader field of machine learning, is distinguished by its reliance on artificial neural networks composed of multiple interconnected layers. This layered architecture allows models to capture and represent highly complex, non-linear relationships in data with remarkable effectiveness. In recent years, DL has gained increasing prominence as a transformative tool, particularly in domains where large and heterogeneous datasets are prevalent. [39].

Fundamental Architectures

Several deep learning architectures have become foundational across a wide range of domains, each designed to address different data types and problem settings:

- **Convolutional Neural Networks (CNNs)**: CNNs are particularly well suited for image-related tasks, as they can effectively capture spatial hierarchies and local patterns. They have been extensively used in image classification, object detection, and computer vision more broadly, achieving state-of-the-art performance in many benchmarks [39].

- **Recurrent Neural Networks (RNNs)**: RNNs, including advanced variants such as Long Short-Term Memory (LSTM) networks, are designed for sequential and temporal data. They are widely applied in natural language processing, speech recognition, and time-series modeling, where the ability to capture dependencies across time is essential [40].

- **Autoencoders (AEs)**: Autoencoders are used primarily for unsupervised feature learning and dimensionality reduction. They are commonly employed for tasks such as anomaly detection, data compression, and denoising, where reconstructing meaningful latent representations of input data is advantageous [41].

- **Generative Adversarial Networks (GANs)**: GANs generate synthetic data by learning to approximate complex data distributions. They have proven highly effective for data augmentation, realistic image synthesis, and style transfer, providing valuable support in scenarios where labeled data is limited [42].

- **Transformers**: Based on self-attention mechanisms, transformers have revolutionized deep learning by enabling efficient modeling of long-range dependencies. Originally developed for natural language processing, they are increasingly applied to computer vision, multimodal learning, and other domains requiring integration of diverse data types [43].

Mathematical Formulation

At the foundation of deep learning models lies the optimization of a loss function $\mathcal{L}$ defined over a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes the input features and $\mathbf{y}_i$ the corresponding target labels. The central aim is to determine a function f_θ, parameterized by θ, that minimizes the average loss across the training set:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i) \tag{2.12}$$

In practice, the choice of loss function depends on the nature of the task. For instance, cross-entropy loss is widely applied in classification problems, while mean squared error remains a standard choice for regression settings.

2.4 Interpretable and Explainable AI (XAI)

As Artificial Intelligence (AI) systems are increasingly deployed in domains where decisions carry significant consequences, the importance of transparency and interpretability has grown substantially. Explainable AI (XAI) seeks to meet this demand by developing approaches that make the functioning of complex models more understandable. By clarifying how models generate their outputs, XAI contributes to building confidence in the technology while providing users with clearer grounds for evaluation and action [44].

2.4.1 Importance Across Domains

The relevance of XAI extends across a wide range of application areas, where transparency and interpretability are not only desirable but often necessary:

- **Healthcare:** In medical diagnostics and treatment planning, understanding the basis of AI-generated predictions is essential for clinical reliability and for preserving the confidence of both practitioners and patients [45].

- **Finance:** In financial services, interpretable models are central to credit scoring and fraud detection. Clear reasoning behind model outputs is necessary for meeting regulatory requirements and for maintaining trust in decisions that can significantly affect customers [45].

- **Legal Systems:** Within judicial and legal contexts, explainable models help safeguard fairness and accountability by making automated decisions transparent and open to review when individual rights are involved [45].

- **Agriculture:** As AI systems are increasingly used for tasks such as crop monitoring and yield estimation, interpretability ensures that farmers and other stakeholders can make sense of the outputs and apply them with confidence [46].

2.4.2 Foundations of Explainable AI

Explainable AI (XAI) brings together a range of methods aimed at making the internal workings of AI models more transparent. These methods are commonly grouped into two broad categories:

- **Intrinsic Interpretability:** Models that are transparent by design, such as decision trees or linear regression, where the reasoning process can be directly followed without additional tools.

- **Post-hoc Explanations:** Approaches applied after model training to shed light on complex systems, including deep neural networks. Widely used examples are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) [47].

2.4.3 Post-hoc Explanations

Post-hoc explanation methods are introduced once a model has already been trained, with the aim of clarifying how predictions are generated without modifying the model's internal design. Such approaches are particularly useful when working with highly complex and accurate models that often function as "black boxes" to users and practitioners [48]. Broadly, post-hoc techniques can be organized into the following categories:

Model-Specific Methods

Model-specific approaches make use of the internal structure and parameters of a model to derive explanations, thus tailoring the interpretation to the architecture being analyzed. Representative techniques include:

- **Saliency Maps:** These methods identify and visualize the regions of an input, such as areas within an image, that exert the greatest influence on the model's prediction. They are particularly common in convolutional neural networks, where spatial hierarchies are central to learning [49].

- **Integrated Gradients:** This technique attributes the outcome of a deep network to its input features by computing and aggregating gradients along a path that interpolates between a baseline reference input and the actual input, thereby offering a more principled assessment of feature relevance [50].
- **Attention Mechanisms:** By assigning varying levels of weight to different parts of the input, attention mechanisms highlight which features are most influential during prediction. This enhances model performance while simultaneously offering a clearer perspective on the decision-making process [51].

Model-Agnostic Methods

Model-agnostic methods approach the learning system as a black box, examining only the relationships between inputs and outputs without reference to the internal architecture. Because of their flexibility, these techniques can be applied to a wide range of model types and are therefore widely used in practice:

- **LIME (Local Interpretable Model-Agnostic Explanations):** Provides local interpretability by fitting a simplified proxy model around a particular prediction, thus clarifying the factors that played the greatest role in shaping that outcome [52].
- **SHAP (SHapley Additive exPlanations):** Grounded in cooperative game theory, this approach assigns each feature a contribution score, quantifying its role in shaping an individual prediction [53].
- **Partial Dependence Plots (PDP):** Depict the average marginal effect of one or two selected features on the predicted response, providing a global view of feature influence [54].
- **Individual Conditional Expectation (ICE) Plots:** Complement PDPs by visualizing how predictions change at the level of individual instances, which uncovering heterogeneity in feature effects [55].
- **Counterfactual Explanations:** Explore minimal modifications to input variables that would alter the model's output, which makes them especially valuable for gen-

erating practical interpretations and clarifying how decision boundaries are formed [56].

- **Permutation Feature Importance:** Evaluates the relevance of each feature by measuring the reduction in predictive accuracy when its values are randomly permuted, offering a measure of its contribution to the overall model.[57].

2.4.4 SHAP for Model Interpretability

Within the family of model-agnostic interpretability techniques, **SHAP (SHapley Additive exPlanations)** has gained wide recognition as a rigorous and well-founded approach. Its strength lies in a solid theoretical basis drawn from cooperative game theory, combined with desirable properties such as local accuracy and consistency, which make it particularly reliable for both research and applied settings [58].

SHAP: SHapley Additive exPlanations

- Derived from the concept of *Shapley values* in game theory, SHAP attributes the contribution of each feature to a prediction by systematically considering all possible feature combinations.
- Ensures feature attributions that are both additive and consistent across predictions.
- Offers interpretability at different levels, ranging from individual predictions to overall model behavior.
- Can be applied to virtually any machine learning model, with specialized and efficient implementations available for tree-based models such as TreeSHAP.
- Produces outputs that are standardized and comparable across models as well as individual instances [58].

2.4.5 Visualization and User Interfaces

Visualization is central to making model explanations understandable, particularly for stakeholders who may not have a technical background. Well-designed visual tools can translate abstract computational processes into intuitive representations, thus narrowing the gap between complex model behavior and human interpretation [59].

- **SHAP Visualizations:**
  - *Summary Plot:* Integrates feature importance with the distribution of effects across the dataset, conveying both the magnitude and direction of influence for each feature.
  - *Force Plot:* Demonstrates how individual features push a prediction upward or downward, making it especially valuable for case-specific explanations.
  - *Dependence Plot:* Depicts the relationship between a selected feature and the model's output, while also marking potential interaction effects with other variables.
  - *Decision Plot:* Particularly relevant for tree-based models, tracing the sequential influence of features as they combine to yield a final prediction.
- **LIME Visualizations:**
  - Typically presented as bar charts that display feature weights in the local surrogate model, indicating positive or negative contributions to a prediction.
  - While less comprehensive than SHAP for global analysis, LIME visualizations remain effective for quick and targeted, instance-level interpretation.
- **User Interfaces:**
  - **Interactive Dashboards:** Frameworks such as SHAP's integration with `Plotly`, or broader platforms like `Streamlit` and `Dash`, enable users to explore predictions interactively and examine patterns in real time.
  - **Custom Interfaces for Domain Experts:** In applied domains, tailored visualization tools can significantly improve usability, for example, dashboards for agronomists or farmers that drawing attention to high-risk zones on maps or illustrate how particular features influence expected yield.

2.5 Blockchain Technology for Data Security

2.5.1 Fundamentals of Blockchain

Blockchain is a decentralized and distributed ledger system designed to provide secure, transparent, and tamper-resistant record-keeping without reliance on a central authority. At its core, it operates as a continuously expanding chain of data records, known as *blocks*, which are linked together through cryptographic hashing [60].

- **Block Structure:** Each block contains a set of transactions, a timestamp, a cryptographic hash of the preceding block, and a nonce used in consensus mechanisms such as Proof of Work.

- **Chaining Process:** Blocks are connected in sequence, such that altering the contents of one block would require simultaneous modification of all subsequent blocks, making tampering computationally prohibitive.

- **Decentralized Network:** The ledger is maintained collectively by a distributed network of nodes, each of which stores a full copy of the blockchain, thereby avoiding single points of failure.

- **Consensus Mechanisms:** Protocols such as Proof of Work or Proof of Stake enable participating nodes to reach agreement on the validity of transactions and the addition of new blocks.

- **Transparency and Immutability:** Once data is validated and recorded, it becomes immutable and publicly verifiable, ensuring both trustworthiness and long-term integrity.

2.5.2 Types of Blockchain Networks

Blockchain systems can be classified according to their access policies and governance structures. Each category reflects a different balance between decentralization, performance, and control, which determines their suitability for specific applications [60].

- **Public Blockchains:**
  - Open for anyone to join, read, or validate transactions (e.g., Bitcoin, Ethereum).
  - Operate in a fully decentralized environment.
  - Provide strong transparency and security, though often at the expense of scalability and energy efficiency.
- **Private Blockchains:**
  - Participation is restricted to approved or invited members.
  - Typically governed by a single organization or administrative entity.
  - Enable faster transaction throughput and improved privacy, but reduce the level of decentralization.
- **Consortium Blockchains:**
  - Managed collectively by a group of organizations or institutions.
  - Aim to strike a balance between decentralization and efficiency.
  - Well-suited to collaborative sectors such as supply chains, healthcare networks, or agricultural cooperatives.

The selection of an appropriate blockchain model ultimately depends on the requirements of the application, including its needs for trust, transparency, efficiency, and governance.

2.5.3 Security and Privacy Features

Blockchain technology establishes a robust framework for secure and reliable data management through its cryptographic foundations and distributed architecture. The following features are central to safeguarding sensitive information and preserving the overall integrity of the system [61].

- **Data Integrity:** Each block incorporates a cryptographic hash of the preceding block, creating a chain that is resistant to tampering. Any attempt to alter a block would invalidate the subsequent sequence unless consensus across the network is re-established.

- **Authentication of Participants:** Digital signatures verify the identity of transaction initiators, ensuring that only authorized entities are able to submit valid records.
- **Confidentiality of Information:** While most public blockchains operate transparently, sensitive data can be protected through encryption or stored off-chain, a practice particularly common in private and consortium-based networks.
- **Non-Repudiation of Transactions:** Once a transaction is confirmed and cryptographically signed on the blockchain, the originator cannot plausibly deny having initiated it.
- **System Availability:** Because the ledger is replicated across multiple nodes, the network remains operational and data accessible even in the presence of node failures or malicious attacks.
- **Auditability and Traceability:** Transactions are permanently recorded with time stamps, enabling full traceability and facilitating regulatory or organizational audits.

2.6 Conclusion

This chapter presented the key concepts that form the foundation of this thesis, including Data Science, Machine Learning, Explainable AI, and Blockchain technology. These topics provide the theoretical and methodological basis for the approaches developed in the later chapters. The next chapter reviews smart predictive agriculture and examines the current state of the art in smart farming. It identifies important research gaps and practical challenges, setting the stage for the proposed contributions of this work.

Chapter 3

Smart Agriculture: State of the Art

3.1 Introduction

Agriculture is entering a period of rapid transformation as established farming practices intersect with emerging technologies such as artificial intelligence, data science, and blockchain. This chapter examines the evolution of agriculture from conventional methods to modern, data-driven systems, with attention to the key challenges and technological advances shaping current developments. Traditional farming methods often face limits in productivity, exposure to climate variability, and risks related to data security, creating a clear demand for innovative solutions. Recent progress, including AI-based crop selection techniques, predictive approaches for greenhouse production, and blockchain frameworks for secure and transparent data management, is beginning to address these pressing needs. The chapter first introduces the global importance of agriculture and the main constraints that continue to affect traditional systems, including limited yields, changing climate conditions, and pressures on natural resources. It then examines the role of artificial intelligence and data-driven methods in improving farming practices, describing the principal technologies, data requirements, and their influence on productivity and sustainable management.

Subsequent sections provide a detailed review of three key areas. The first explores crop selection systems, assessing current methods, their strengths and limitations, and the research gaps that remain. The second focuses on crop yield prediction in greenhouse environments, outlining existing approaches and opportunities to improve forecast accuracy and adaptability. The final section discusses blockchain applications in different sectors, with attention to their potential to enhance security, transparency, and trust in collaborative farming networks.

3.2 Traditional Agriculture: Challenges and Limitations

3.2.1 The Global Importance of Agriculture

Agriculture holds a central position in shaping global socioeconomic development, ensuring food and nutritional security, and supporting environmental sustainability. As one of the oldest and most essential human activities, it continues to provide the foundation for survival and well-being across all regions of the world. At present, farming directly supports the livelihoods of about 2.5 billion people, with the majority living in rural areas of developing nations [62]. Beyond providing food, the agricultural sector contributes significantly to economic growth, representing around 4% of global Gross Domestic Product (GDP). In many low-income countries, this share often rises above 25%, demonstrating its critical role in national development and poverty reduction strategies [7].

Economically, agriculture continues to serve as the world's largest source of employment, sustaining the livelihoods of an estimated 892 million people as of 2022 and accounting for approximately 26.2% of total global employment [6]. The sector's significance is even more pronounced in certain regions: in Africa, nearly 48% of the population is employed in agriculture, while in South Asia the proportion remains above 39% [63]. Employment patterns over the period 2020–2025 are summarized in Table 3.1. These figures show that agriculture is both a driver of economic activity and a key factor in reducing poverty and maintaining rural stability. [64].

The resilience of agricultural systems has been particularly evident during recent global disruptions, such as the COVID-19 pandemic, when the sector acted as a buffer against economic shocks and maintained relative growth at a time when many other industries contracted [65]. Furthermore, the global distribution of agricultural employment aligns closely with regions experiencing the highest levels of food insecurity, a correlation illustrated in Figure 3.1 [66].

Table 3.1: Agricultural employment as a share of total employment (2020–2025) [1].

Year	Global (%)	Sub-Saharan Africa (%)	South Asia (%)
2020	27.0	54.0	43.0
2021	26.5	53.5	42.5
2022	26.2	48.0	40.0
2023	25.8	47.5	39.5
2024	25.5	47.0	39.0
2025	25.2	46.5	38.5

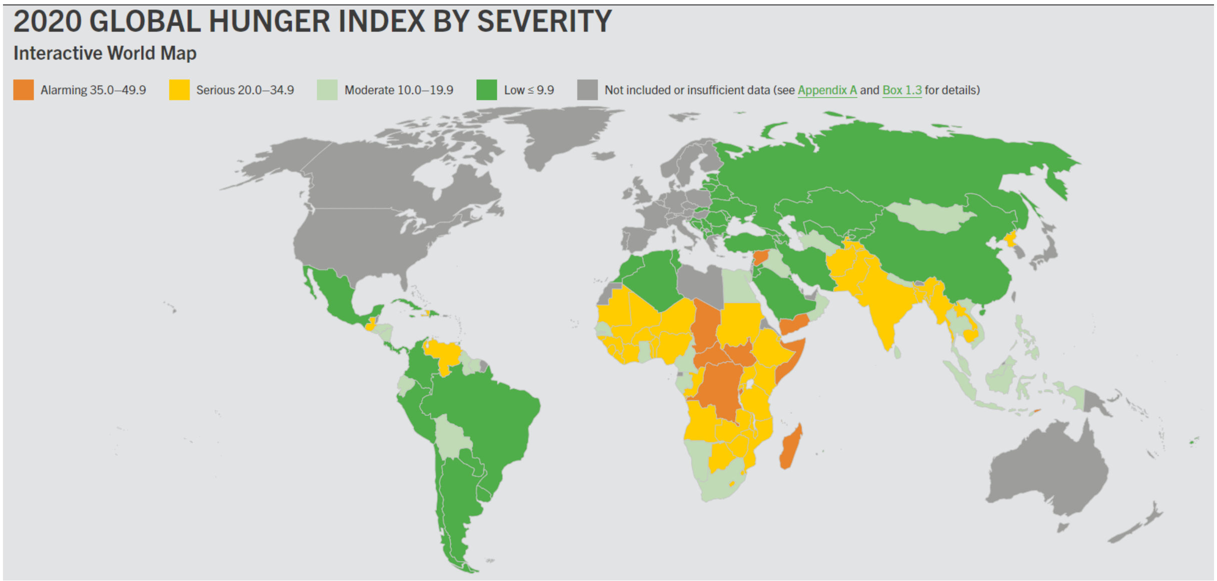

2020 GLOBAL HUNGER INDEX BY SEVERITY
Interactive World Map
Alarming 35.0–49.9
Serious 20.0–34.9
Moderate 10.0–19.9
Low ≤ 9.9
Not included or insufficient data (see Appendix A and Box 1.3 for details)

Figure 3.1: Global Hunger Index by severity, 2020 [3].

Agriculture's contribution to Gross Domestic Product (GDP) varies markedly across countries, reflecting differences in income levels and structural dependence on the sector. In 2022, agriculture represented approximately 4.1% of global GDP, yet this aggregate figure masks substantial disparities: in low-income countries, the sector's share can reach or exceed 24%, while in high-income economies it averages only about 1.3%. These contrasts underscore the continued centrality of agriculture in driving economic development and reducing poverty within the world's most vulnerable regions (see Table 3.2).

Environmentally, agriculture exerts a profound influence on global ecosystems, contributing essential services such as soil formation, carbon sequestration, water regulation, and

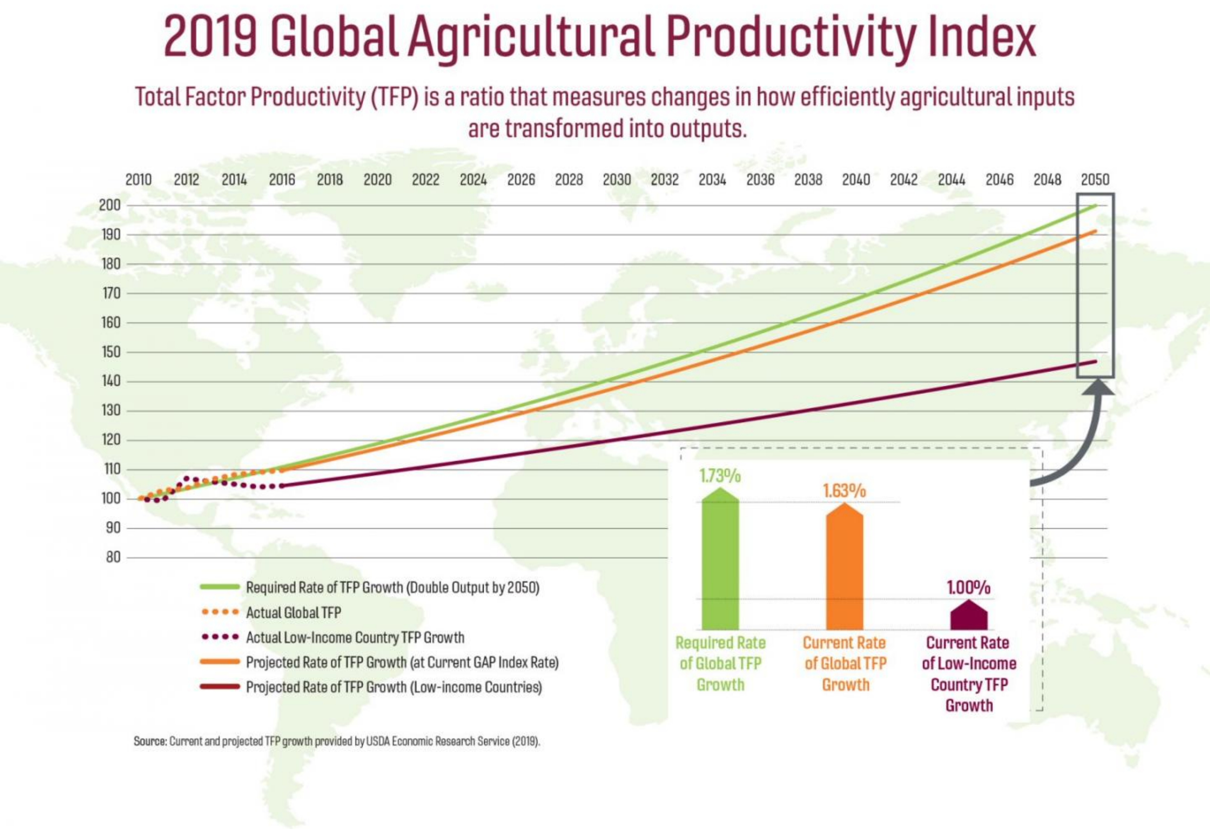
Table 3.2: Agriculture, forestry, and fishing value added as a share of GDP (2020–2025) [2].

Year	Global (%)	Low-Income Countries (%)	High-Income Countries (%)
2020	4.3	25.0	1.5
2021	4.2	24.5	1.4
2022	4.1	24.0	1.3
2023	4.0	23.5	1.2
2024	3.9	23.0	1.1
2025	3.8	22.5	1.0

the maintenance of biodiversity [67, 68]. Agricultural activity shapes landscapes across more than one-third of the planet's land surface, linking farming practices directly to questions of long-term environmental sustainability [69]. While sustainable management techniques can enhance carbon sequestration and mitigate climate change impacts [70, 71], the sector continues to face the pressing challenge of reconciling the demand for higher food production with the protection of soil quality, freshwater resources, and biological diversity. Current projections suggest that growth in total factor productivity (TFP) is lagging behind the pace required to meet the goal of doubling global agricultural output by 2050, with the shortfall being most acute in low-income countries (see Figure 3.2). This widening productivity gap reinforces the need for innovation and the adoption of strategies that enable sustainable intensification [72].

From a nutritional standpoint, agriculture remains central to ensuring global food security, with worldwide food demand expected to increase by nearly 70% by 2050. Yet, despite notable advances in technology and productivity, hunger continues to affect large segments of the population. In 2023, it was estimated that 733 million people experienced hunger, with the highest prevalence occurring in regions where agricultural livelihoods are most widespread. These enduring disparities in food availability and nutritional outcomes underscore the dual challenge of expanding production while at the same time fostering more equitable and resilient food systems [73, 74].

Agriculture is fundamental to economic development, rural livelihoods, global food security, and the health of the environment. Meeting the ambitious goals of reducing hunger, fostering economic growth, and ensuring ecological sustainability will require sustained



Source: Current and projected TFP growth provided by USDA Economic Research Service (2019).

Figure 3.2: 2019 Global Agricultural Productivity (GAP) Index [4].

innovation and carefully directed investment, particularly in regions where vulnerabilities are most acute.

3.2.2 Key Types of Agricultural Challenges

The agricultural sector is confronted with a wide range of complex and interdependent challenges that can be grouped into environmental, economic, and technological domains. These interconnected issues exert significant influence on global productivity, long-term sustainability, and the overall resilience of agricultural systems [11].

Environmental Constraints: Climate change represents one of the most pressing threats to agricultural productivity, exerting a direct influence on both crop performance and the stability of farming systems. Rising average temperatures, shifts in precipitation patterns, and the growing frequency of extreme weather events, such as droughts, floods, and heatwaves, have been shown to reduce yields, compromise crop quality, and place additional pressure on farm incomes [75, 76]. Empirical studies suggest, for instance, that an increase of 2.15–4.13 could lower wheat yields by approximately 9.14–10.20% in certain regions [77]. Beyond yield reductions, extreme weather accelerates processes of soil erosion and land degradation, diminishing the availability of fertile land and thereby threatening long-term food security [78, 76]. At the same time, intensive agricultural practices often intensify these problems, contributing to biodiversity loss, declining soil health, and weakened ecosystem resilience [79].

Economic Pressures: Agriculture remains highly sensitive to economic fluctuations, particularly in relation to market volatility, unstable commodity prices, and uneven access to financial resources and infrastructure. Since 2020, global food prices have risen by roughly 30%, a significant rise primarily attributed to disruptions caused by the COVID-19 pandemic as well as ongoing geopolitical tensions. These dynamics have destabilized food supply chains and reduced affordability for consumers worldwide [80]. The impact is especially severe for smallholder farmers, who constitute a significant share of global producers. Their vulnerability comes from limited access to formal markets, low bargaining power, and long-term underinvestment in basic inputs and rural infrastructure, which together reduce their ability to adapt and stay competitive. [81].

Technological Constraints: Although agricultural technologies are advancing at an fast-growing rate, their adoption across the sector remains highly uneven, particularly in developing regions. Persistent barriers such as inadequate infrastructure, high costs, and shortages of technical expertise continue to limit the reach of these innovations. A large proportion of farms worldwide, particularly small-scale farms in low- and middle-income countries, still lack reliable digital connectivity, which restricts their capacity to benefit from precision agriculture, smart farming tools, and data-driven decision support systems [82]. This digital divide represents a significant obstacle to sustainable productivity growth. Furthermore, the integration of advanced technologies often requires substantial capital investment, dependable data infrastructure, and specialized expertise, resources that are rarely accessible to smallholder farmers [83].

3.2.3 Traditional Agricultural Practices and Their Limitations

For centuries, traditional agricultural methods such as crop rotation, polyculture, agroforestry, and the use of natural soil enrichment have sustained human societies and supported ecological balance. These practices provided resilience in local food systems and contributed to the preservation of biodiversity and soil fertility. However, in the context of today's rapidly growing population, climate variability, and market-oriented production, such approaches reveal important shortcomings. While valuable for maintaining subsistence farming, they are often insufficient to meet the scale, efficiency, and stability required by modern agricultural systems [84].

Constraints on Productivity: Traditional farming systems generally produce lower yields when compared to intensified or mechanized approaches. Comparative studies suggest that smallholder farms relying on conventional techniques may achieve up to 50% less output than farms adopting modern agronomic practices, largely due to restricted input use, dependence on manual labor, and limited access to improved technologies [85].

Exposure to Climatic Variability: Conventional agricultural practices often lack the technological and infrastructural resilience needed to withstand changing climate conditions and extreme weather events. Heavy reliance on rainfall for irrigation, without the support of supplementary water management systems, makes these systems particularly vulnerable to prolonged droughts and flooding, risks that are intensifying under current climatic shifts [75].

Environmental Sustainability Challenges: While many traditional methods promote soil fertility and biodiversity, certain practices, most notably slash-and-burn agriculture, contribute to serious environmental degradation. Such methods can accelerate deforestation, soil erosion, and biodiversity loss, with repeated cycles of slash-and-burn cultivation driving long-term land degradation and ecosystem instability in tropical regions [11].

Economic and Market Barriers: Farmers relying on traditional systems frequently experience economic disadvantages due to limited integration into markets, lack of reliable market information, and weak logistical infrastructure. In addition, restricted access to financial services, modern inputs, and technical support constrains their competitiveness and reduces profitability, leaving smallholder communities economically vulnerable [81].

3.3 AI-Driven Transformation of Agriculture

3.3.1 From Traditional Practices to Intelligent Systems

The progression of agriculture from conventional methods to intelligent, technology-enabled systems represents a profound paradigm shift shaped by both scientific innovation and the growing demand for sustainable food production. Traditional farming, long reliant on manual labor and experience-based decision-making, is now increasingly complemented and, in many contexts, transformed by data-driven approaches. These approaches employ Artificial Intelligence (AI) and Machine Learning (ML) to optimize resource use, improve productivity, and strengthen the resilience and sustainability of agricultural systems [8].

Key Drivers of the Transition: Multiple forces are propelling the shift from traditional agriculture toward intelligent, technology-enabled systems. Global population growth, projected to reach 9.7 billion by 2050, is placing unprecedented pressure on food production systems. At the same time, escalating challenges such as climate change, resource scarcity, and shortages in agricultural labor require innovative and sustainable responses. Artificial Intelligence (AI) provides a suite of tools capable of meeting these demands by supporting precise resource allocation, generating predictive knowledge through advanced analytics, and automating tasks that have historically relied on intensive human labor [62, 7].

Applications of AI in Agriculture: Artificial Intelligence is now widely applied across diverse areas of agricultural practice, where it supports more efficient management and decision-making processes [9].

- *Crop Monitoring and Management:* AI-based platforms draw on satellite imagery together with data from field sensors to track crop health, anticipate yield outcomes, and detect the presence of pests or diseases at an early stage. These findings make it possible for farmers to act promptly and reduce potential losses.

- *Soil and Water Management:* Machine learning models process soil characteristics to generate recommendations for fertilization strategies, while intelligent irrigation systems regulate water distribution in response to real-time weather information and moisture levels within the soil.

- *Precision Farming:* By integrating data on soil heterogeneity, crop growth patterns, and environmental conditions, AI enables site-specific management practices that optimize input use and contribute to higher productivity.

- *Supply Chain Optimization:* Predictive tools enhance efficiency along the agricultural supply chain by anticipating demand patterns, coordinating logistics more effectively, and lowering post-harvest losses.

Impact and Future Prospects: The incorporation of Artificial Intelligence into agricultural systems has already produced measurable improvements that extend beyond experimental trials and into practical applications. Studies report that the use of AI tools has resulted in yield gains of as much as 30% while simultaneously reducing water consumption by approximately 20% in specific production contexts. Looking ahead, as digital technologies continue to evolve and become more widely accessible, the role of AI in supporting sustainable, productive, and resilient farming practices is expected to expand further, offering a pathway to address pressing challenges in food security and environmental management [9].

3.3.2 Core AI Technologies in Smart Agriculture

A number of Artificial Intelligence technologies play a central role in shaping modern smart agriculture.

1. Machine Learning and Predictive Analytics: Machine learning models are applied to large and complex agricultural datasets in order to forecast crop yields, detect emerging plant diseases, and guide the efficient use of resources. For example, predictive models have demonstrated strong accuracy in estimating crop productivity, which allows farmers to plan cultivation strategies more effectively and reduce potential losses [10].

2. Computer Vision Systems: Computer vision tools supported by AI enable continuous monitoring of crop health, soil status, and pest activity. By applying image recognition techniques, these systems can identify symptoms of plant stress or disease at very early stages, making it possible to intervene promptly and preserve both yields and quality [86].

3. Internet of Things (IoT): Networks of IoT devices gather continuous data on soil conditions, crop growth, and local climate. When combined with Artificial Intelligence, these measurements support precise management of irrigation, fertilization, and pest control, which in turn helps conserve resources and improve overall efficiency [10].

4. Cloud Computing: Cloud-based platforms provide the extensive storage capacity and computing power required to handle the vast datasets produced in modern agriculture. They also make it possible to deploy advanced AI models at scale, enabling farmers and researchers to access real-time analytics and informed decision-making tools [86].

5. Blockchain Systems: Blockchain technology strengthens transparency and accountability across agricultural supply chains. By recording transactions and data in secure, tamper-resistant ledgers, it contributes to food safety, supports quality assurance, and fosters greater trust among producers, distributors, and consumers [87].

6. Data Science Approaches: Methods drawn from data science are used to process and interpret complex agricultural datasets, allowing the discovery of patterns and relationships that would otherwise remain hidden. Such analyses guide decision-making in areas such as crop choice, market forecasting, and risk management [88].

7. Robotics and Automation: Robotics supported by AI enable the automation of tasks including planting, harvesting, and weed management. These systems perform with high accuracy and efficiency, reducing reliance on manual labor while increasing productivity, particularly in large-scale operations [89].

8. Generative AI: Generative AI models synthesize information from multiple datasets to provide tailored recommendations on crop planning, planting schedules, and resource allocation. Such tools can support farmers in adjusting practices to changing environmental and economic conditions [88].

3.3.3 Impact on Productivity and Sustainability

The integration of artificial intelligence (AI) and advanced digital technologies into agriculture has produced measurable improvements in both productivity and sustainability. These effects are becoming increasingly evident across diverse agricultural systems worldwide. By combining tools such as machine learning, remote sensing, and the Internet of Things (IoT) with data-driven decision support, farms can allocate resources more efficiently, carry out timely diagnostics, and implement adaptive management strategies that respond directly to changing environmental and production conditions.

Productivity Gains: Empirical evidence shows that farms adopting AI-enabled precision practices achieve yield increases of 18% to 34%, depending on crop type and agro-ecological conditions. A multi-country study in Asia and Sub-Saharan Africa, for example, reported average gains of 22% in smallholder rice and maize systems, largely through better timing and dosage of inputs. In addition, AI-supported pest and disease detection has been shown to reduce crop losses by 14 to 21% in key horticultural supply chains [93].

Resource Efficiency and Environmental Sustainability: AI-guided variable-rate technologies and sensor-based irrigation systems contribute to substantial reductions in input use and environmental impacts. Studies document water savings of 18 to 25% and fertilizer reductions of up to 28% without yield penalties, reflecting the benefits of more precise and adaptive management. Likewise, predictive analytics and monitoring tools strengthen integrated pest management, decreasing pesticide applications and supporting ecological resilience [93].

Supply Chain and Food Loss Reduction: The integration of cloud-based analytics with blockchain platforms has improved transparency, traceability, and logistics across agri-food supply chains. Such systems have reduced post-harvest losses by 10 to 15% through real-time tracking and optimized distribution, thereby enhancing food security and promoting more circular production models [87].

Societal and Environmental Implications: Beyond immediate gains in productivity and resource use, these digital innovations contribute to broader development objectives, including poverty reduction, climate action, and sustainable consumption. For instance, smallholders using AI-driven systems have reported average income increases of around 12% due to lower production costs and improved market access. Environmentally, smarter input management reduces nitrogen leaching and eutrophication risks, with studies noting up to 20% lower nitrate runoff in areas where smart agriculture platforms are deployed [92].

3.4 Literature Review on Crop Selection Systems

3.4.1 Descriptive Analysis

This section provides a detailed descriptive overview of existing research on crop selection systems. The discussion reviews major dimensions of the literature, including publication trends, document types, geographical distribution, and the frequency of recurring keywords. By examining these aspects, the analysis seeks to clarify how the field has developed over time and to characterize its present state. Such an approach offers a clearer understanding of dominant research themes while also pointing to areas that remain underexplored and may serve as directions for future studies.

The review began with a systematic search of the Scopus database, which is recognized as the largest source of peer-reviewed scientific literature. The search was restricted to the period between 2020 and 2024 and was carried out using a set of predefined keywords such as ”crop recommendation system,” ”crop selection system,” and ”machine learning in crop recommendation.” To preserve consistency and ensure relevance, the results were filtered to include only publications written in English, while dissertations and other non-journal sources were excluded. Applying these criteria produced a final set of 310 articles, which constitute the foundation for the analysis presented in this study.

An analysis of the yearly distribution shows a steady increase in research on crop selection systems Figure 3.3). The field began with 22 publications in 2020 and grew steadily, reaching 52 publications by 2022. In 2023, the number rose sharply to 114, a growth that can be associated with progress in machine learning, the Internet of Things, and smart agriculture technologies. Although a modest decline was recorded in 2024 with 92 publications, the volume remains well above the earlier years, reflecting the continued interest of the research community in this area.

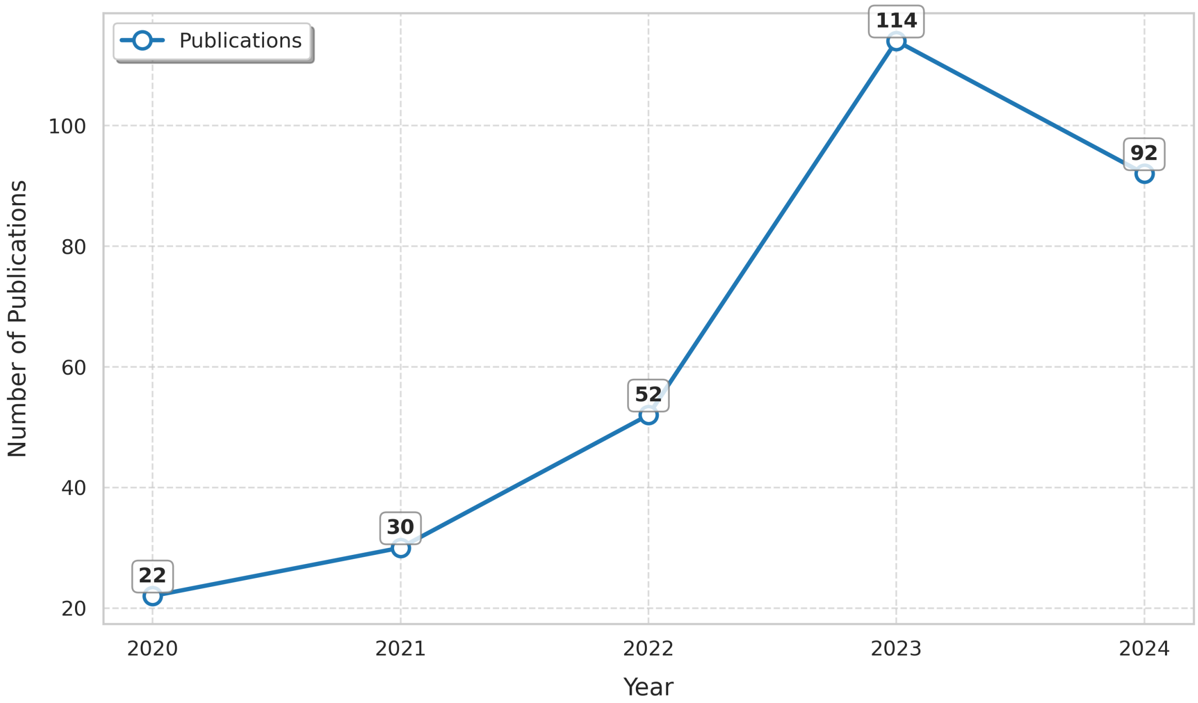


Figure 3.3: Yearly distribution of publications on crop selection systems from 2020 to 2024.

The body of literature consists of several categories of publications, including conference papers, journal articles, book chapters, and review papers, each contributing in a distinct way to the development of the field (Figure 3.4). Conference papers are the most numerous, with 217 contributions, indicating that much of the research has been shared through venues that prioritize recent advances and rapid communication of results. Journal articles make up 75 publications, offering more comprehensive, peer-reviewed studies that provide depth and methodological rigor. The collection also contains 13 book chapters that deliver specialized discussions on particular aspects of crop selection systems, along with 5 review papers that synthesize existing knowledge and outline potential research directions. Taken together, this distribution reflects the active and evolving character of the field, with conferences serving as a primary platform for presenting emerging work.

A geographical examination of the literature reveals that research on crop selection systems is distributed across a wide range of countries (Figure 3.5). India stands out with 249 publications, reflecting sustained efforts to apply agricultural technologies in response to diverse climatic and agronomic conditions. The United States follows with 14 contribu-

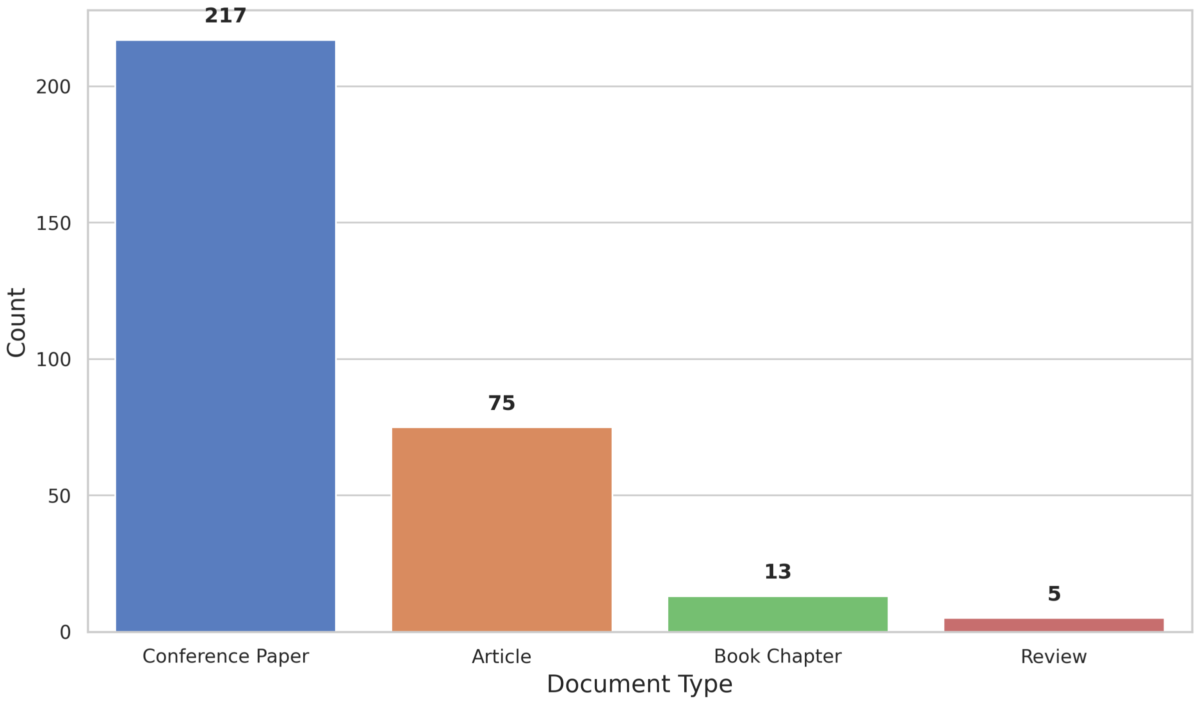


Figure 3.4: Publications classified by document type within the field of crop selection systems.

tions, while Bangladesh accounts for 11, indicating notable engagement from both regions. Other countries, including China (5), Egypt (5), and Sri Lanka (5), also demonstrate research activity directed toward improving agricultural productivity in their respective contexts. Nations such as Algeria, Iraq, and Italy, although represented by fewer studies, point to a growing interest in the topic. Moreover, contributions from Australia, France, and Ethiopia confirm that the subject has attracted attention across multiple continents, even though research intensity varies according to national capacity and available resources.

An analysis of keywords provides an overview of the main themes and recurring patterns within the literature on crop selection systems (Figure 3.6). Frequently occurring terms such as ”crops,” ”crop selection,” ”crop recommendation,” and ”learning systems” point to the central focus on applying artificial intelligence to improve decision-making in agriculture. Keywords including ”Internet of Things (IoT),” ”precision farming,” and ”machine learning” demonstrate how data-driven technologies are being incorporated into agricultural practices to promote efficiency and long-term sustainability. Additional groups of terms, such as ”soil conditions,” ”fertilizers,” and ”agricultural productivity,” draw atten-

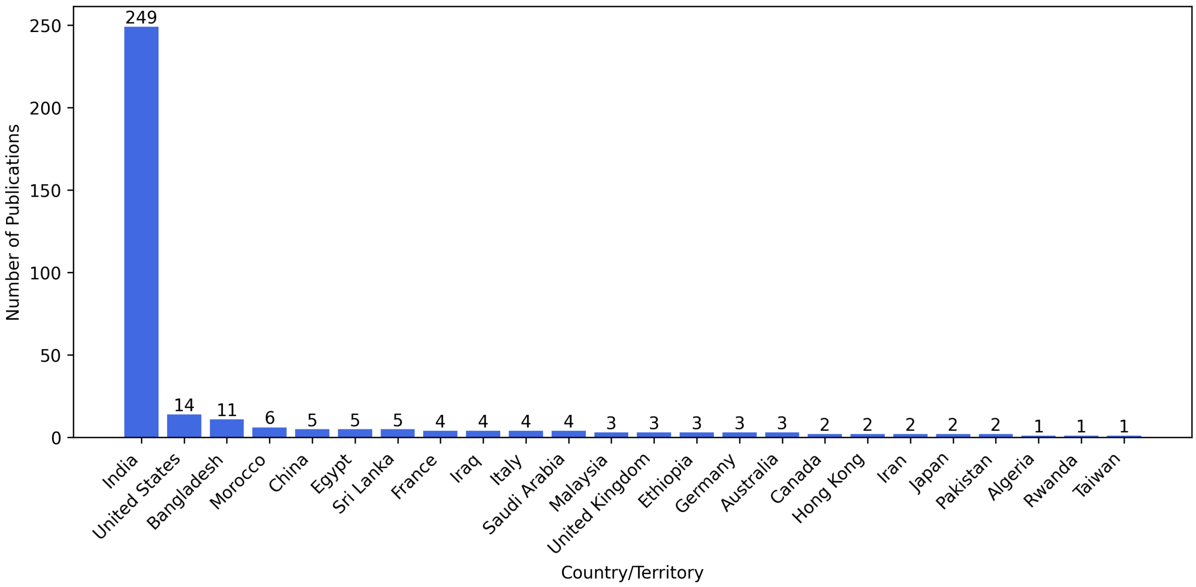


Figure 3.5: Geographical distribution of publications on crop selection systems between 2020 and 2024.

tion to the importance of environmental and resource management when designing crop selection models. The presence of keywords like ”decision trees,” ”support vector regression,” and ”genetic algorithms” shows the range of machine learning methods employed in this field. Broader interdisciplinary themes, represented by terms such as ”economics,” ”logistics,” and ”food supply,” reveal the strong connections between technological development and socioeconomic considerations. Furthermore, keywords including ”climate conditions,” ”weather prediction,” and ”soil moisture” reflect the growing attention given to external environmental factors that shape agricultural choices. Together, these clusters demonstrate how artificial intelligence, environmental sciences, and agricultural economics intersect to advance research on crop selection systems.

3.4.2 Related Works

The evolution of crop recommendation systems has progressed steadily through the application of machine learning (ML), deep learning (DL), ensemble techniques, IoT integration, and hybrid predictive frameworks. Early studies frequently relied on conventional ML algorithms, which were appreciated for their ease of implementation and capacity to establish baseline predictive performance. Within this context, Alsowaiq et al. [94] examined several classifiers, including Random Forest, Support Vector Machine (SVM),

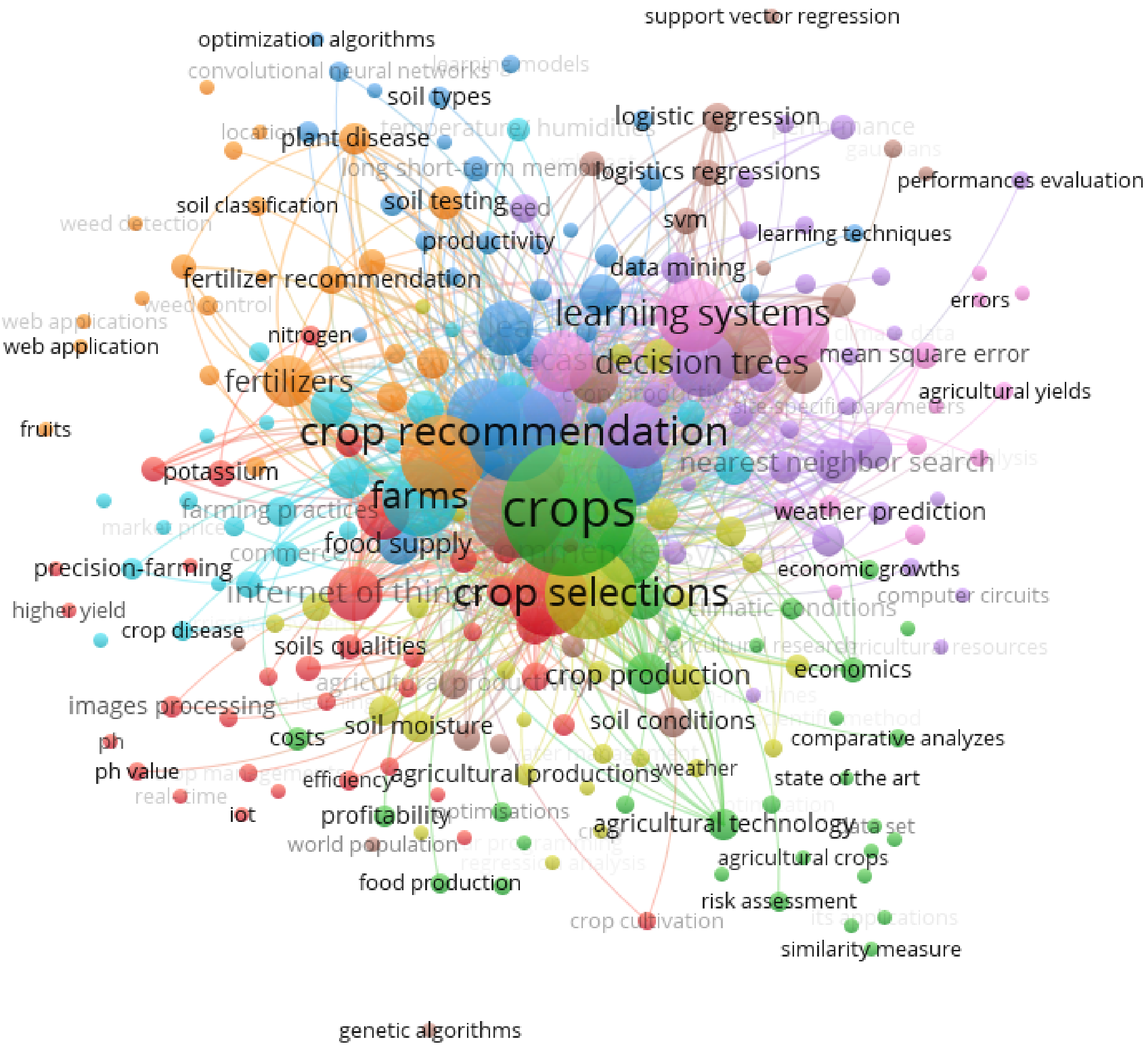


Figure 3.6: Keyword network analysis of research on crop selection systems between 2020 and 2024.

Decision Tree, K-Nearest Neighbors (KNN), and Naïve Bayes. Their analysis demonstrated that Random Forest offered the highest predictive accuracy (99.45%) in identifying appropriate crops for arid regions when standard agronomic features were used as input variables.

As research advanced, greater emphasis was placed on incorporating multiple and heterogeneous data sources to enhance both accuracy and practical utility. Palle and Raut [95] developed a multi-stage framework that combined weather forecasting, implemented through ARIMA models, with profitability assessment. Their system relied on logistic regression classifiers and achieved an accuracy of 94.2%. A limitation of their approach, however, was the dependence on synthetic crop price data, which reduced its applicability in real-world agricultural markets. Expanding on this direction, Janrao and Shah [96] introduced a return-on-investment-driven framework that employed several regression techniques, including an optimized multilayer perceptron regressor, and demonstrated highly consistent predictive performance ($R^2 > 0.999$).

Ensemble learning methods have attracted growing attention in the development of crop recommendation systems. Bandi et al. [97] used a voting ensemble that combined Decision Tree, Random Forest, and KNN models, and their approach reached an accuracy of 99.3%. While the results were strong, the absence of hyperparameter optimization showed that further refinement was still needed. Kumar et al. [98] designed a stacking ensemble that brought together Random Forest and Naïve Bayes, using Random Forest again as a meta-learner. This framework clearly outperformed the individual models, achieving an accuracy of 99.54%. Extending these efforts, Motamedi and Villányi [99] introduced Bayesian-optimized decision trees enhanced with PCA-based dimensionality reduction, which produced an F1-score of 99.54%.

The adoption of deep learning methods marked another step forward in the development of crop recommendation systems, especially through the application of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Elghamrawy et al. [100] designed an 18-layer CNN optimized with Grey Wolf Optimization to generate crop recommendations under climate change conditions, reporting predictive accuracies between 98.2% and 98.7%. The model proved highly effective in handling complex climate-related variables, but its lack of attention to socio-economic aspects limited its wider practical use. In a related study, Rani et al. [101] employed Long Short-Term Memory (LSTM) networks for weather forecasting, which subsequently improved crop recommendation when combined with a Random Forest model, achieving 97.24% accuracy.

The integration of the Internet of Things (IoT) has also become an important direction in the design of crop recommendation systems, mainly because it supports real-time data collection and decision-making. Bakthavatchalam et al. [102] developed an IoT-based precision agriculture framework that combined sensor data with machine learning classifiers, including Multilayer Perceptron (MLP), JRip, and Decision Table, and reported an accuracy of 98.2%. While effective, this framework did not provide clear mechanisms for interpretability. Building on this idea, Villanueva et al. [103] introduced IoT-driven soil analytics integrated with artificial neural networks, offering user-friendly interfaces and achieving 98.62% accuracy. In another contribution, Abdullahi et al. [104] used IoT sensor networks together with Decision Trees, which produced recommendations with an accuracy of 99.2%, although the system was limited by the availability of regional data. Palakshappa et al. [105] further advanced this line of work by combining IoT integration with Random Forest models within digital platforms designed for practical use, reaching an accuracy of 98%.

Hybrid predictive systems have increasingly made use of advanced optimization methods to improve both accuracy and adaptability in crop recommendation. Kiruthika and Karthika [106] introduced a framework that applied Improved Distribution-based Chicken Swarm Optimization (IDCSO) for feature selection together with a Weight-based LSTM for prediction, achieving 92.68% accuracy. In a related study, Mahale et al. [107] combined expectation maximization preprocessing with Random Forest classification and LSTM-based weather forecasting, which resulted in a system that produced 92.7% accuracy.

Progress has also been made in region-specific frameworks that integrate agronomic and economic considerations. Musanase et al. [108] presented a system tailored to Rwanda that used neural network-based recommendations along with rule-based fertilizer guidance, reaching 97% accuracy.

The comparative analysis in Table 3.3 shows that most earlier studies on crop selection concentrate on achieving high predictive accuracy, with reported values typically ranging between 92% and 99%, without giving attention to model interpretability. Many of these works apply machine learning classifiers such as Random Forest, Decision Tree, KNN, and neural networks. In some cases, additional modules for weather prediction, including LSTM or ARIMA models, are used to improve the quality of recommendations. A smaller group of studies experiments with ensemble approaches such as voting or stacking, which provide strong predictive performance. Despite these promising results, the limited trans-

Table 3.3: Comparison of previous studies on crop selection systems.

Study	Proposed Model	Crop Recommendation Dataset	Performance	Ensemble Learning	Interpretability
[100]	Optimized Convolutional Neural Network (CNN)	✗	Accuracy: Wheat 98.2%, Maize 98.7%, Rice 98.1%	✗	✗
[94]	Random Forest, SVC, Decision Tree, KNN, Naïve Bayes	✓	Accuracy: 99.45%, F1-score: 99.5% (RF)	✓	✗
[95]	ARIMA for weather/price prediction, one-vs-rest logistic regression for crop recommendation	✓	Classification accuracy: 94.2%; RMSE (weather): 2.25	✗	✗
[97]	Ensemble (Voting Classifier: Decision Tree, Random Forest, KNN)	✓	Accuracy: 99.3%	✓	✗
[102]	IoT+ML: Multilayer Perceptron	✓	Accuracy: 98.2% (MLP), ROC: 1.0	✗	✗
[98]	Stacking: Random Forest, Naive Bayes, Random Forest meta-learner	✓	Accuracy: 99.54%, Precision: 99.54%, Recall: 99.53%, F1: 99.52%	✓	✗
[101]	LSTM RNN for weather prediction, Random Forest for crop selection	✗	Acc: 97.24% (RF), RMSE: 5.02–8.24% (LSTM)	✓	✗
[103]	ANN	✓	Accuracy: 98.62%	✗	✗
[104]	Decision Tree, Random Forest, KNN	✓	DT: Acc 99.2%, Prec/Rec/F1: 99%	✓	✗
[96]	Optimized MLP regressor	✗	RMSE: 12.32, $R^2 > 0.999$	✗	✗
[106]	WLSTM neural network	✓	Accuracy: 92.68%, Precision: 90.88%, Recall: 91.98%	✗	✗
[108]	Neural network	✓	Accuracy: 97%, per-class F1 > 0.95	✗	✗
[107]	LSTM (weather), Random Forest (crop rec.)	✓	Acc: 92.7%, F1/Prec/Rec: 93%	✓	✗
[99]	Bayesian-optimized ensemble decision trees	✓	Accuracy: 99.5%, F1: 99.54%, Precision: 99.55%, Recall: 98.59%	✓	✗
[105]	IoT-enabled, RF, SVM, NB, DT classifiers	✓	RF: 98% (SVM: 93%, NB: 96%, DT: 89%)	✓	✗
Our proposed approach	**CS-AdaRF-SHAP**	✓	Accuracy 99.72%	✓	✓

parency of these models reduces their usefulness in agricultural practice, since farmers may be reluctant to rely on predictions that are not supported by clear explanations. To address this issue, Explainable AI (XAI) methods are needed to make model outputs more understandable and to encourage adoption in real farming environments.

3.4.3 Research Gaps and Contribution

A key gap in the current literature is the absence of an effective balance between predictive accuracy and interpretability. Many previous works achieve high performance with machine learning or deep learning models, yet these models function as black boxes and provide no explainable AI (XAI) mechanisms to clarify the reasoning behind their recommendations. This is a serious limitation in agriculture, where farmers and decision makers need transparent explanations to trust automated suggestions. Without such reasoning, it becomes difficult to justify why one crop is recommended while another is not, which discourages adoption even when accuracy is high.

Another shortcoming is the limited attention to the impact of different types of prediction errors. Most studies emphasize overall accuracy but rarely examine the consequences of specific errors. In particular, false positives are critical in a crop recommendation context. A false positive occurs when the system advises planting a crop that is unsuitable for the local soil or climate. Such an error can lead to wasted resources, lower yields, and loss of confidence in data-driven systems.

The present work addresses these gaps by proposing a crop selection framework, CS-AdaRF-SHAP, that aims to combine high predictive performance with clear interpretability while reducing false positive errors. The system uses the AdaBoost algorithm as the main classifier and Random Forest as the base learner. AdaBoost iteratively adjusts the weight of misclassified samples, forcing the model to focus on difficult cases and thereby reducing systematic mistakes such as repeated false positives. Random Forest contributes robustness against noisy data and captures complex, nonlinear relationships among soil properties, weather factors, and nutrient levels.

To overcome the black-box nature of ensemble models, the framework integrates SHapley Additive Explanations (SHAP) to provide transparent reasoning for each recommendation. SHAP produces feature-level explanations that show how variables such as nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall influence the predicted suitability of each crop. Two types of explanations are generated. Global explanations reveal which features generally increase or decrease the likelihood of selecting a crop, while local explanations clarify why, in a specific case, one crop is recommended over another. By combining robust prediction with detailed explanations, the proposed system supports trustworthy and informed decision making in real agricultural settings.

3.5 Literature Review on Data-Driven Crop Yield Prediction

Accurate prediction of crop yields is a key factor in ensuring food security and supporting the economic stability of agricultural systems worldwide. Reliable forecasts allow farmers, policymakers, and supply chain stakeholders to plan cultivation schedules, manage resources efficiently, and reduce production risks. However, yield prediction remains a complex task due to the interaction of many variables, including soil conditions, climate patterns, farming practices, and crop-specific growth characteristics [109, 110].

Traditional yield estimation methods have generally relied on manual field inspections, historical yield records, and expert judgment. While these approaches have been widely used, they are prone to inconsistencies and often fail to capture the intricate relationships between environmental factors and plant growth [111]. In greenhouse and open-field settings alike, such methods may produce inaccurate forecasts, limiting their usefulness for precision farming and large-scale production planning.

Recent developments in smart agriculture are transforming the way crop production is monitored and managed. Technologies such as the Internet of Things (IoT), Artificial Intelligence (AI), blockchain-based systems, and robotics have introduced new opportunities to collect and analyze large volumes of agricultural data [112, 113]. Within this context, Machine Learning (ML) has emerged as a critical tool for data-driven yield prediction. By learning from historical and real-time data, ML models can uncover hidden patterns and complex relationships, enabling more accurate and timely predictions. These advances support more efficient resource allocation, improved decision-making, and farming practices that are both economically viable and environmentally sustainable [114, 115].

This section reviews published research on crop yield prediction to provide context for recent advances in modeling strategies across different crops. Several studies have applied a range of machine learning and deep learning techniques to improve forecasting accuracy. For example, the authors of [100] evaluated multiple machine learning and deep learning methods for winter wheat yield prediction. Using a dataset that combined weather, soil, and phenological information from 271 German counties collected between 1999 and 2019, they compared deep neural networks (DNNs), convolutional neural networks (CNNs), decision trees, random forests, XGBoost, and linear regression. Among these models, the CNN achieved the best performance, reducing RMSE by 7–14%, lowering MAE by 3–15%, and improving correlation coefficients by 4–50% compared with the other approaches.

In [94], an early yield estimation method for tomato crops was introduced by combining Decision Tree Ensembles (DTE) with data captured by Unmanned Aerial Vehicles (UAVs). Their DTE-Bag model achieved a prediction accuracy of 92.5%, demonstrating the potential of UAV-based data for supporting farm management decisions.

In another study [95], a transformer-based model was applied to rice yield prediction using satellite observations and climate variables. The model outperformed four other machine learning techniques (LASSO, RF, XGBoost, and AtLSTM), achieving the highest R^2 (0.78), the lowest RMSE (0.44 t/ha), the lowest MAPE (16.56%), and an overall accuracy of 0.72. The authors noted, however, that soil characteristics, tillage practices, and fertilizer inputs were not included as predictive features, which may limit the model's generalizability.

Similarly, [97] proposed a hybrid framework for greenhouse yield forecasting that combined outputs from a biophysical model (Tomgro) with a deep learning model. The Tomgro component used environmental inputs such as temperature, humidity, and light, while the CNN-RNN network was trained on historical yield and environmental data. The combined approach delivered the highest accuracy, with mean RMSE, R^2, and Nash–Sutcliffe efficiency (NSE) values of 17.69 ± 3.47.

Finally, in [102], climatic factors including temperature, rainfall, and solar radiation were used to estimate national wheat yields. Among the tested models, Random Forest achieved the best performance, yielding RMSE values of 9.1% for Brazil, 6.7% for France, and 6.4% for Russia.

In [98], rice yield prediction was evaluated using multiple linear regression (MLR), random forest (RF), and a traditional regression (TR) method, based on agronomic traits such as plant density and plant height. Field experiments conducted in Jilin Province, China, showed that the RF model achieved the highest accuracy under varying conditions.

In [110], several regression algorithms were applied to tomato yield prediction, including Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Lasso Regression, and Linear Regression, across multiple datasets. Among these datasets, the third proved to be the most reliable and comprehensive. Although RF required more computational resources, it delivered the best predictive accuracy, while KNN and Lasso Regression provided competitive performance with lower computational costs.

Building on the results of previous studies, the present research seeks to improve crop yield prediction by advancing modeling techniques with an emphasis on ensemble learning. A Stacked Ensemble Model is applied to combine multiple algorithms for daily yield estimation, using tomato production as a case study. This approach leverages the strengths of different models while reducing their individual limitations, leading to predictions that are more robust and reliable across diverse datasets.

3.6 Literature Review on Blockchain Applications for Data Security

The rapid growth of smart agriculture has increased the importance of collecting, storing, and protecting large volumes of digital data. In modern farming systems, Internet of Things (IoT) devices continuously generate real-time information that supports data-driven decision-making. Reliable raw data forms the backbone of these systems, and secure storage is essential to ensure that decisions are accurate, traceable, and resistant to tampering.

Similar concerns about data security and integrity extend well beyond agriculture to many scientific and industrial fields. Researchers handling large datasets often face significant challenges related to privacy, transparency, and trust. Blockchain technology has emerged as a promising solution to these issues because of its decentralized architecture, which ensures immutability and strengthens data integrity. By recording transactions across a distributed network, blockchain provides an auditable and tamper-resistant ledger that enhances trust among stakeholders.

Despite these advantages, blockchain has technical and economic limitations when applied to data-intensive tasks. Storing large datasets directly on-chain is impractical because of high storage costs, limited capacity, and slower transaction verification as file sizes increase. These constraints reduce performance and hinder the scalability of blockchain-based systems, making it unsuitable as a standalone solution for applications that require frequent handling of large files.

To address these limitations, recent studies have explored integrating blockchain with distributed storage systems such as the InterPlanetary File System (IPFS). IPFS enables efficient off-chain storage by distributing files across a peer-to-peer network while maintaining a unique cryptographic hash for each file. The hash is stored on the blockchain, creating a permanent and verifiable link between the ledger and the stored content. This hybrid approach allows blockchain to maintain its strengths in security and transparency, while IPFS provides scalable and cost-effective file management.

Several research efforts illustrate the effectiveness of this integration. For example, [94] proposed a framework that combines blockchain with IPFS to enhance the management of Open Educational Resources (OER). In their system, providers create and share educational content, while consumers access and use these materials. Providers generate a digital contract containing metadata such as the resource title, creation time, creator identity, and content hash. The resource itself is uploaded to IPFS, and its hash is permanently recorded on the blockchain. This design ensures data provenance and provides a secure, verifiable record of the resource without overloading the blockchain with large files.

A related study in [95] applied the same blockchain–IPFS combination to the field of healthcare data management. In this design, sensitive medical files are encrypted and stored on IPFS, while associated metadata, including file identifiers, patient IDs, and cryptographic hashes, are recorded on an Ethereum blockchain. Only the hashes are stored on-chain, which reduces storage requirements and transaction costs. When a file is requested by an authorized party such as a doctor, technician, or patient, the system retrieves it from IPFS and compares its hash with the blockchain entry. If the two values match, the file is decrypted and made available to the user, thereby ensuring both data integrity and patient privacy.

In the field of e-learning, the security and privacy of Electronic Learning Records (ELRs) remain a challenge, largely because of dependence on third-party storage platforms. To address these risks, [98] proposed MOOCs Chain, a blockchain-based framework designed for the management of ELRs in Massive Open Online Courses (MOOCs). In this model, only course providers are required to join the blockchain network, while learners remain anonymous to preserve their privacy. Core components of ELRs are stored on the blockchain, whereas the original course materials are kept on IPFS. The framework also introduces inter-authentication, anonymization, and strong mechanisms to ensure secure storage and controlled distribution of ELRs.

Similarly, [102] introduced a blockchain-based prototype for supply chain management, aiming to improve transparency, traceability, scalability, and the security of third-party transactions. Since storing large records directly on the blockchain is inefficient, the authors employed IPFS as a distributed storage layer. This hybrid setup enabled process automation and supported secure and reliable data exchange across different points in the supply chain.

A review of these studies shows that IPFS is widely adopted as the distributed storage layer in blockchain-based data management systems. Its popularity stems from its scalability, efficient peer-to-peer architecture, and flexibility to integrate with diverse applications that require secure and verifiable storage of large files.

Building on this foundation, our work proposes a unified framework that connects multiple agricultural sites with government institutions through a secure blockchain-enabled architecture. The framework is designed to manage the collection, storage, and controlled exchange of greenhouse data while preserving data integrity and confidentiality. Access is strictly regulated so that one site cannot retrieve or modify another site's information without explicit authorization, while institutional oversight ensures that data sharing remains transparent and accountable.

3.7 Conclusion

This chapter reviewed the development of smart agriculture, outlining the limitations of traditional farming and the potential of digital technologies to overcome these challenges. The next chapter presents the first contribution of this thesis, an interpretable crop selection system designed to combine predictive accuracy with explainability to support reliable farming decisions.

Chapter 4

Contribution 1: Interpretable Crop Selection for Optimized Farming Decisions

4.1 Introduction

Deciding *what to plant* is the first and most fundamental challenge in agriculture, as it shapes the entire production cycle and strongly influences profitability, resource management, and environmental sustainability. Farmers must make this decision before any other management step, yet traditional approaches to crop selection often fall short when facing changing conditions such as soil variability, shifting climate patterns, and fluctuating nutrient availability. Artificial intelligence (AI) offers considerable potential to support this crucial choice by analyzing diverse sources of information and adapting to complex environmental factors. Nevertheless, a major obstacle remains: most AI-based systems do not provide clear explanations for their recommendations. Farmers, whose income and long-term planning depend on this initial decision, are often reluctant to adopt systems that deliver predictions without transparent reasoning, even when those predictions are statistically sound.

In this context, interpretability means the ability to explain why a specific crop is recommended. For instance, a model should indicate whether factors such as nitrogen, phosphorus, potassium levels, rainfall, or temperature had the greatest influence on its decision. Transparency is closely related and refers to making the internal reasoning of the model understandable, such as showing how much each factor contributes to the final recommendation. Without interpretability and transparency, AI systems may be seen as "black boxes," which weakens trust and limits their practical value. Many current approaches focus mainly on predictive accuracy while giving little attention to explainability, creating a gap between technical performance and farmers' readiness to use these tools. Agricultural decisions demand accurate predictions together with clear explanations that farmers can understand and use in their planning.

To answer the fundamental question of *"what we plant?"* and respond to the challenge of providing both accuracy and transparency, this chapter presents a crop selection system that integrates strong predictive performance with clear interpretability. The proposed CS-AdaRF-SHAP framework combines an ensemble learning approach with explainable AI techniques to deliver recommendations that are both dependable and understandable. Adaptive boosting is employed to improve predictions by concentrating on harder-to-classify cases, which strengthens the model's ability to handle varied environmental conditions. In addition, feature attribution methods are applied to measure the influence of variables such as soil nutrients and climate conditions on final outcomes.

The remainder of this chapter is organized as follows. First, an exploratory analysis of the dataset is presented to describe its key characteristics. Next, the preprocessing steps and feature selection process are explained in detail. This is followed by a comparison of several machine learning models for crop selection to assess and validate the preprocessing strategy. Finally, the proposed system is introduced and evaluated with respect to both predictive accuracy and clarity of explanations, and the chapter concludes with a brief summary of the main outcomes.

4.2 Methodology Overview

The proposed system CS-AdaRF-SHAP is designed to provide strong predictive accuracy while also offering clear and practical explanations of its recommendations. This dual focus supports the use of AI-based tools in agricultural decision making and helps farmers understand the reasoning behind each crop suggestion. As shown in Figure 4.1, the system architecture is organized into two main phases: *offline phase* and *online phase*.

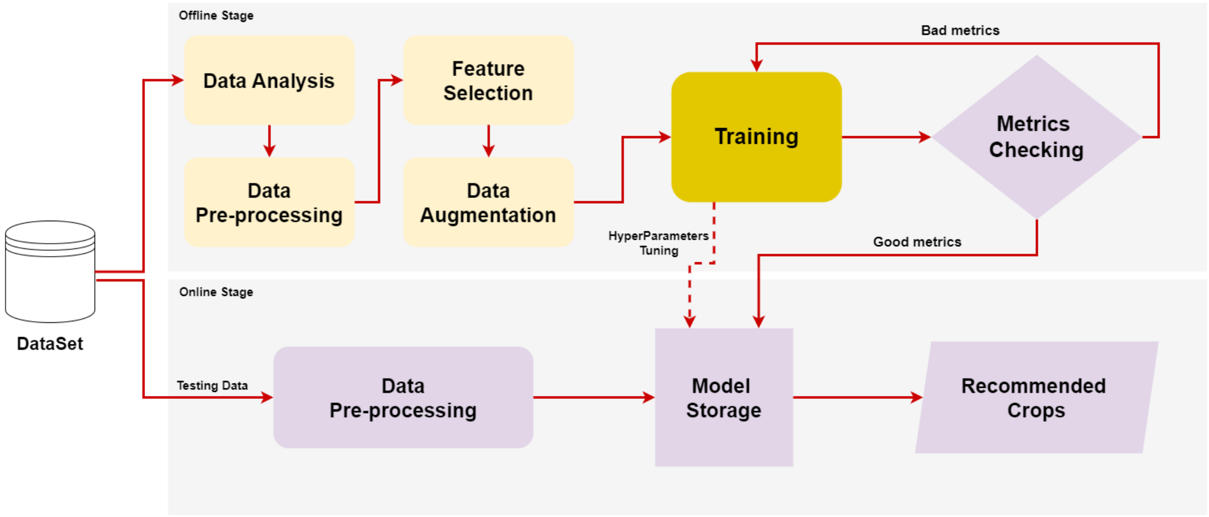


Figure 4.1: The general architecture of the proposed system.

Offline Phase The offline stage focuses on constructing a reliable crop selection model. The process begins with an exploratory data analysis (EDA) to check the distribution of each variable, study univariate and bivariate analysis, and measure correlations. After this analysis, data preparation includes detecting and handling outliers, the use of Min–Max scaling to align feature ranges, and the creation of additional synthetic samples to expand each crop class from 100 to 300 records. All seven agronomic variables (nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall) are kept to maintain essential soil and climate information. To confirm that each preprocessing step contributes to better predictions, several scenarios were examined using five machine learning algorithms (RF, DT, Naïve Bayes, SVM, and KNN). The results of these tests guided the construction of the final preprocessing pipeline.

The cleaned and enriched dataset is then used to train the AdaBoost classifier, which combines a series of decision tree learners to form a strong ensemble model.

Online Phase In the online phase, the system operates in real time to provide farmers and agricultural practitioners with crop recommendations. The trained AdaBoost model processes new input data and returns suitability predictions with very low response time, allowing users to make timely decisions. Each recommendation is accompanied by interpretability measures produced with SHapley Additive exPlanations (SHAP). SHAP calculates and displays how each soil and climate variable contributes to the suggested crop, giving clear and practical explanations for every prediction. Interactive visual tools present these explanations in an accessible way, helping users compare the results with their own field conditions and build confidence in the model's guidance.

4.3 Exploratory Data Analysis

4.3.1 Data Acquisition and Characteristics

The dataset used in this study was obtained from a public repository available on Kaggle [116]. It contains a total of **2,200 records**, distributed equally across **22 crop species of agricultural importance**. Each crop class is represented by exactly 100 entries, which provides a balanced distribution across categories. This balance is particularly important for supervised learning, as it reduces bias during training and supports fair evaluation of model performance.

The crop categories represent a wide range of agronomic groups. For clarity of analysis, they can be organized into four main sectors:

- **Cereals**: Staple food crops including rice, wheat, and maize.
- **Legumes**: Protein-rich pulses such as chickpea, lentil, and kidney beans.
- **Fruits**: Seasonal and perennial fruit crops including watermelon, muskmelon, papaya, and mango.
- **Plantation or Cash Crops**: Crops of high economic value, such as cotton, jute, and coffee.

Each record in the dataset includes seven independent variables that play an essential role in determining crop growth and suitability (Figure 4.2).

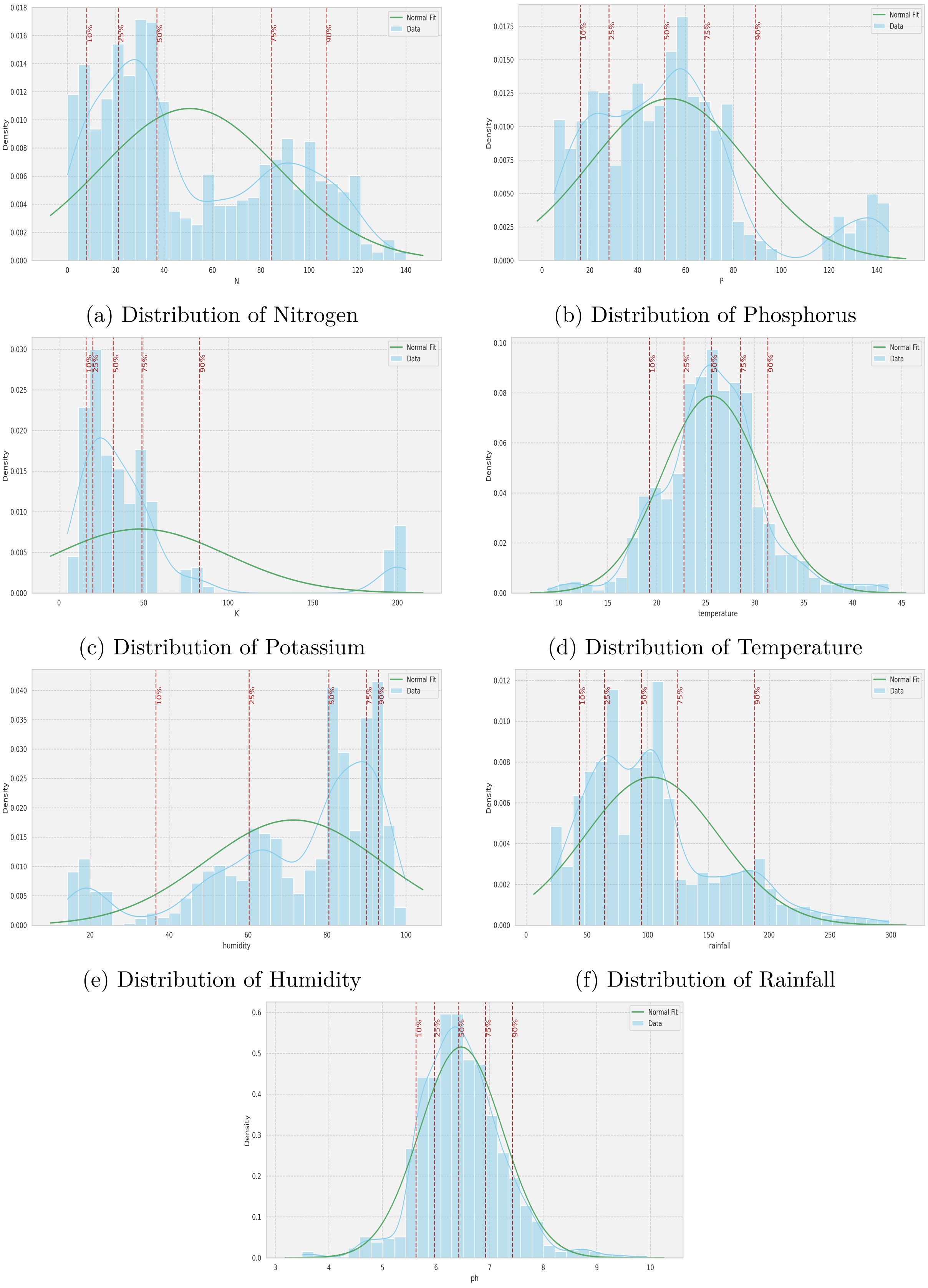
Table 4.1: Descriptive statistics of agricultural parameters.

Features	Min	Max	Median	SD	Mean	Skewness	Kurtosis
N (mg/kg)	0	140	37	36.9	50.55	0.5	-1.05
P (mg/kg)	5	145	51	33.05	53.36	1.01	0.85
K (mg/kg)	5	205	32	50.6	48.14	2.4	4.4
Temp. (°C)	8.8	43.7	25.6	5.06	25.61	0.18	1.2
Humidity (%)	14.3	100	80.5	22.3	71.48	-1	0.3
pH	3.5	9.94	6.43	0.77	6.46	0.3	1.6
Rainfall (mm)	20	299	95	55	103.46	0.96	0.6

4.3.2 Univariate Analysis and Distribution Visualization

Univariate analysis examines each variable independently to describe its general behavior and statistical properties. This step helps reveal the distribution, central tendency, variability, and overall shape of the data. It also assists in detecting skewed features, unusual values, or quality issues that may influence the performance of predictive models. Descriptive statistics are provided in Table 4.1, and feature-level patterns are considered in the discussion that follows.

- **Nitrogen (N):**
  - Range: 0 to 140 mg/kg; Median = 37 mg/kg.
  - The distribution is mildly right-skewed (skewness $\approx$ 0.5), and kurtosis is negative (-1.05), indicating a relatively flat distribution with few extreme values.
  - *Interpretation:* Over half of the samples show nitrogen values at or below 37 mg/kg, suggesting that many crops in the dataset grow under low to moderate N conditions. The slightly right-skewed shape indicates the presence of soils with higher nitrogen, which are likely associated with crops that require greater nutrient input.
- **Phosphorus (P):**
  - Range: 5 to 145 mg/kg; Mean = 53.36 mg/kg; SD = 32.99 mg/kg.
  - Right-skewed distribution (skewness $\approx$ 1.01) with some high-value outliers.



(a) Distribution of Nitrogen

(b) Distribution of Phosphorus

(c) Distribution of Potassium

(d) Distribution of Temperature

(e) Distribution of Humidity

(f) Distribution of Rainfall

(g) Distribution of Soil pH

Figure 4.2: Distributions of soil macronutrients (N, P, K), environmental factors (temperature, humidity, rainfall), and soil acidity (pH).

  - *Interpretation:* The mean and median place most samples in a moderate phosphorus range. The right skew reveals fewer but notable cases of high-P soils, which may correspond to crop groups with stronger phosphorus demand. This pattern indicates that while moderate P conditions are common, certain crops adapt to higher levels.

- **Potassium (K):**
  - Range: 5 to 205 mg/kg; Median = 32 mg/kg; Mean = 48.15 mg/kg.
  - Strongly right-skewed (skewness = 2.40) and leptokurtic (kurtosis = 4.4).
  - *Interpretation:* With a median near 32 mg/kg, most samples fall into a low to moderate potassium range, which suits many of the crops represented. The pronounced skewness and high kurtosis reflect a small fraction of samples with very high potassium, likely linked to crop types requiring stronger K availability.

- **Temperature (°C):**
  - Range: 8.83 to 43.68°C; Mean = 25.62°C.
  - Near-normal distribution with slight right skew.
  - *Interpretation:* Most values lie between 22 and 29 °C, which represents optimal conditions for many of the crops included in the dataset. The higher values reflect environments suitable for heat-tolerant crops, while the lower values correspond to species adapted to cooler climates.

- **Humidity (%):**
  - Range: 14.26% to 99.98%; Mean = 71.48%.
  - Bimodal distribution.
  - *Interpretation:* The presence of two peaks indicates that the dataset covers both dry and humid conditions. This suggests inclusion of crops grown in arid environments as well as crops requiring high atmospheric moisture.

- **Rainfall (mm):**
  - Range: 20.21 to 298.56 mm; Mean = 103.46 mm.
  - Strong right-skewed distribution.
  - *Interpretation:* Most records fall below 150 mm, showing that many crops in the dataset grow under moderate rainfall. The presence of a few very high values reflects crops cultivated in regions with heavy rainfall.

  Soil pH:
  - Range: 3.50 to 9.94; Mean = 6.47.
  - Nearly normal distribution centered around 6.4.
  - *Interpretation:* Most samples fall within a neutral to slightly acidic range. This range is favorable for a wide group of crops. A smaller number of samples at the extremes show strongly acidic or alkaline soils, suggesting conditions suited only for crops adapted to those specific environments.

4.3.3 Bivariate Analysis

A bivariate analysis was carried out to explore the relationships between each independent variable and the target crop label. This examination helps to understand how much each numerical feature varies across different crop classes and to detect variables that may be redundant or strongly discriminative. The analysis was organized into two main parts according to the types of variables: (1) relationships between pairs of numerical variables, and (2) relationships between numerical variables and the categorical crop label.

Numerical–Numerical Analysis

To examine relationships among the numerical features and to check for possible multicollinearity, a Pearson correlation matrix was computed (Figure 4.3). The coefficient values range from -1 (perfect negative correlation) to $+1$ (perfect positive correlation). The main observations are as follows:

- **Phosphorus (P)** and **Potassium (K)** show a moderately strong positive correlation ($r \approx 0.74$). This means that higher levels of P are often accompanied by higher

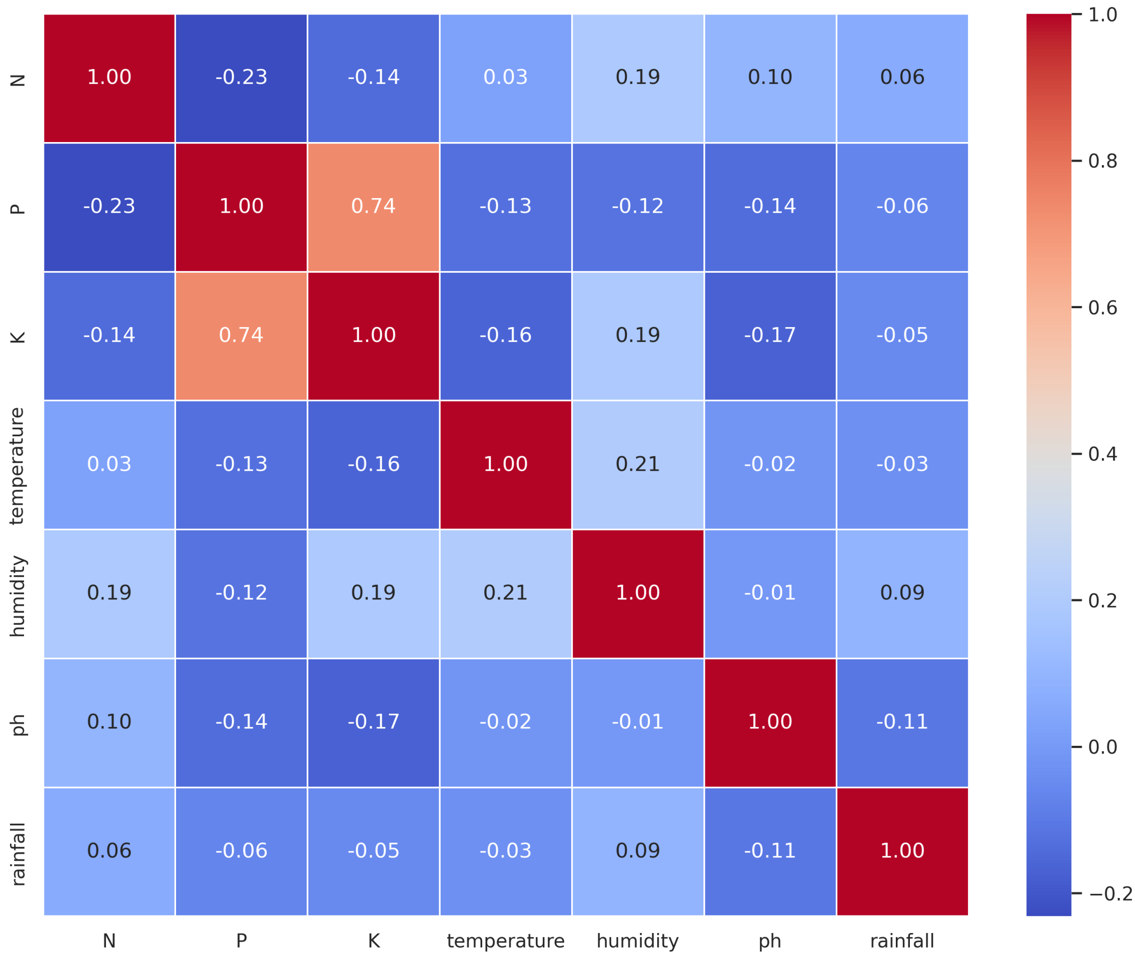


Figure 4.3: Correlation matrix showing pairwise relationships among numerical features.

levels of K across the samples. In machine learning applications, such a correlation may introduce redundancy, particularly in linear models such as Logistic Regression or Linear Discriminant Analysis, where multicollinearity can affect performance.

- **Temperature** and **Humidity** have a weak negative correlation ($r \approx -0.14$), suggesting only a slight inverse relationship between the two variables.

- Other variable pairs, including **Nitrogen** with **Rainfall** or **pH**, show negligible correlation ($|r| < 0.1$). Such low associations reduce the risk of overlapping information and help maintain model stability.

Interpretation: The correlations observed are generally low to moderate, which indicates that most features contribute distinct information. This diversity among predictors is beneficial for building machine learning models, as it reduces redundancy and supports clearer interpretation of results.

Numerical–Categorical Analysis

This analysis explored how each numerical feature varies across the 22 crop categories in the dataset. Two complementary approaches were applied to evaluate the ability of these features to separate crop types. First, one-way ANOVA was used to test whether the mean values differ significantly among crops, which is helpful for detecting linear separability relevant to models such as logistic regression. Second, Mutual Information (MI) was calculated to measure the strength of non-linear relationships between each feature and the crop labels, a property that can guide the use of non-linear algorithms such as decision trees.

1. One-Way ANOVA (Analysis of Variance):

A one-way ANOVA test was applied to check whether the mean of each numerical feature varies significantly across the 22 crop types. The results are presented in Table 4.2.

Table 4.2: One-Way ANOVA Results for Numerical Features

Feature	**F-statistic**	**p-value**
Nitrogen (N)	897.57	$< 1 \times 10^{-300}$
Phosphorus (P)	1885.66	$< 1 \times 10^{-300}$
Potassium (K)	27,238.36	$< 1 \times 10^{-300}$
pH	60.34	6.49×10^{-199}
Temperature	102.19	4.02×10^{-305}
Humidity	3103.71	$< 1 \times 10^{-300}$
Rainfall	605.53	$< 1 \times 10^{-300}$

Interpretation: All p-values are extremely small, showing that the mean values of every feature differ significantly across the crop categories. Potassium (K) recorded the highest F-statistic, making it the most discriminative variable among the tested features. Humidity and Phosphorus also display strong variation between crops, indicating that these factors play an important role in distinguishing growing conditions.

2. Mutual Information (MI):

Mutual Information (MI) was used to measure how strongly each numerical feature is related to the crop labels while also capturing possible non-linear relationships. The ranked scores in Table 4.3 show the strength of these feature–label connections.

Table 4.3: Mutual Information Scores for Feature–Label Relationships

Feature	**Mutual Information Score**
Humidity	1.730
Rainfall	1.637
Potassium (K)	1.630
Phosphorus (P)	1.298
Temperature	1.018
Nitrogen (N)	0.993
pH	0.686

Interpretation: Humidity, Rainfall, and Potassium show the strongest connections with crop type, each with MI values above 1.6. Nitrogen, although important in the ANOVA results, ranks lower here, which points to a relationship that is more linear and therefore less captured by MI. The lowest score belongs to pH, suggesting that this feature varies less across crops and is generally more uniform in its effect.

3. Feature Distributions Across Crops:

To support the statistical findings, graphical summaries were prepared to show how the numerical features vary among the 22 crop types. Boxplots were created for each variable grouped by crop label (Figures 4.4, 4.5). These visualizations help reveal natural groupings, differences in spread, and possible extreme values.

- **Rainfall:** Crops such as rice and jute require much higher rainfall, while legumes and pulses remain tightly clustered at lower values.
- **Humidity:** Separates water-demanding crops like rice and sugarcane from dry-land crops such as lentil and gram.
- **pH:** Most crops grow best in a near-neutral pH range (around 6.0–7.5), although crops like coffee and grapes tolerate a wider range.

Interpretation: These visual patterns support the ANOVA and MI outcomes by showing clear differences in key variables across crop groups. They also help identify outliers and overlapping regions, which is valuable when selecting features or preparing data for classification models.

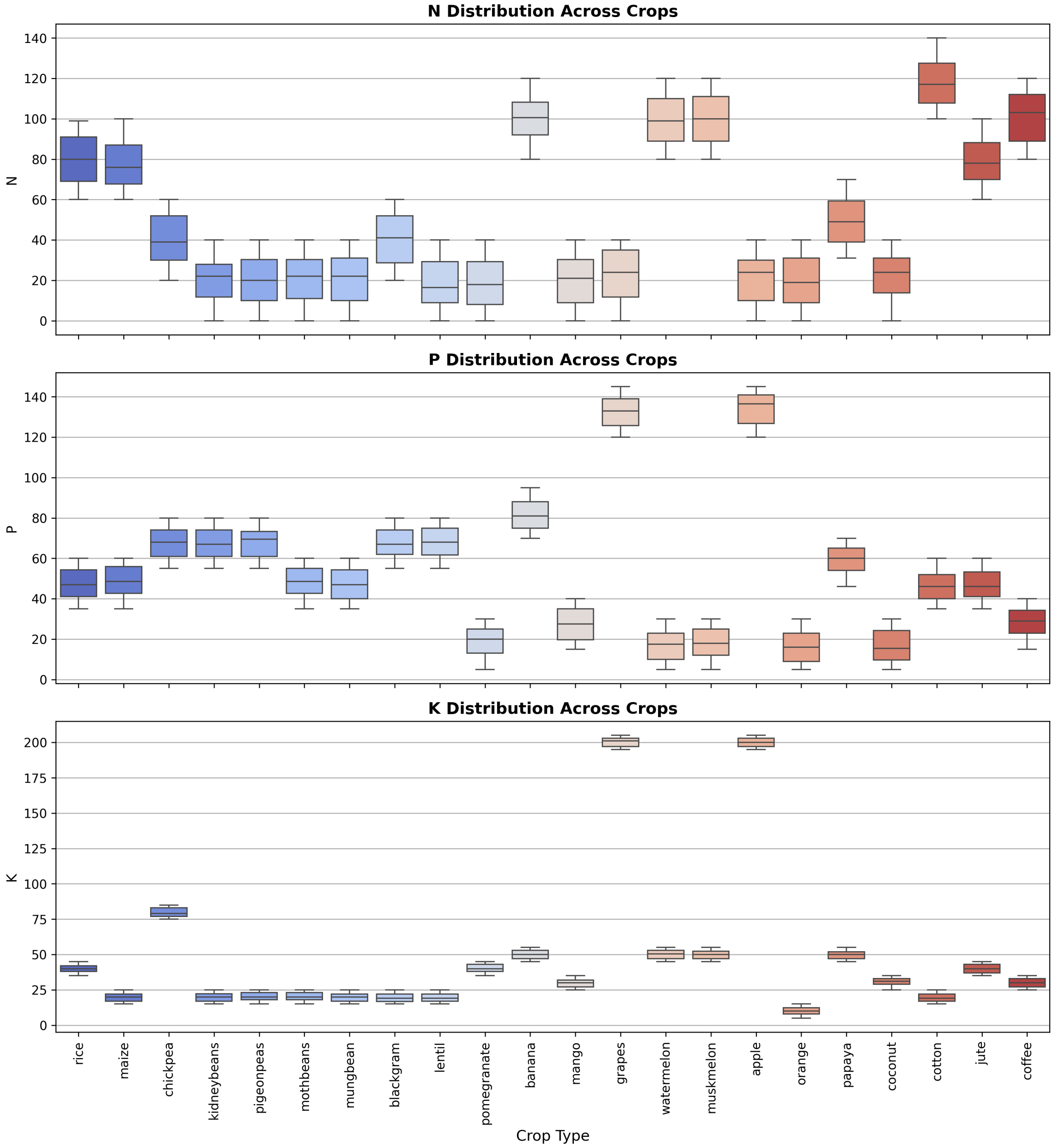


Figure 4.4: Feature distributions across crops for Nitrogen (N), Phosphorus (P), and Potassium (K).

4.3.4 EDA-Driven Strategy

Table 4.4 summarizes the key observations from the exploratory data analysis and the actions taken during dataset preparation. Each action is supported by statistical tests and visual checks to improve data quality and guide later modeling steps.

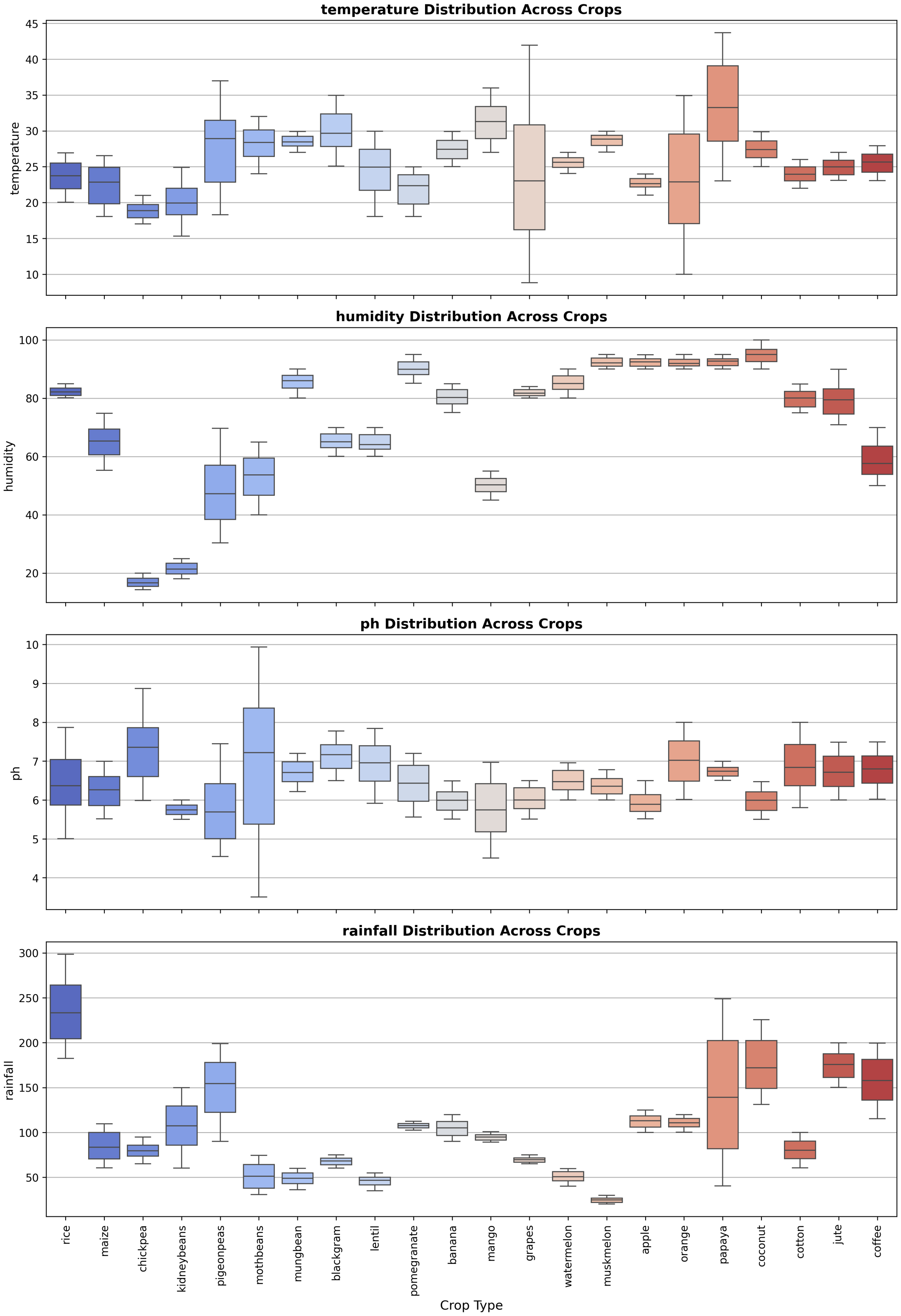


Figure 4.5: Feature distributions across crops for Temperature, Humidity, pH, and Rainfall.

Table 4.4: Summary of EDA Results and Applied Actions

Analysis	**Main Observations**	**Actions / Decisions**
Numerical–Numerical Correlation	Most feature pairs show very low correlation. Only Phosphorus and Potassium have a moderate positive relationship ($r \approx 0.74$).	Keep all features but monitor Phosphorus and Potassium when using linear models. If multicollinearity affects model stability, remove or combine one of them.
One-Way ANOVA	All numerical features differ strongly in their mean values across the 22 crop types (p-values close to zero). Potassium, Humidity, and Phosphorus show the highest F-statistics.	Retain all features for modeling. Give priority to Potassium and Humidity as main predictors. If dimensionality reduction is needed, start by dropping features with lower F-statistics.
Mutual Information (MI)	Humidity, Rainfall, and Potassium show the strongest non-linear relationship with crop labels. Nitrogen has moderate scores, and pH records the lowest value.	Use Humidity, Rainfall, and Potassium as key inputs for non-linear models. Consider removing pH if feature reduction is required since it carries limited information.
Feature Distributions	Rainfall, Humidity, and pH vary clearly across crops. Some crops share overlapping ranges and a few outliers appear in nutrient levels.	Apply scaling to maintain balanced influence of all variables. Identify and treat extreme nutrient values to reduce the effect of outliers.
Class Balance	Each crop class contains 100 samples.	No action required; keep the natural class balance without oversampling or undersampling.

4.4 Data Preprocessing

This section explains the operations performed to prepare the dataset before applying machine learning models. The objective was to ensure that all features were clean, complete, and ready for analysis.

4.4.1 Data Cleaning

Missing Value Detection

The dataset was examined to confirm that each numerical feature and the crop label were fully recorded. Let x_{ij} represent the value of feature j for observation i, and let the indicator function be

$$M_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is missing} \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

The number of missing entries for each feature is then calculated as $\sum_{i=1}^{n} M_{ij}$, where n is the total number of records.

Result: All seven numerical predictors (*Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH, Rainfall*) and the crop label were complete, with no missing entries detected. Because the dataset is fully populated, no imputation or record removal was required.

The same inspection was applied to check for duplicate rows and inconsistent values. No duplicates or irregular entries were found, confirming that the raw data could be used directly in later preprocessing steps.

Outlier Detection

Potential outliers were examined using the **Z-score method**, a standard statistical approach that measures how far a value lies from the mean of a given feature. For each observation x_i of a feature x, the Z-score is computed as

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{4.2}$$

where μ is the mean and σ is the standard deviation of the feature. Values with an absolute Z-score greater than three ($|Z_i| > 3$) were flagged as possible outliers. This threshold represents the outer 0.3% of values in a normal distribution, based on the empirical 68–95–99.7 rule. This procedure allows the detection of unusually high or low measurements that may result from recording errors or extreme natural variation.

Global Outlier Detection and Observations The univariate analysis in Section 4.3.2 showed that several features depart from a normal distribution, as reflected by their skewness and kurtosis. These measures correspond to the number of extreme values identified by the Z-score method:

- **Potassium (K)** displayed the highest skewness (2.40) and kurtosis (4.40). This pattern matches its large count of Z-score outliers and points to the presence of many very high measurements.
- **Phosphorus (P)** and **Rainfall** presented moderate positive skew (around 1.0), which agrees with their moderate outlier counts driven by occasional high readings.
- **pH** showed low skewness and a nearly normal distribution, in line with the small number of detected outliers.

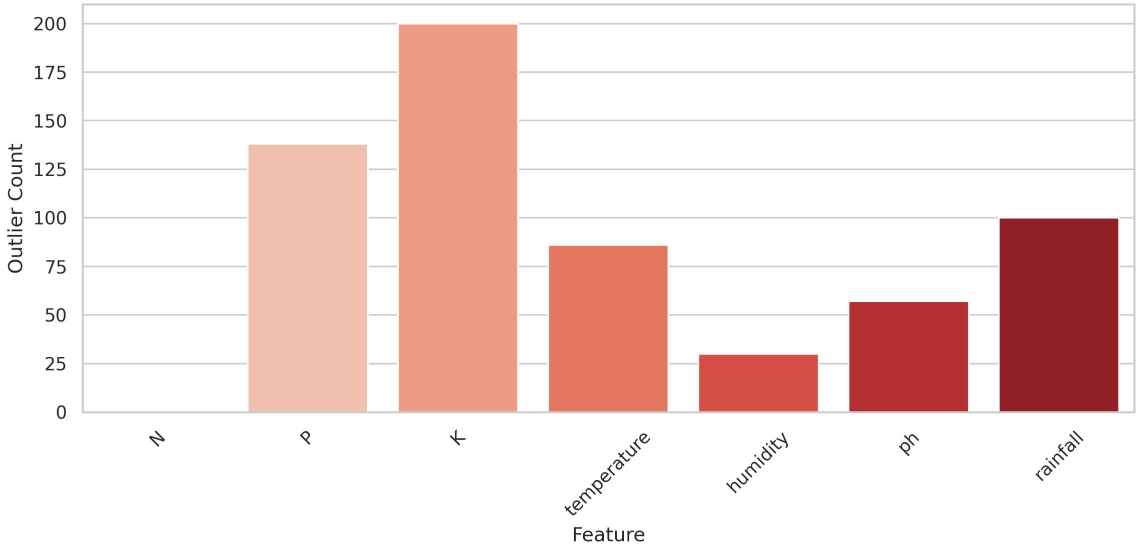


Figure 4.6: Global outlier count for each feature using Z-score.

The global Z-score analysis (Figure 4.6) showed that **Potassium (K)** contained about 200 extreme values, **Phosphorus (P)** around 135, **Temperature** about 85, **pH** nearly 60, **Rainfall** close to 100, and **Humidity** roughly 30. In contrast, **Nitrogen (N)** displayed no outliers under the standard Z-score threshold ($|z| > 3$), which indicates a stable distribution for this feature.

Limits of Global Z-Scores Global Z-score analysis can detect unusual values when the data come from a single population, but this condition does not fully match the present dataset. The data include 22 different crop types, each with its own physiological traits and environmental needs. As a result, values that appear extreme when viewed across the entire dataset may be normal within certain crop groups. This limits the use of a single global threshold for detecting outliers in such a varied, class-dependent setting. Figures 4.4 and 4.5 illustrate this point by showing the distribution of key numerical features for each crop type. The boxplots display the interquartile range (IQR) through their whiskers, providing a clearer view of variability within individual crop classes.

To address this variability, outlier detection was refined using a **class-conditional Z-score**, where the score is computed separately for each crop type. For a given crop c and feature j, the Z-score is calculated as

$$Z_{ij}^{(c)} = \frac{x_{ij}^{(c)} - \mu_j^{(c)}}{\sigma_j^{(c)}} \quad \text{for } i = 1, \ldots, n_c \tag{4.3}$$

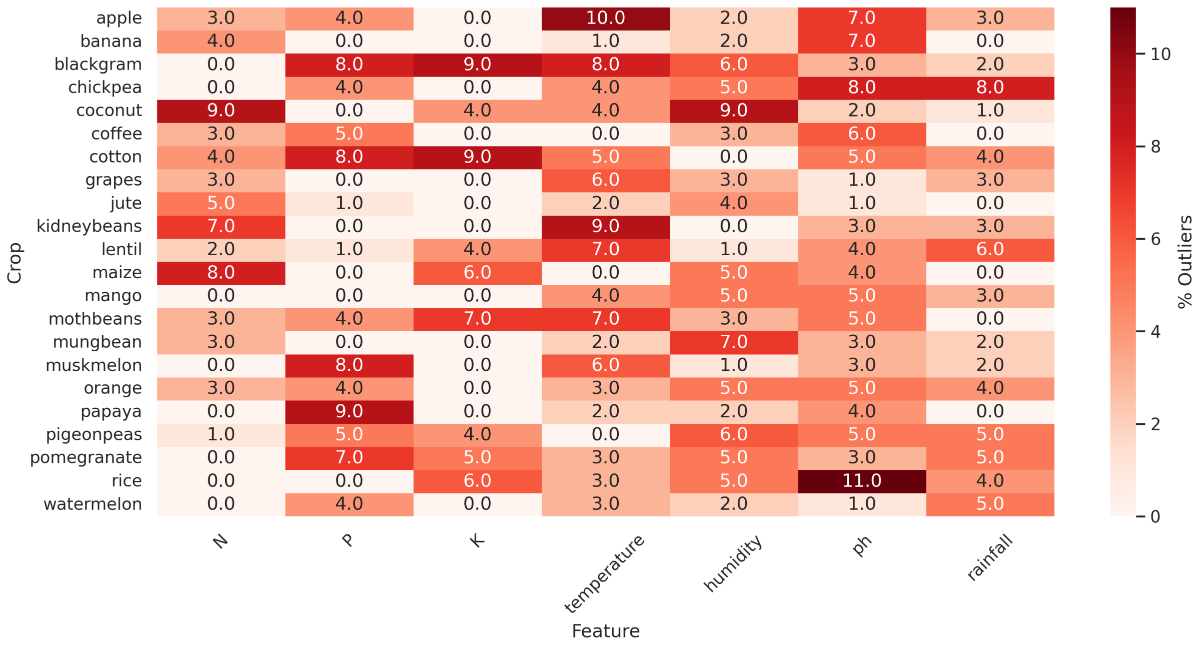
Here, $\mu_j^{(c)}$ and $\sigma_j^{(c)}$ represent the mean and standard deviation of feature j within crop c, and n_c is the number of observations for that crop. This method identifies extreme values relative to the natural distribution of each crop and avoids labeling valid crop-specific measurements as outliers.

The percentage of detected outliers for each crop and feature is shown in Figure 4.7. This figure provides a clear view of how the share of extreme values changes across crops and measured features.

Detected outliers were handled using **median imputation**. Each extreme value was replaced with the median of the same feature within the corresponding crop, which keeps the typical value of each group while reducing the effect of rare extreme points.

4.4.2 Feature Scaling

To give all numerical features an equal effect during model training, Min–Max scaling was applied. This method rescales each feature to the range $[0, 1]$ using its observed minimum and maximum values. Scaling prevents variables with wide numeric ranges



Crop	N	P	K	temperature	humidity	ph	rainfall
apple	3.0	4.0	0.0	10.0	2.0	7.0	3.0
banana	4.0	0.0	0.0	1.0	2.0	7.0	0.0
blackgram	0.0	8.0	9.0	8.0	6.0	3.0	2.0
chickpea	0.0	4.0	0.0	4.0	5.0	8.0	8.0
coconut	9.0	0.0	4.0	4.0	9.0	2.0	1.0
coffee	3.0	5.0	0.0	0.0	3.0	6.0	0.0
cotton	4.0	8.0	9.0	5.0	0.0	5.0	4.0
grapes	3.0	0.0	0.0	6.0	3.0	1.0	3.0
jute	5.0	1.0	0.0	2.0	4.0	1.0	0.0
kidneybeans	7.0	0.0	0.0	9.0	0.0	3.0	3.0
lentil	2.0	1.0	4.0	7.0	1.0	4.0	6.0
maize	8.0	0.0	6.0	0.0	5.0	4.0	0.0
mango	0.0	0.0	0.0	4.0	5.0	5.0	3.0
mothbeans	3.0	4.0	7.0	7.0	3.0	5.0	0.0
mungbean	3.0	0.0	0.0	2.0	7.0	3.0	2.0
muskmelon	0.0	8.0	0.0	6.0	1.0	3.0	2.0
orange	3.0	4.0	0.0	3.0	5.0	5.0	4.0
papaya	0.0	9.0	0.0	2.0	2.0	4.0	0.0
pigeonpeas	1.0	5.0	4.0	0.0	6.0	5.0	5.0
pomegranate	0.0	7.0	5.0	3.0	5.0	3.0	5.0
rice	0.0	0.0	6.0	3.0	5.0	11.0	4.0
watermelon	0.0	4.0	0.0	3.0	2.0	1.0	5.0

Figure 4.7: Percentage of detected outliers for each crop and feature using class-conditional Z-scores.

from dominating those with smaller ranges, improves numerical stability, and helps the optimization process converge more smoothly. The transformation is defined as

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{4.4}$$

where X is the original feature value, and X_{min} and X_{max} are the minimum and maximum values of that feature.

4.4.3 Categorical Encoding

The dataset includes categorical variable, the crop type labels, which need to be expressed as numbers to be used in machine learning models. Label encoding was applied to give each crop type a unique integer value. This approach keeps the class distinctions clear and allows categorical and numerical features to be combined during model training and evaluation.

4.4.4 Feature Selection

Feature selection was guided by EDA using correlation checks, ANOVA tests, mutual information, and examination of feature distributions.

Correlation results revealed one moderate relationship between Phosphorus and Potassium ($r \approx 0.74$), while all other feature pairs were nearly independent. To reduce the risk of multicollinearity in linear models, Potassium was kept and Phosphorus considered for removal.

ANOVA showed that all numerical features differ across crop types, with Potassium, Humidity, and Phosphorus giving the highest F-statistics. Mutual information supported this by ranking Humidity, Rainfall, and Potassium as the most informative, while pH had the lowest score and showed strong overlap between crop classes.

Based on these results, four features were selected for model training:

- **Potassium**: high ANOVA score and strong mutual information,
- **Humidity**: high ANOVA score and highest mutual information,
- **Rainfall**: strong mutual information and clear separation between classes,
- **Temperature**: moderate but stable predictive value.

The remaining features **pH**, **Phosphorus**, and **Nitrogen** were removed. pH contributed little useful information and overlapped heavily across classes. Phosphorus was dropped to avoid redundancy with Potassium, which showed stronger predictive strength. Nitrogen, although important for soil analysis, provided only moderate mutual information and lower discriminative value than Temperature.

To confirm whether the four selected features can achieve performance similar to the full seven-feature set, the next section compares machine learning models trained on both configurations.

4.4.5 Data Augmentation

The dataset was augmented to raise the number of samples in each crop class from 100 to 300, giving a total of 6,600 records. This step widened the range of feature values and provided a stronger basis for model training. The target of 300 samples per class was chosen as a compromise between diversity and computational cost. Preliminary tests showed that 200 samples offered limited variability, while 400 samples added little benefit but increased training time.

The augmentation process is given by

$$R_{\text{augmented}} = N_{\text{original}} + (R_{\text{target}} - R_{\text{original}}) \times C \tag{4.5}$$

where

- $R_{\text{augmented}}$ is the total number of rows after augmentation $(6,600)$,
- N_{original} is the initial number of rows $(2,200)$,
- R_{target} is the desired rows per class (300),
- R_{original} is the rows per class before augmentation (100),
- C is the number of crop classes (22).

This expansion allowed evaluation of model performance on a larger training set while keeping computation manageable. Because the extra records are synthetic, model results were validated on the augmented data to ensure reliability.

4.5 Assessment of Data Preprocessing

This section reviews the preprocessing steps applied to the dataset and examines their effect on model reliability. A set of baseline machine learning models is evaluated to provide a performance benchmark and to verify the effectiveness of the preprocessing procedure.

4.5.1 Experimental Design

The effect of the preprocessing pipeline described in Section 4.4 was examined through a set of controlled experiments. Each experiment modified a single preprocessing step while keeping the remaining steps unchanged, allowing a clear view of how individual choices influence model behavior.

The prepared dataset was then used to train and test several machine learning algorithms for crop recommendation. The evaluated models include Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB). Their performance serves as a reference for validating the preprocessing approach.

The experiments addressed the following aspects:

- **Dataset Splitting:** Tested 50:50 and 80:20 partitions to study how training set size influences performance.
- **Data Cleaning:** Compared median imputation of class-conditional outliers with direct removal to examine the impact on model stability and data coverage.
- **Feature Scaling:** Tested Min-Max scaling to assess its effect on model performance.
- **Feature Selection:** Trained models first on the key variables identified during exploratory analysis and then on the complete feature set to measure the contribution of additional attributes.
- **Data augmentation:** Evaluated the effect of increasing the number of samples per crop on prediction accuracy.

4.5.2 Results

Before presenting the detailed comparisons, it is useful to describe the starting conditions that served as a reference for all experiments. The first set of tests was carried out using the following baseline configuration:

- **Data cleaning:** The raw dataset was used as collected, with no cleaning or imputation of missing or extreme values.
- **Dataset split:** Training and testing sets were divided evenly using a 50:50 ratio.
- **Data augmentation:** No additional samples were generated or added to the dataset.

- **Feature selection:** Only four key features were selected from the seven available variables.

This setup provides a practical reference point for examining how later preprocessing choices influence model accuracy and stability.

Dataset Splitting

Table 4.8 shows the effect of changing the train–test split from 50:50 to 80:20. All classifiers obtained higher accuracy when a larger share of the data was used for training, which allowed the models to learn more representative patterns before evaluation. The largest improvement was observed for KNN, which increased from 82.45% to 87.72% (a gain of 5.27%). Decision Tree and Random Forest also benefited, with increases of 2.45% and 1.59%, respectively. SVM improved by 1.91%, while Naïve Bayes showed only a slight change of 0.04%.

Table 4.5: Impact of dataset splitting on classification accuracy.

Classifier	50:50 Split	80:20 Split	Accuracy Gain
RF	94.54%	96.13%	+1.59%
DT	92.27%	94.72%	+2.45%
NB	94.27%	94.31%	+0.04%
SVM	68.09%	70.00%	+1.91%
KNN	82.45%	87.72%	+5.27%

Data Cleaning

To examine how different treatments of outliers influence model accuracy, three strategies were applied:

1. **Raw data**: The dataset was used without any cleaning to reflect original field conditions.

2. **Outlier removal**: Records with extreme values were deleted in an attempt to reduce skewed distributions.

3. **Median imputation**: Outliers were replaced with the feature-wise median to keep the full sample size while limiting the effect of extremes.

Table 4.6: Impact of Outlier Processing Techniques on Accuracy with Gains or Losses Relative to Raw Data.

Classifier	Raw Data	Outlier Removal (Gain/Loss)	Median Imputation (Gain/Loss)
RF	94.54%	93.00% (-1.54%)	94.72% (+0.18%)
DT	92.27%	89.54% (-2.73%)	92.27% (+0.00%)
NB	94.27%	91.72% (-2.55%)	94.27% (+0.00%)
SVM	68.09%	61.72% (-6.37%)	68.27% (+0.18%)
KNN	82.45%	77.45% (-5.00%)	82.72% (+0.27%)

A direct comparison between the raw dataset and the version with outlier removal shows a drop in accuracy for every classifier (Table 4.6). The reduction is small for RF (-1.54%) but large for SVM (-6.37%), which suggests that deleting extreme records removed some data points that carry useful information. When outliers were replaced by the median, the models kept nearly the same accuracy as the raw data and in a few cases achieved slight gains. For example, RF rose from 94.54% to 94.72% (+0.18%), and KNN improved from 82.45% to 82.72% (+0.27%).

Feature Scaling

To reduce bias caused by different feature ranges, we tested Min–Max scaling in the range [0,1] against models trained on raw values:

- **Unscaled**: Original feature ranges kept without adjustment.
- **Scaled**: All features transformed to the [0,1] range using the Min–Max method.

Table 4.7: Impact of Feature Scaling on Accuracy

Classifier	Without Scaling	With Scaling	Accuracy Gain
RF	94.54%	95.18%	+0.64%
DT	92.27%	92.81%	+0.54%
NB	94.27%	94.27%	+0.00%
SVM	68.09%	88.63%	+20.54%
KNN	82.45%	88.36%	+5.91%

Normalization mainly improved models that depend on distance calculations (Table 4.7). SVM gained 20.54% because kernel functions are sensitive to unequal feature ranges. KNN increased by 5.91% as neighbor comparisons require features on a similar scale.

Tree-based methods showed only minor changes (RF: +0.64%, DT: +0.54%), and Naïve Bayes remained unchanged, reflecting their scale-invariant structure.

Feature Selection

We compared two different feature sets using an 80:20 train–test split to examine the effect of dimensionality on model accuracy:

- **Selected features**: a reduced set containing Potassium, Temperature, Humidity, and Rainfall.
- **All features**: the complete group of seven variables.

Table 4.8: Impact of Feature Count on Accuracy

Classifiers	4 features (80:20 split)	7 features (80:20 split)	Accuracy Gain
RF	96.13%	99.09%	+2.96%
DT	94.72%	98.18%	+3.46%
NB	94.31%	99.31%	+5.00%
SVM	70.00%	96.59%	+26.59%
KNN	87.72%	97.50%	+9.78%

Expanding the input to seven features improved accuracy for all models (Table 4.8). Gains ranged from +2.96% for RF to +26.59% for SVM. Tree-based models such as RF and DT showed smaller gains, while SVM and KNN benefited most from the extra variables. This shows that the four-feature set retains most of the useful signal, but using all seven features gives the best overall accuracy.

Data Augmentation

Two dataset versions were evaluated using an 80:20 train–test split to assess how data augmentation affects model accuracy:

- **Without augmentation**: The original dataset contained 2,200 records, with 100 samples for each of the 22 crop classes.
- **With augmentation**: The dataset was expanded to 6,600 records by generating synthetic samples so that each class increased from 100 to 300 samples.

Table 4.9: Impact of Data Augmentation on Accuracy

Classifier	Without Augmentation (80:20 split)	With Augmentation (80:20 split)	Accuracy Gain
RF	96.13%	96.85%	+0.72%
DT	94.72%	95.41%	+0.69%
NB	94.31%	94.88%	+0.57%
SVM	70.00%	71.26%	+1.26%
KNN	87.72%	88.65%	+0.93%

As shown in Table 4.9, adding synthetic samples gave a modest accuracy increase for every model. Tree-based classifiers such as RF and DT gained about 0.7%, which reflects their ability to use a larger training set even when their performance was already high. Naïve Bayes improved by only 0.57%, a small change that matches its lower dependence on sample size once class probabilities are well estimated. Distance-based methods benefited the most: KNN accuracy rose by 0.93%, and SVM achieved the largest gain of 1.26%, suggesting that extra data helped these algorithms draw more precise decision boundaries in the feature space.

Preprocessing Outcomes

Testing showed that an 80:20 split, median imputation, Min–Max scaling, full seven-feature input, and data augmentation each improved model accuracy. Scaling mainly boosted SVM and KNN, while tree models gained from median imputation and larger training data. These choices were combined to build the final preprocessing pipeline for the AdaBoost crop selection model.

4.6 Development and Workflow of the CS-AdaRF-SHAP Model

We introduce **CS-AdaRF-SHAP**, an AdaBoost ensemble with Random Forest base learners and post hoc explanations provided by SHAP.

4.6.1 CS-AdaRF Development and Optimization

Rationale for Selecting AdaBoost

In agricultural decision-making, particularly in crop selection, **classification errors can have serious effects**. A **false positive (FP)**, which occurs when the system recommends an *unsuitable crop*, may lead to wasted inputs, poor yields, and financial loss. In contrast, a **false negative (FN)**, where a *suitable crop* is not recommended, can cause farmers to miss profitable and productive options.

AdaBoost is well suited because it trains models in sequence and gives more weight to difficult samples, helping to lower both types of errors. Its exponential loss function places strong penalties on confident mistakes, adding an extra layer of protection against costly outcomes.

When combined with strong base learners such as Random Forests, AdaBoost provides a solid compromise between predictive power and resistance to overfitting. Because minimizing wrong recommendations and maintaining farmer confidence are critical for a practical crop selection system, AdaBoost was chosen as the main model even when other algorithms achieved similar accuracy in early experiments.

CS-AdaRF Framework

AdaBoost (Adaptive Boosting) is an ensemble learning method that builds a strong predictive model by combining many weak classifiers in a sequential manner. As shown in Figure 4.8, the algorithm begins by assigning the same weight to every training sample. During each iteration, a weak base learner is trained on the weighted data. After each round, the weights of misclassified samples are increased so that the next learner pays more attention to the cases that are hardest to classify. Through this adaptive weighting, the model gradually focuses on the most challenging observations and improves its ability to separate the classes. In the final step, AdaBoost merges the outputs of all weak learners, usually through a weighted voting scheme, to produce a single classifier that achieves higher accuracy than any individual learner.

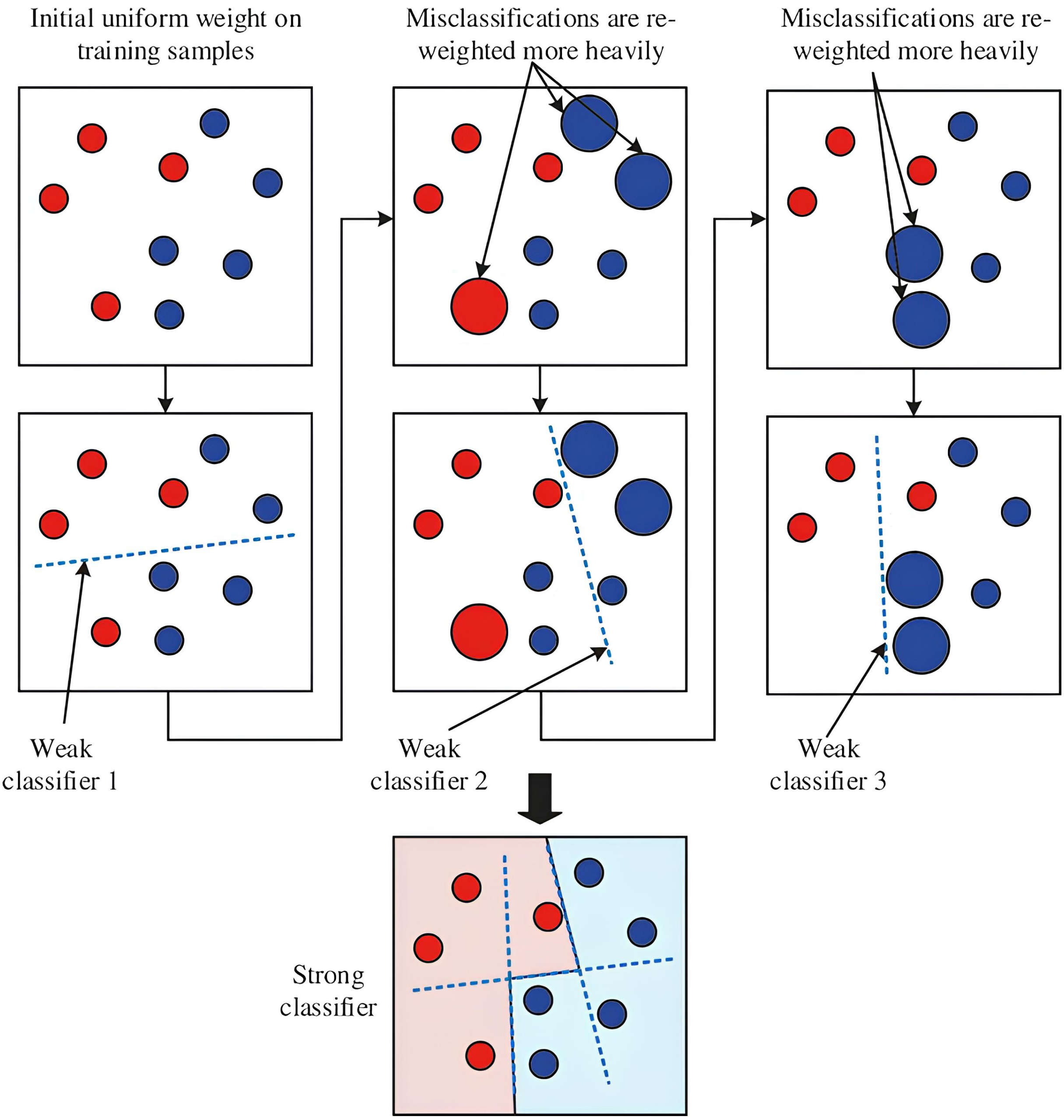

Initial uniform weight on training samples
Misclassifications are re-weighted more heavily
Misclassifications are re-weighted more heavily
Weak classifier 1
Weak classifier 2
Weak classifier 3
Strong classifier

Figure 4.8: AdaBoost architecture.

Mathematical Foundations and Learning Procedure

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ denotes the feature vector and $y_i \in \{1, 2, \ldots, 22\}$ the crop class label, CS-AdaRF operates as follows:

1. **Initialization:** Assign each sample an equal initial weight:

$$w_i^{(1)} = \frac{1}{N}, \quad \forall i = 1, \ldots, N.$$

This ensures every data point is equally important at the start.

2. **Boosting Rounds:** For each iteration $t = 1, \ldots, T$:

   (a) **Base Classifier Training:** Train a Random Forest classifier $h_t(x)$ using the weighted dataset. Samples that were harder to classify in previous rounds will have higher weights.

   (b) **Weighted Error Calculation:** Compute the weighted error rate:

   $$\epsilon_t = \frac{\sum_{i=1}^{N} w_i^{(t)} \cdot \mathbb{I}(h_t(x_i) \neq y_i)}{\sum_{i=1}^{N} w_i^{(t)}}$$

   where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the prediction is incorrect.

   (c) **Model Weight Computation:** Calculate the importance (weight) of the current model:

   $$\alpha_t = \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) + \ln(K - 1)$$

   where $K = 22$ is the number of classes. This step ensures that models with lower error rates contribute more to the final prediction.

   (d) **Weight Update:** Increase the weights of misclassified samples so that future classifiers focus on them:

   $$w_i^{(t+1)} = w_i^{(t)} \cdot \exp\left(\alpha_t \cdot \mathbb{I}(h_t(x_i) \neq y_i)\right)$$

   (e) **Normalization:** Normalize weights so they sum to 1:

   $$\sum_{i=1}^{N} w_i^{(t+1)} = 1$$

   This maintains the weights as probabilities for the next round.

3. **Final Ensemble Prediction:** For a new sample x, the ensemble predicts the class

with the highest weighted sum of votes:

$$H(x) = \arg \max_{k \in \{1,\ldots,22\}} \sum_{t=1}^{T} \alpha_t \cdot \mathbb{I}(h_t(x) = k)$$

Each $h_t(x)$ is a Random Forest model trained in the t-th boosting round.

Algorithm 1 CS-AdaRF-SHAP for Crop Selection

```
**Input:** Preprocessed dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; number of boosting rounds $T$; Random Forest hyperparameters
**Output:** Final ensemble classifier $H(x)$
Initialize sample weights: $w_i^{(1)} \leftarrow 1/N$ for all $i$
**for** $t = 1$ to $T$ **do**
    Train Random Forest $h_t(x)$ with weights $w_i^{(t)}$
    Compute weighted error: $\epsilon_t \leftarrow \frac{\sum_{i=1}^{N} w_i^{(t)} \mathbb{I}(h_t(x_i) \neq y_i)}{\sum_{i=1}^{N} w_i^{(t)}}$
    Compute model weight: $\alpha_t \leftarrow \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) + \ln(K-1)$
    **for** $i = 1$ to $N$ **do**
        Update: $w_i^{(t+1)} \leftarrow w_i^{(t)} \exp\left(\alpha_t \mathbb{I}(h_t(x_i) \neq y_i)\right)$
    **end for**
    Normalize $w^{(t+1)}$ so $\sum_{i=1}^{N} w_i^{(t+1)} = 1$
**end for**
**Prediction:** For test sample $x$,
    $H(x) = \arg \max_{k \in \{1,\ldots,22\}} \sum_{t=1}^{T} \alpha_t \mathbb{I}(h_t(x) = k)$
Compute SHAP values for $H(x)$ to explain predictions.
```

Hyperparameter Selection and Optimization

To obtain strong predictive performance, the training setup and model hyperparameters were adjusted through several trial runs until a stable and accurate configuration was reached. The main parameter settings are listed below:

- **Number of estimators:** $n\_estimators = 50$

  Defines how many weak learners (Random Forest classifiers) are combined in the ensemble to build a reliable prediction model.

- **Base estimator:** `RandomForestClassifier`

Serves as the weak learner within AdaBoost, providing the ability to capture complex feature patterns and reduce variance.

- **Learning rate:** 0.001

  Regulates the weight given to each weak learner when forming the final ensemble, allowing careful control of the learning process and helping the model generalize.

- **Random state:** 0

  Sets the random seed for all stochastic operations so that experiments can be repeated and results can be reproduced.

4.6.2 SHAP-Based Interpretability

In our system, CS-AdaRF builds an ensemble of Random Forest classifiers, where each successive model pays more attention to the errors made by the previous ones. This iterative process produces a model with strong predictive accuracy but leaves the decision process as a black box. To make the decision process understandable, SHAP (SHapley Additive exPlanations) is applied *post hoc* after the CS-AdaRF model has been fully trained. SHAP is used only during model evaluation on the test data and does not affect the training procedure or the optimization of the model.

Theoretical Basis of SHAP: SHAP builds on Shapley values from cooperative game theory, which assign to each feature its fair share of the average contribution to a model's prediction. In the CS-AdaRF-SHAP system, SHAP evaluates the adaptive, weighted outputs of all base classifiers and produces explanations that satisfy the key properties of local accuracy, additivity, and missingness. These properties are important for producing clear and reliable interpretations in agricultural decision support.

For a problem with $K = 22$ crop classes and $d = 7$ input features (soil nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall), the SHAP value for feature x_i and class k is given by:

$$\mathrm{SHAP}(f, x_i, k) = \sum_{S \subseteq F \setminus \{x_i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \big[f_k(x_S \cup \{x_i\}) - f_k(x_S)\big] \tag{4.6}$$

Where:

- f: the trained CS-AdaRF model,
- x_i: the i-th input feature,
- k: the predicted crop class ($k = 1, \ldots, 22$),
- F: the complete set of input features,
- S: a subset of F that does not contain x_i,
- $f_k(x_S)$: the model's predicted probability (or score) for class k using only the features in S,
- $|S|!$, $(|F| - |S| - 1)!$, $|F|!$: factorial terms used to average fairly over all possible subsets of features.

Algorithmic Workflow:

Algorithm 2 Interpretable Crop Selection with SHAP

```
Require: Trained AdaBoost model $f$, evaluation dataset $X$, number of crop classes $K = 22$
for $k \leftarrow 1$ to $K$ do
    $explainer_k \leftarrow$ Initialize SHAP explainer for class $k$
    $shap\_values_k \leftarrow$ Compute SHAP values for all $x \in X$ and class $k$
end for
for $k \leftarrow 1$ to $K$ do
    $Feature\_importance_k \leftarrow$ Aggregate SHAP values for class $k$
end for
return $Feature\_importance_k$
```

Algorithm Description: After training the proposed **CS-AdaRF** model, SHAP values are computed for every feature and crop class using the evaluation data. The procedure includes:

- Creating a SHAP explainer for each crop class.
- Calculating the SHAP value of each feature for every sample to measure its effect on the model prediction.

- Aggregating the computed values over all samples to obtain the overall importance of each feature for each crop class.

These aggregated SHAP values reveal which soil and environmental variables play the largest role in each crop recommendation and provide a clear basis for practical agricultural decisions.

Reading the SHAP Values:

- **Positive values**: Indicate that the feature raises the probability of selecting a specific crop.
- **Negative values**: Indicate that the feature lowers the probability of selecting that crop.
- **Magnitude**: Shows how strongly the feature affects the model's recommendation, with larger values meaning a greater effect.

4.7 Results

4.7.1 Evaluation of the CS-AdaRF Model

This section presents a detailed assessment of the predictive ability of the proposed CS-AdaRF model, which combines AdaBoost with Random Forest base learners for multi-class crop selection. The evaluation covers accuracy, precision, recall, F1-score, training time, and class-wise results, and includes a direct comparison with other classification methods.

Training and Testing Behavior:

Figure 4.9 shows the progression of accuracy and error rate for the CS-AdaRF model during training and testing. The error steadily decreases from 0.06 to 0.003 on the training set and from 0.054 to 0.004 on the testing set, indicating efficient learning and strong generalization. At the same time, both training and testing accuracy increase from about 0.95 to nearly 0.999, demonstrating effective reduction of misclassification without signs of overfitting.

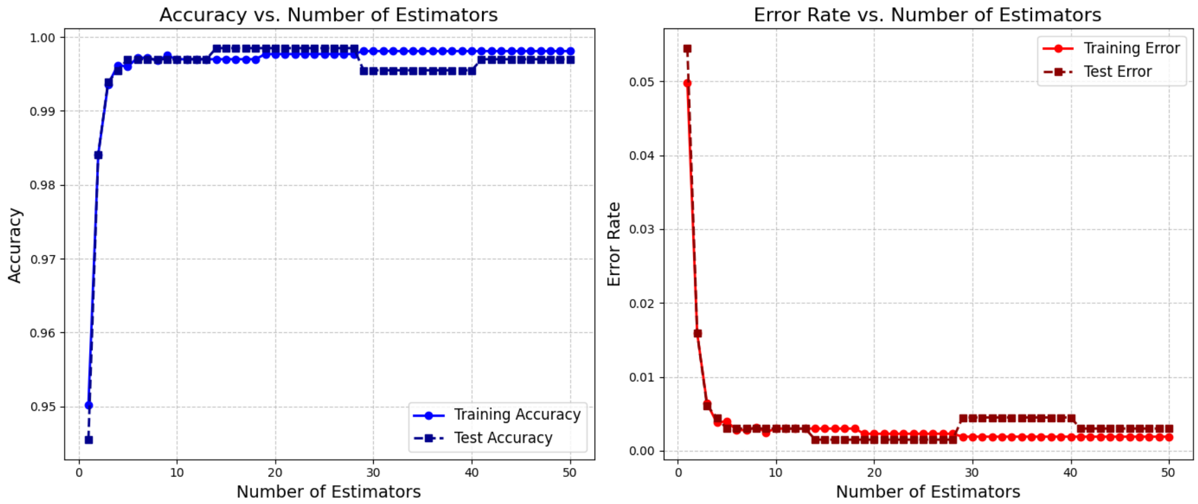


Figure 4.9: Accuracy and error rate of the CS-AdaRF Model during training and testing.

Comparison with Alternative Models

Table 4.10 and Figure 4.10 present the comparative evaluation of CS-AdaRF against several well-established classifiers, including SVM, KNN, Decision Tree (DT), Bagging, XGBoost, and LightGBM. CS-AdaRF achieves the highest overall accuracy (99.77%) and records perfect values for precision, recall, and F1-score. These results show that the model produces highly accurate predictions while maintaining balanced control over false positives and false negatives. In addition, the recorded testing time of 0.57 seconds demonstrates the suitability of CS-AdaRF for practical deployment.

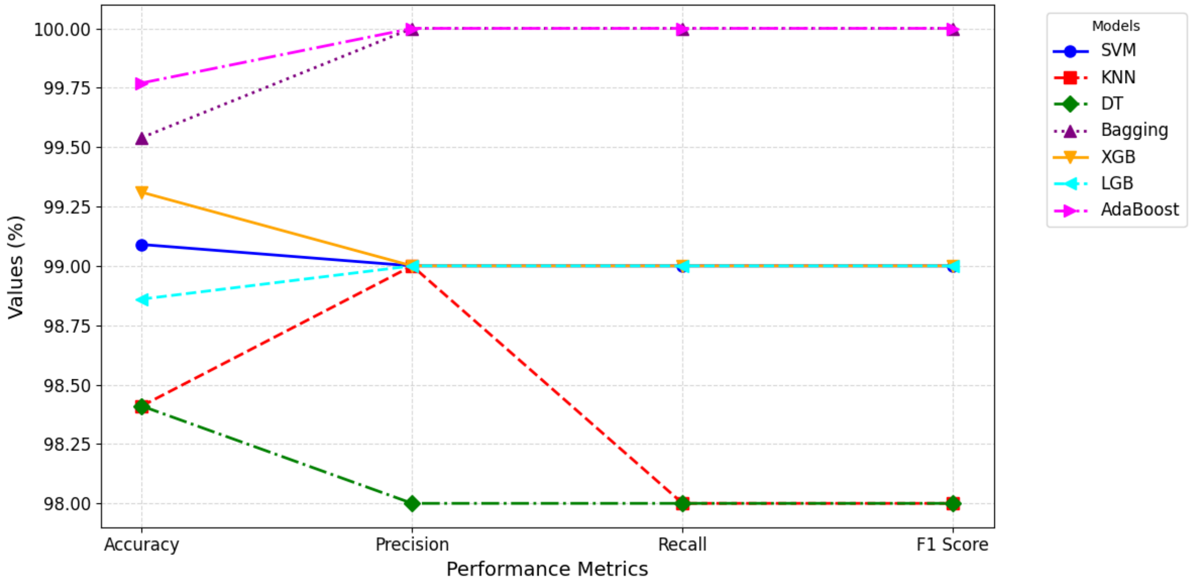


Figure 4.10: Performance metrics comparison across multiple models.

Table 4.10: Performance metrics comparison across multiple models.

Models	Correct Instances	Incorrect Instances	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Testing time (s)
SVM	1308	12	99.09	99	99	99	0.07
KNN	1299	21	98.41	99	98	98	0.003
DT	1299	21	98.41	98	98	98	0.037
Bagging	1314	6	99.54	100	100	100	9.7
XGBoost	1311	9	99.31	99	99	99	12.3
LightGBM	1305	15	98.86	99	99	99	4.5
CS-AdaRF	**1317**	**3**	**99.77**	**100**	**100**	**100**	**0.57**

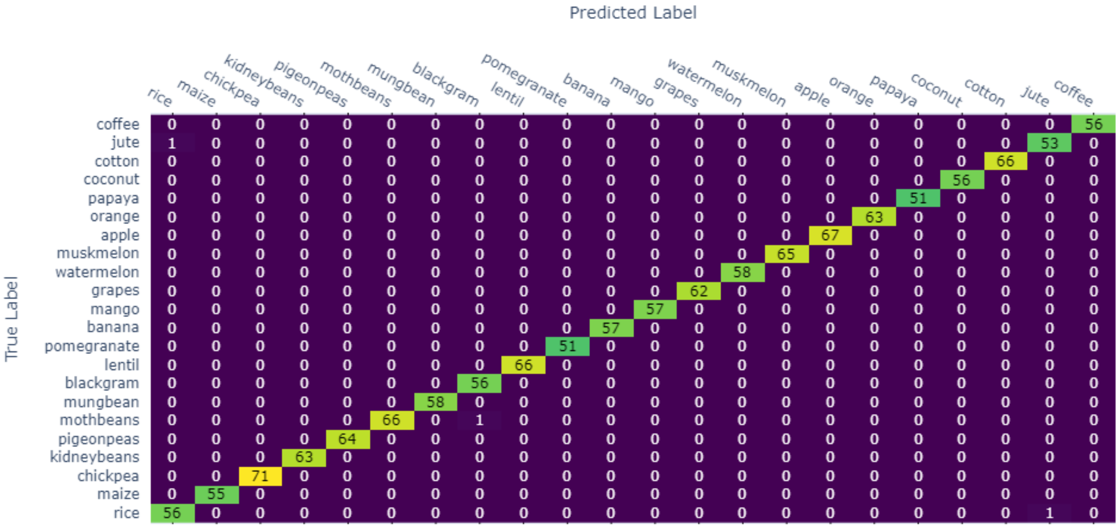


Figure 4.11: Confusion matrix for the CS-AdaRF model.

Analysis of Misclassifications

The confusion matrix in Figure 4.11 provides a detailed view of classification errors across the 22 crop categories. The proposed model misclassifies only three samples, a very small number considering the complexity of the task. These errors are not random but occur in crop pairs that share closely related agronomic characteristics:

- **Rice predicted as Jute:** Rice and jute are often cultivated in similar floodplain areas and require comparable soil nutrients and climate conditions. The single error most likely reflects a data point located near the decision boundary in the feature space, where high rainfall and overlapping nutrient profiles could describe either crop.

- **Blackgram predicted as Mothbeans:** Both crops belong to the legume family and thrive under similar soil and environmental conditions, especially in regions with moderate rainfall and similar nitrogen needs. Their feature representations in the dataset are so close that even a strong classifier may confuse a small number of cases.

Practical Impact of Misclassification and Computational Efficiency

In agricultural decision support, even rare classification errors can have practical consequences. For example, recommending jute instead of rice could lead to lower yield or inefficient use of inputs if the field is better suited to rice. Similarly, confusing blackgram with mothbeans may affect fertilizer selection, irrigation planning, and marketing decisions. However, the very small number of errors (3 out of 1320 samples) and the fact that these mix-ups occur between crops with similar biological and environmental requirements indicate that the model is highly reliable and presents minimal risk to farmers.

The proposed CS-AdaRF model achieves top predictive performance with near-perfect classification across all crop types and a testing time of only 0.57 seconds. This level of efficiency stands out when compared to other strong ensemble methods: Bagging (9.7 seconds), XGBoost (12.3 seconds), and LightGBM (4.5 seconds) require considerably longer training while delivering slightly lower accuracy and F1 scores.

CS-AdaRF reaches the highest accuracy (99.77%) and perfect precision, recall, and F1-score, while maintaining a training time far shorter than Bagging and XGBoost. Although simpler models such as SVM (0.07 s), KNN (0.003 s), and Decision Tree (0.037 s) train somewhat faster, they do so at the cost of reduced predictive power and a higher rate of misclassification.

This combination of strong predictive accuracy and low computational cost makes CS-AdaRF well suited for real-world crop selection tasks, particularly when rapid retraining and scalability are important. Its efficiency supports quick deployment and model updates while lowering computing requirements, which is valuable for agricultural decision support in environments with limited resources.

4.7.2 Assessment of Model Interpretability

Understanding which input features guide the model's decisions is essential for scientific validation and for building trust among agricultural practitioners. To explore this aspect, the permutation feature importance method was applied to the CS-AdaRF model in order to measure how each variable affects prediction accuracy. The ranking shown in Figure 4.12 reveals the features that contribute most to changes in prediction error.

Despite its usefulness, permutation importance has known limitations. When features are strongly correlated, the method can produce biased rankings. In this dataset, for example, phosphorus (P) and potassium (K) show a correlation coefficient of 0.74, which may cause their importance to be overestimated or underestimated. Such multicollinearity complicates the interpretation of their individual roles. Moreover, permutation importance does not indicate whether a feature promotes or suppresses a particular crop recommendation, nor does it express the strength or direction of its effect. These missing details are important for translating model outputs into practical agronomic advice.

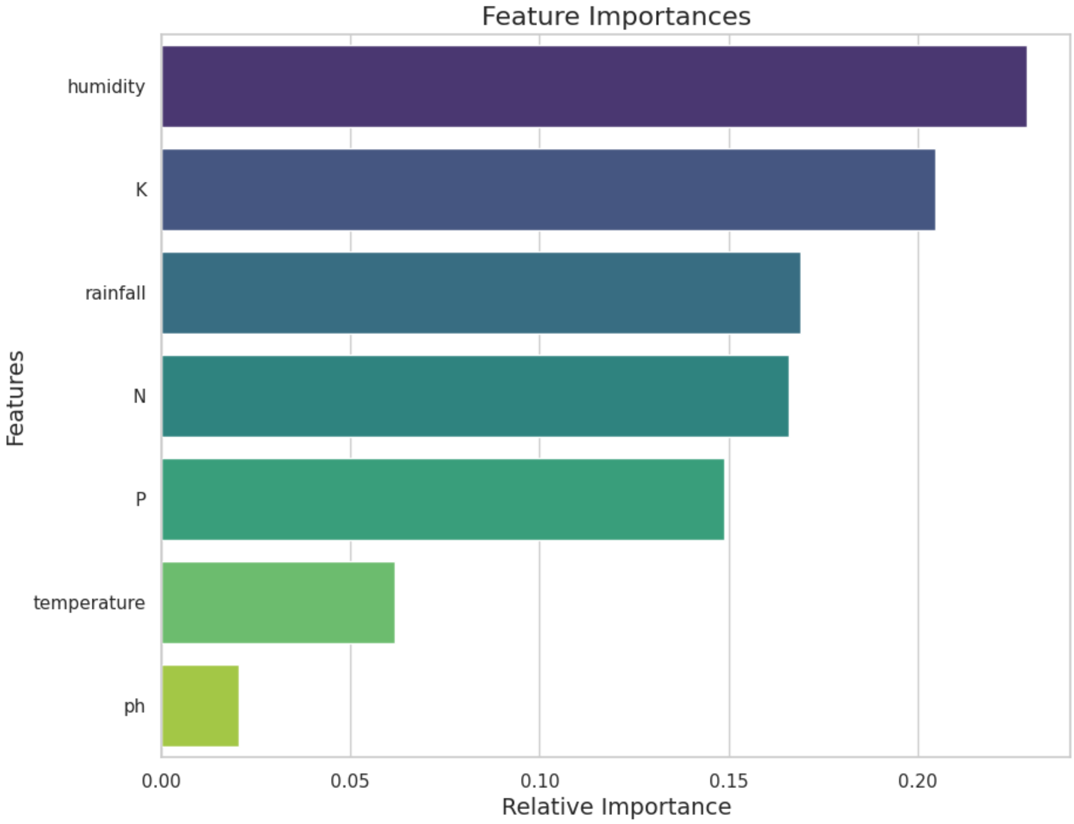


Figure 4.12: Permutation-based ranking of feature importance.

SHAP overcomes the limitations of permutation importance by using concepts from game theory to assign each feature a precise contribution to every individual prediction. In contrast to permutation importance, SHAP offers two key advantages:

- It can reliably separate the contribution of each feature even when strong correlations are present, producing stable and meaningful attributions.
- It provides both the direction of influence (whether a feature increases or decreases the probability of selecting a specific crop) and the magnitude of this effect.

The SHAP summary plot in Figure 4.13 presents the overall impact of all input variables on crop recommendations. Humidity appears as the most influential factor, followed by nitrogen (N) and potassium (K). These results are consistent with well-known agronomic relationships and also reveal data-driven details about how soil and climate conditions shape crop suitability.

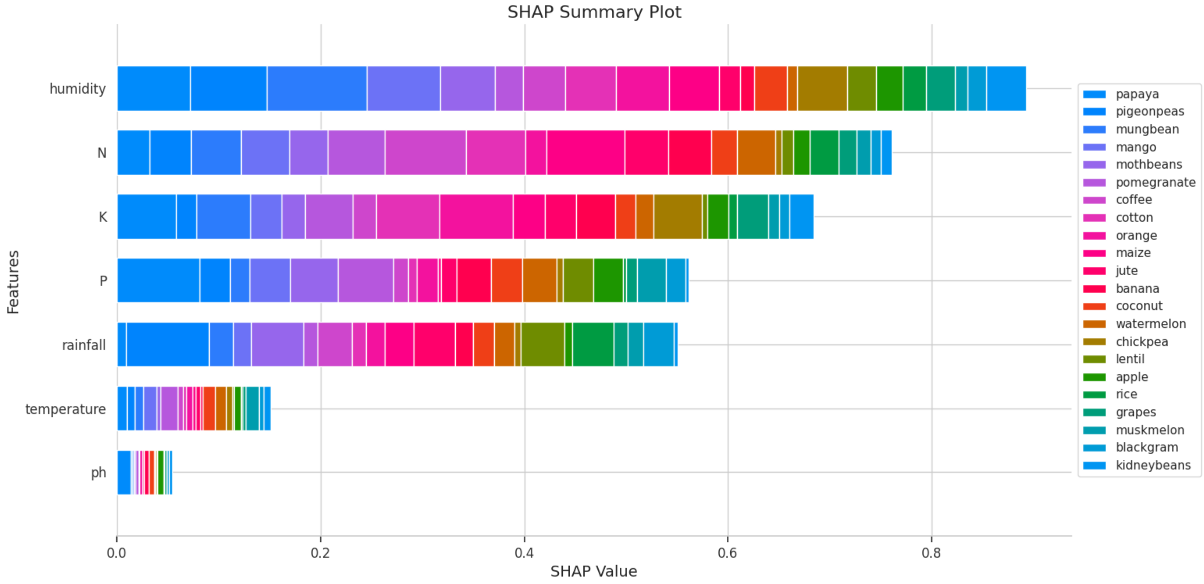


Figure 4.13: SHAP-based analysis of feature importance for crop recommendations.

Crop-Specific Explanations and Case Studies

A key advantage of SHAP is its ability to provide explanations for individual predictions and specific crop classes. To demonstrate this capability, SHAP values were examined for four representative crops, rice, maize, chickpea, and banana (Figures 4.14 and 4.15). The analysis shows the following patterns:

- **Rice:** Rainfall is the primary factor driving suitability, reflecting rice's high water requirement. Nitrogen and humidity also contribute, though to a lesser extent. The model correctly reduces the likelihood of rice selection in areas with low rainfall, matching known agronomic limits.

- **Maize:** Nitrogen availability is the strongest positive driver, consistent with maize's high demand for N. Humidity and potassium also support suitability, while excessive rainfall slightly lowers the recommendation because maize is prone to waterlogging.

- **Chickpea:** Potassium and moderate humidity play the most important roles. Very high humidity or low potassium reduce the predicted suitability, showing the model's ability to balance interacting environmental and nutrient factors.

- **Banana:** Nitrogen, potassium, and phosphorus are all essential. Rainfall has a moderate but complex effect, where both excess and shortage of key nutrients can reduce the predicted suitability.

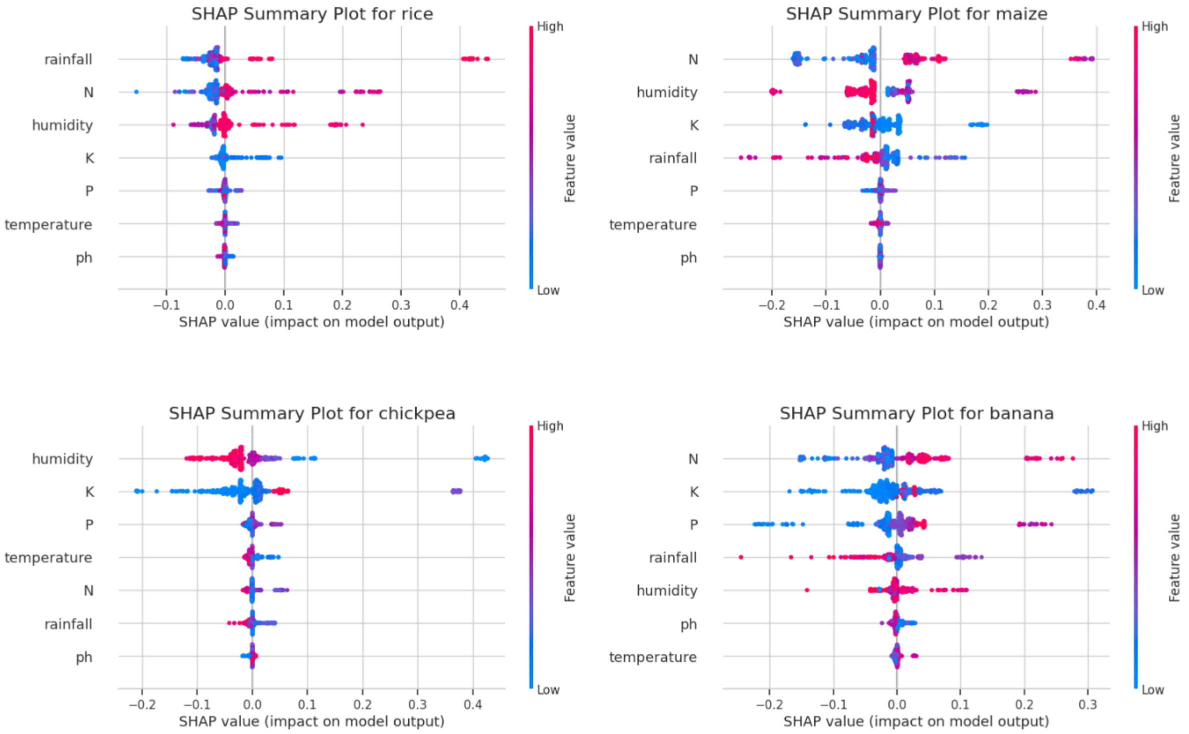


Figure 4.14: SHAP summary plot showing feature influence for rice, maize, chickpea, and banana.

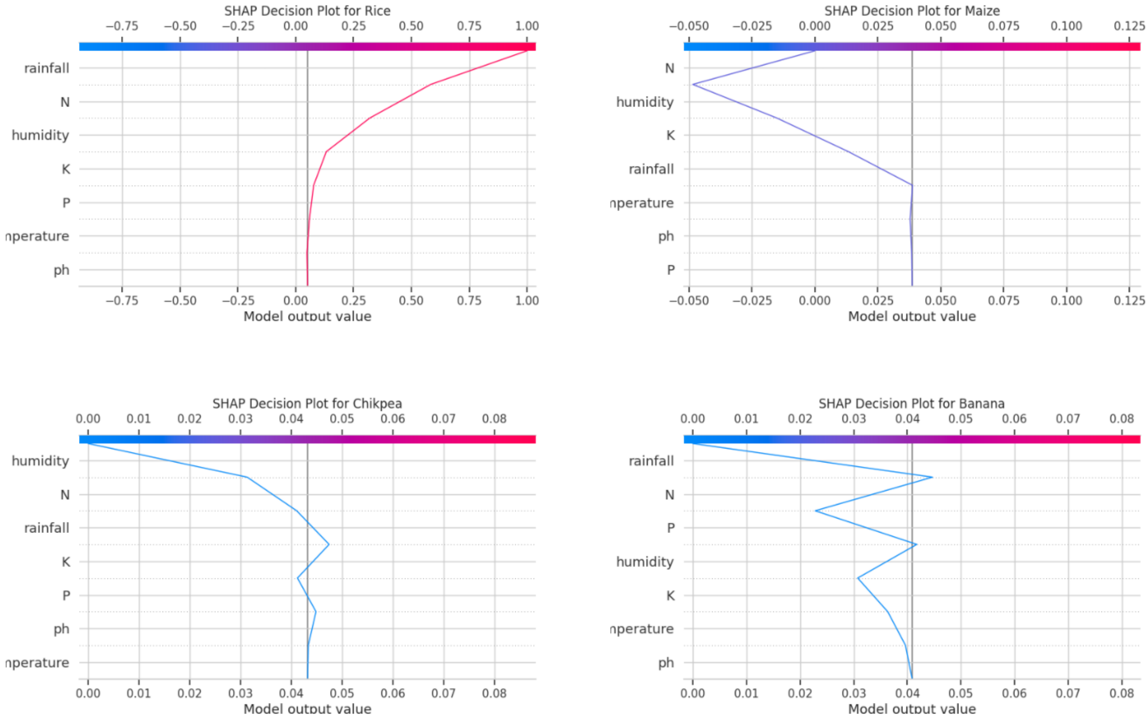


Figure 4.15: Decision plot illustrating the contribution of key features to individual predictions for the same crops.

The use of SHAP explanations makes the decision process transparent by identifying how each input variable contributes to each crop recommendation. This level of detail allows farmers, agronomists, and policy makers to verify the reasoning behind every prediction and ensures that the system operates as a clear and reliable decision support tool grounded in agronomic principles.

4.8 Discussion

Improving crop selection systems in modern agriculture requires a careful balance between predictive accuracy and interpretability. Because soils differ widely, climate conditions change over time, and farming decisions carry significant economic risk, the practical use of machine learning (ML) models depends not only on high performance but also on clear explanations. Incorrect crop recommendations can lead to yield reduction, wasted resources, and reduced confidence among farmers and stakeholders. For this reason, both reliable prediction and understandable reasoning are essential to support real-world agricultural decisions.

The proposed CS-AdaRF-SHAP framework shows strong performance, mainly through its adaptive reweighting strategy that gives greater attention to difficult cases during training. This mechanism allows the ensemble to handle variation within crop classes and to separate crops with similar characteristics, such as maize and chickpea. The model achieves very low rates of false positives and false negatives, a property that is especially important in agricultural applications where even a single incorrect recommendation, for example, suggesting maize in a nitrogen-deficient area, can lead to economic loss and environmental harm. The high accuracy and minimal error rates observed in the test results indicate a direct and practical benefit for farming decisions.

A comparative evaluation shows that the CS-AdaRF model performs better than the other tested methods, reaching a test accuracy of 99.77%. This score exceeds the results of Random Forest (99.45%), IoT-based frameworks (98%) [94, 102], and ACRM (98.7% for maize and 98.1% for rice) [100]. The model also achieved perfect values for F1-score, precision, and recall, a result supported by the balanced dataset and the careful design of the experiments. In addition, the testing time was efficient at 0.57 seconds, providing a clear advantage over more computationally demanding approaches such as XGBoost and Bagging.

Interpretability provided by SHAP analysis is a key element of the model's usefulness. SHAP ranks feature importance at the global level and also measures how each input, such as humidity, nitrogen, or potassium, affects individual crop selection. The model's reasoning agrees with established agronomic knowledge. For example, the strong role of humidity in SHAP results matches its well-known influence on crop water use, while nitrogen and potassium remain essential nutrients for healthy plant growth. This agreement with agricultural science strengthens trust and supports reliable recommendations.

Clear explanations are also offered through SHAP visualizations, which allow farmers to see the soil and climate factors that guide each recommendation. For instance, the system shows how rainfall affects rice selection or how humidity influences mung bean decisions. These visual outputs connect machine learning results to practical farming choices, encouraging user confidence, easing technology adoption, and supporting the design of decision tools that serve real agricultural needs.

Despite the strong performance of the proposed framework, some limitations remain. Although the dataset covers a wide range of crops and environmental conditions, it does not fully represent all global agro-ecological settings. Future studies should examine how well the model adapts to new regions and how robust it remains when inputs contain noise or measurement errors. From the perspective of interpretability, SHAP explanations, while effective, can be computationally demanding for large ensembles and may lose accuracy when features are highly correlated. Addressing these issues will require enlarging the dataset to capture broader variability, testing the model under real-world uncertainties, and exploring advanced interpretability techniques such as feature grouping or dimensionality reduction.

In addition, future research should focus on user-centered evaluation. Structured usability studies with farmers and agricultural advisors are essential to improve how explanations are presented, ensuring that outputs are clear, trusted, and practical. Feedback from stakeholders will play an important role in shaping the next generation of explainable crop recommendation systems and in strengthening both the scientific and practical value of AI in agriculture.

4.9 Conclusion

This chapter introduced an interpretable crop selection system designed to deliver both accurate predictions and clear explanations. The proposed CS-AdaRF-SHAP framework generates dependable crop recommendations while revealing the influence of key soil and climate variables on each decision. The results show that high predictive performance can be combined with transparent reasoning, supporting practical and trusted decision-making in agriculture.

After addressing the question *"what to plant?"*, the next logical challenge is *"how much to expect?"*. The following chapter examines this issue by focusing on crop yield estimation, aiming to predict the expected production level.

Chapter 5

Contribution 2: Data-Driven Crop Yield Prediction

5.1 Introduction

The previous chapter showed how interpretable machine learning can support strategic crop selection by providing farmers and agronomists with transparent, easily explained, and data-based recommendations. However, choosing the most suitable crop is only the first step in the broader set of decisions involved in precision agriculture. Once the question of “what to plant?” is resolved, the next challenge is estimating the expected yield, or “how much to expect?” This stage is essential for guiding farm management practices, planning the use of resources, and preparing for participation in agricultural markets.

In this chapter, we focus on the problem of forecasting tomato yields in greenhouse production. To address this task, we develop and evaluate a Stacked Ensemble Learning Model designed to integrate diverse sources of information and improve predictive accuracy.

The structure of this chapter is organized as follows. Section 5.2 describes the system architecture, including an outline of the stacked ensemble framework, the characteristics of the greenhouse tomato dataset, and the preprocessing steps such as data cleaning, temporal alignment, normalization, augmentation, and feature selection. Section 5.3 examines the predictive performance of the proposed model, comparing it with alternative machine learning approaches and interpreting the outcomes using both numerical indicators and visual analysis.

5.2 Materials and Methods

5.2.1 System Architecture Overview

The proposed system is built on a Stacked Ensemble Learning framework designed to provide reliable daily predictions of tomato yield in greenhouse settings. As shown in Figure 5.1, the architecture is organized into two main phases: an offline phase dedicated to model development and an online phase for real-time prediction.

In the offline phase, historical greenhouse data are collected, including environmental variables, crop growth characteristics, and yield records. These data pass through several preprocessing steps such as cleaning, normalization, and feature engineering. The resulting dataset is then used to train the Stacked Ensemble Model.

In the online phase, the trained model is applied to test data in order to generate yield predictions. Its performance is carefully evaluated and compared with alternative regression models, including KNN, Random Forest, and LightGBM, to determine the most accurate and operationally suitable approach for greenhouse management.

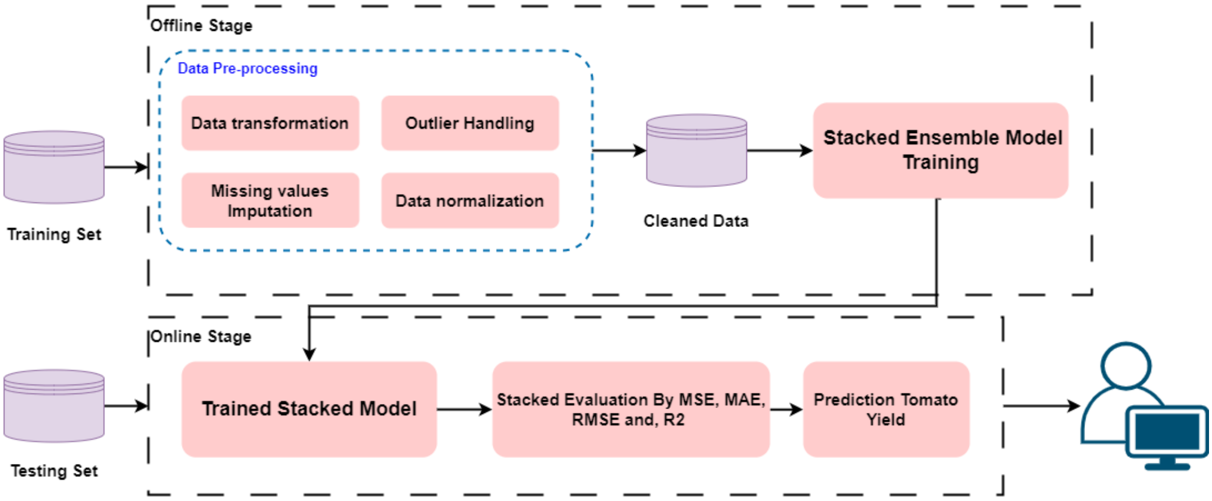


Figure 5.1: General architecture of the proposed system.

5.2.2 Dataset Description

The dataset used in this work is obtained from the second edition of the Autonomous Greenhouse Challenge (AGC) [117]. It offers detailed records of tomato production and crop management carried out under controlled greenhouse conditions. The data were collected between November 1, 2019 and April 30, 2020 from several teams responsible for 96 m^2 greenhouse units at Wageningen University & Research in Bleiswijk. The dataset contains information on key environmental variables such as air temperature, natural and supplemental light, heating inputs, and CO_2 concentration. It also includes cultivation parameters, for example plant density and stem density, which reflect the structural management of the crop.

5.2.3 Data Preprocessing

Preparing the dataset is an important stage in the development of a tomato yield prediction model. This stage includes several tasks designed to improve the quality and reliability of the data before modeling. Missing values are addressed through suitable replacement methods, while unusual or extreme measurements are detected and corrected to reduce their impact. Since the raw data were collected at different time intervals, they are aligned to a daily frequency to create a consistent timeline. All variables are then scaled to a common range so that they can be compared fairly. To increase the amount of training data, augmentation techniques are applied, and finally, the Boruta algorithm is used to select the most informative features for model training.

Handling Missing Values

Missing data were managed through median imputation. For every feature that contained missing entries, the absent values were replaced with the median of the available observations within that feature. The choice of the median, rather than the mean, helps reduce the influence of extreme values and skewed distributions. This method provides a simple yet reliable way to maintain the general characteristics of the dataset without introducing strong biases.

Handling Outliers

Outliers were identified using the interquartile range (IQR) method. In this approach, any observation that fell below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR was marked as an outlier. Instead of removing these values, which could lead to the loss of useful data, they were replaced with the median of the corresponding feature. Using the median in this way reduces the influence of extreme points while preserving the overall distribution of the dataset.

Data Transformation

For the construction of a unified model, it is important that all features in the dataset are aligned in time and expressed on a consistent scale. This requirement is particularly relevant when dealing with heterogeneous time series data. In the present case, the raw dataset contains several subsets recorded at different temporal resolutions, including measurements taken every five minutes, as well as daily and weekly records. Without adjusting these differences, the data cannot be combined in a meaningful way, which makes temporal harmonization an essential part of the preparation process.

To prepare a coherent dataset for supervised learning, all variables were expressed on a common daily interval. This transformation allowed different subsets of the data to be integrated into a single structure suitable for analysis and modeling.

- **5-Minute to Daily Aggregation:** The “Weather” and “Greenhouse Climate” subsets were originally collected at 5-minute intervals. To convert these into daily values, the mean for each variable was calculated across all records for a given day. This step reduces the overall data volume and lowers computational requirements, while still capturing the main daily patterns. In addition, the use of daily averages helps smooth out short-term fluctuations and potential noise from sensors:

  $$\text{Daily Mean}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} x_{i,d} \tag{5.1}$$

  where N_d is the number of 5-minute observations in day d, and $x_{i,d}$ represents the i-th measurement on that day.

- **Weekly to Daily Interpolation:** The “Production” and “Crop Parameter” subsets were recorded on a weekly basis. To align them with the daily series, values were estimated using Lagrange polynomial interpolation. This method produces a smooth daily curve that reflects the underlying variation in the original weekly data. The resulting series allows the inclusion of crop-related variables in day-level analyses and supports the training of predictive models that require uniform temporal resolution.

After temporal harmonization, the various subsets were merged into a unified daily dataset. In this structure, the harvest variable was expressed on a daily scale and aligned with the corresponding predictors, as illustrated in Figure 5.2. This step ensures that each record contains both the input features and the target variable in a synchronized manner, providing a reliable foundation for model development.

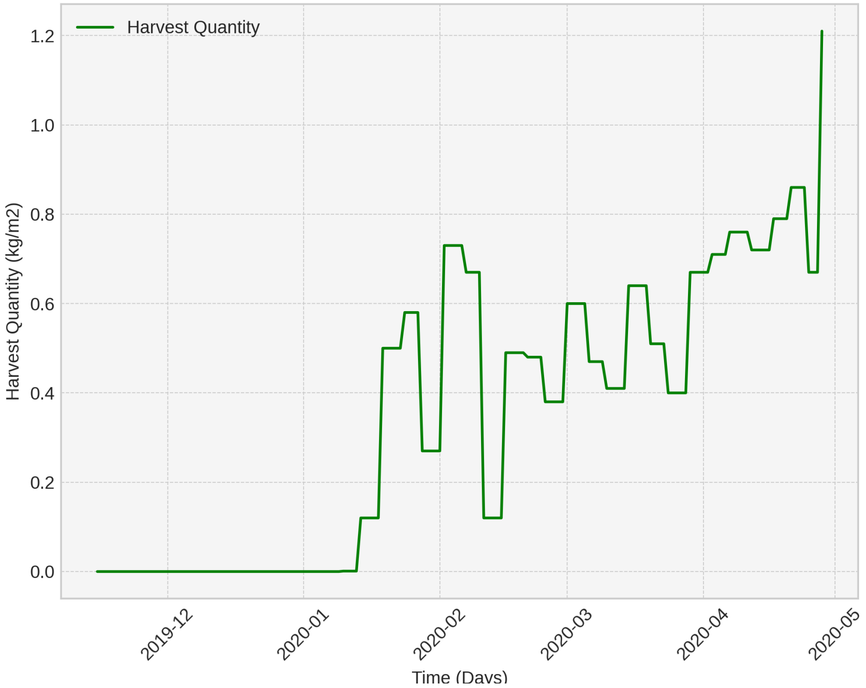


Figure 5.2: Schematic representation of the transformation of harvest data to a daily resolution.

Data Normalization

To place all features on a comparable scale, Min-Max normalization was applied. This method transforms each value into the range $[0, 1]$, which supports stable training and improves the efficiency of the learning process.

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{5.2}$$

where X is the original value, X_{min} and X_{max} are the minimum and maximum values of the feature, and X_{scaled} is the normalized output.

Data Augmentation

Since the dataset contained only 166 samples, data augmentation was applied to increase its size and reduce the risk of overfitting. We used *random noise augmentation* [118], where small variations were added to the original values. This creates new samples that reflect realistic variability in greenhouse conditions, helping the model learn more robust patterns.

The augmentation process is mathematically defined as:

$$X_{\text{augmented}} = X_{\text{original}} + \epsilon \tag{5.3}$$

where

$$\epsilon \sim \mathcal{N}\left(0,\ (0.01 \cdot \text{std}(X_{\text{original}}))^2\right) \tag{5.4}$$

Here, ϵ is a vector of random noise sampled from a normal distribution with zero mean and a standard deviation equal to 1% of the feature's standard deviation. This design ensures that the added noise remains small in scale, preserving the original statistical properties while introducing enough variability to improve model robustness. After augmentation, the dataset was expanded to 500 samples, providing a stronger basis for model training.

Feature Selection

Tomato yield depends on a wide range of environmental, physiological, and management factors. Using too many input variables, however, can make the model unnecessarily complex, increase computational requirements, and raise the risk of overfitting. To reduce these issues, a structured feature selection method was applied in order to retain only those variables that provide meaningful predictive value for yield estimation.

For this purpose, we used the Boruta algorithm [119], a wrapper-based method built around Random Forests. The approach works by comparing the importance of actual features with that of “shadow” features, which are created by randomly permuting the data. Only variables that show statistically significant predictive power compared to the shadow features are kept. The main steps of the Boruta procedure are summarized in Algorithm 3.

Algorithm 3 Boruta Feature Selection

```
Input: Dataset $X$ with $n$ samples and $p$ features, target variable $y$
Output: Subset of important features $S$
Generate $m$ shadow features by randomly permuting each original feature
while feature importance ranking not stable do
    Train a Random Forest regressor on the extended dataset ($X$ + shadow features)
    Compute importance scores for all features
    For each original feature, compare its importance with the maximum importance among the shadow features
    Keep features that show higher importance than the shadow features; remove those that do not
end while
Return Final set $S$ of selected features
```

Applying the Boruta method to our dataset, which originally included 39 features, reduced the number of inputs to 11. This represents a reduction of about 77% in dimensionality. Such a decrease not only lowers computational cost but also improves the clarity of the model by directing attention toward variables that are most relevant to tomato production. Table 5.1 lists the features selected by Boruta for yield prediction, along with their descriptions and measurement units.

5.3 Results and Discussion

This section presents and discusses the results of the comparative study on tomato yield prediction. Four machine learning models were evaluated: K-Nearest Neighbors (KNN), Random Forest, LightGBM, and the proposed Stacked Ensemble Model. The dataset was

Table 5.1: Features selected by Boruta for tomato yield prediction

Feature	**Description**	**Unit**
Tair	Greenhouse air temperature	°C
Rhair	Relative humidity in greenhouse	%
CO2air	CO_2 concentration	ppm
Tot_PAR	Total inside PAR (Sun + HPS + LED)	µmol/m^2 s
pH_drain_PC	Drainage pH	–
EC_drain_PC	Drainage electrical conductivity	dS/m
Cum_irr	Cumulative irrigation per day	L/m^2
Stem_elong	Stem growth	cm/week
Stem_dens	Stem density	Stems/m^2
Plant_dens	Plant density	Plants/m^2
Stem_thick	Stem thickness	mm
Prod	Tomato yield (target)	kg/m^2

divided into two subsets, with 80% allocated for training and 20% reserved for testing. To ensure a fair comparison, all models were tuned through hyperparameter optimization before the evaluation.

5.3.1 Predictive Performance Comparison

Table 5.2 reports the main performance measures used to evaluate the models, namely mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). These results are based on the test set. For clarity, graphical comparisons of the performance are also provided in Figures 5.3, 5.4, and 5.5.

Table 5.2: Performance Evaluation of Tomato Yield Prediction Models

Model	**MSE**	**MAE**	**RMSE**	**R^2**
KNN	0.023	0.110	0.150	0.712
Random Forest	0.009	0.046	0.095	0.884
LightGBM	0.013	0.083	0.114	0.831
Stacked Ensemble	**0.0080**	**0.065**	**0.090**	**0.896**

As shown in Figure 5.3, the Stacked Ensemble Model obtained the highest R^2 value (0.896), followed by Random Forest (0.884), LightGBM (0.831), and KNN (0.712). This outcome suggests that the ensemble approach captured the variability in tomato yield more effectively, offering a stronger and more reliable fit than the other models.

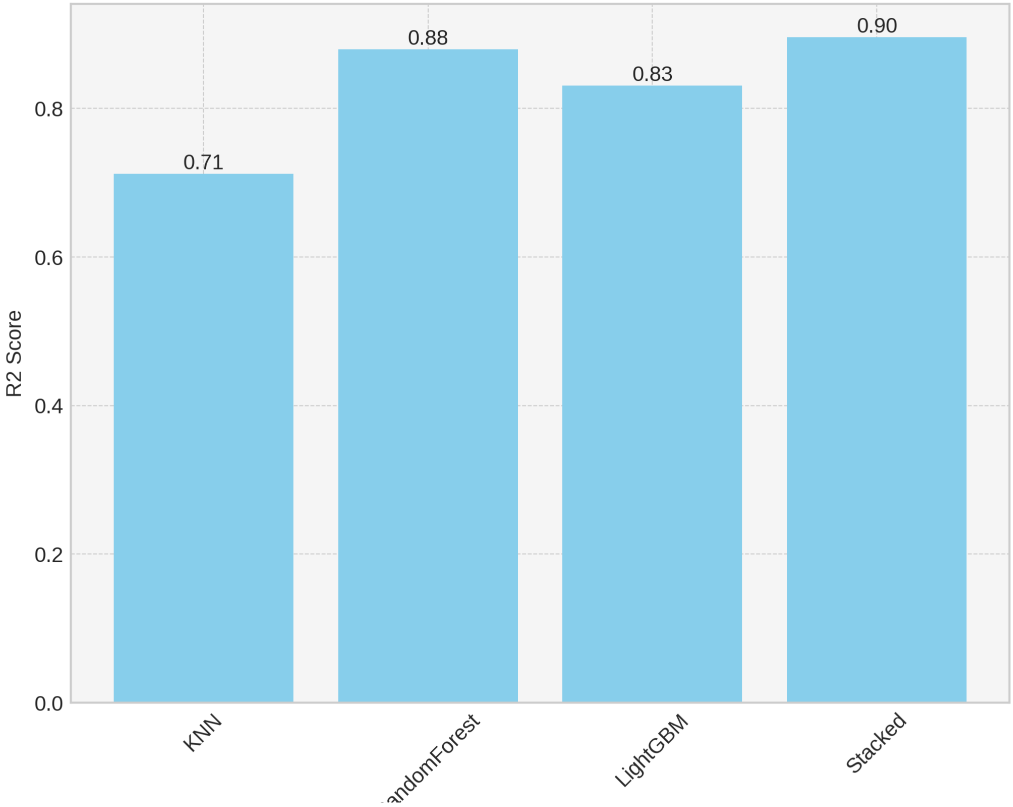


Figure 5.3: R^2 scores for the tested models.

5.3.2 Error Analysis and Robustness

Figure 5.4 presents a comparison of the error metrics for all models. The Stacked Ensemble Model achieved the lowest mean squared error (0.008) and root mean squared error (0.09), showing that it can deliver accurate yield predictions with only small deviations from observed values. Random Forest and LightGBM also performed well, but the stacking approach provided a modest improvement by drawing on the strengths of multiple base learners.

The robustness of the ensemble is reflected in its ability to reduce both bias and variance. By combining predictions from different learners such as Ridge, Random Forest, and XGBoost, the stacked model counterbalances the tendency of single algorithms to either underfit or overfit. This leads to stronger generalization on unseen data, which is particularly important in agricultural applications where variability and complex non-linear relationships are common.

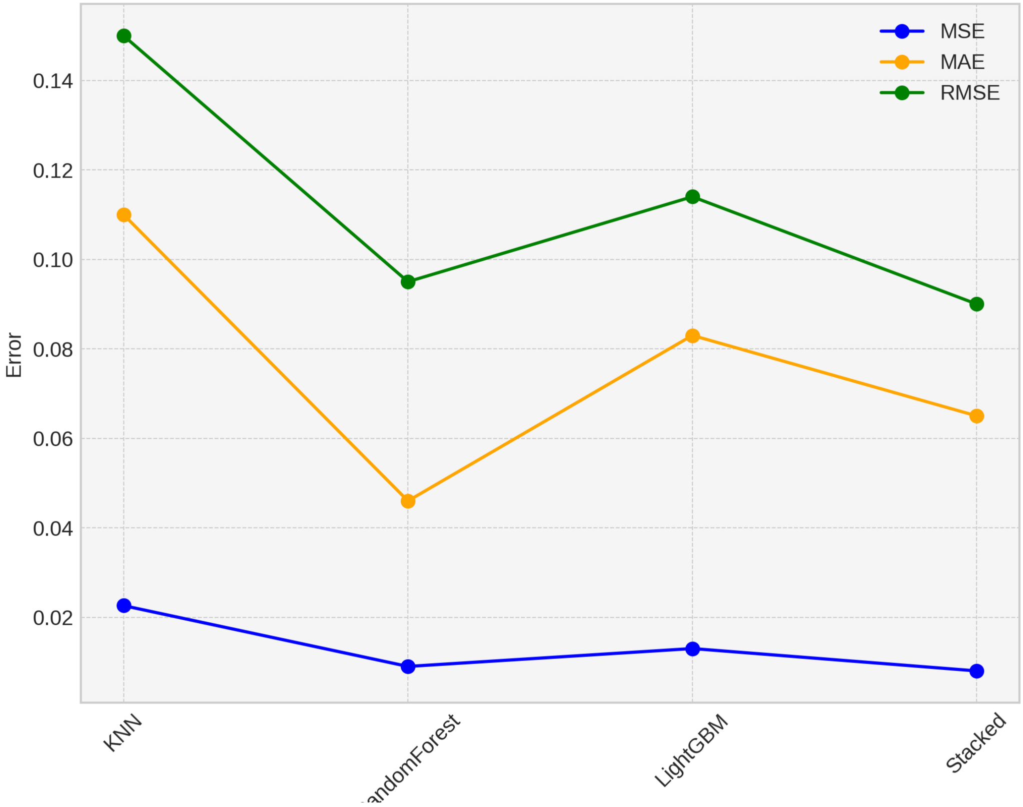


Figure 5.4: Error metrics (MSE, MAE, RMSE) for the evaluated models.

5.3.3 Alignment of Predicted and Actual Yields

Figure 5.5 shows the relationship between predicted and observed yields for all models. In the ideal case, predictions would fall exactly on the diagonal line, representing a perfect match with the measured values.

Among the tested approaches, the Stacked Ensemble Model produced predictions that lie closest to the diagonal, indicating strong agreement with the actual harvest data. The other models show more scattered points and visible deviations, which reflects lower accuracy and reduced reliability in capturing yield variation.

The findings show that the Stacked Ensemble Model offers clear advantages for predicting yield in greenhouse tomato production. Its lower error values and higher R^2 scores indicate that it can be a practical tool for use in precision agriculture. Reliable forecasts of yield can support better planning of resources, guide crop management decisions, and improve marketing strategies, which together can increase productivity, reduce waste,

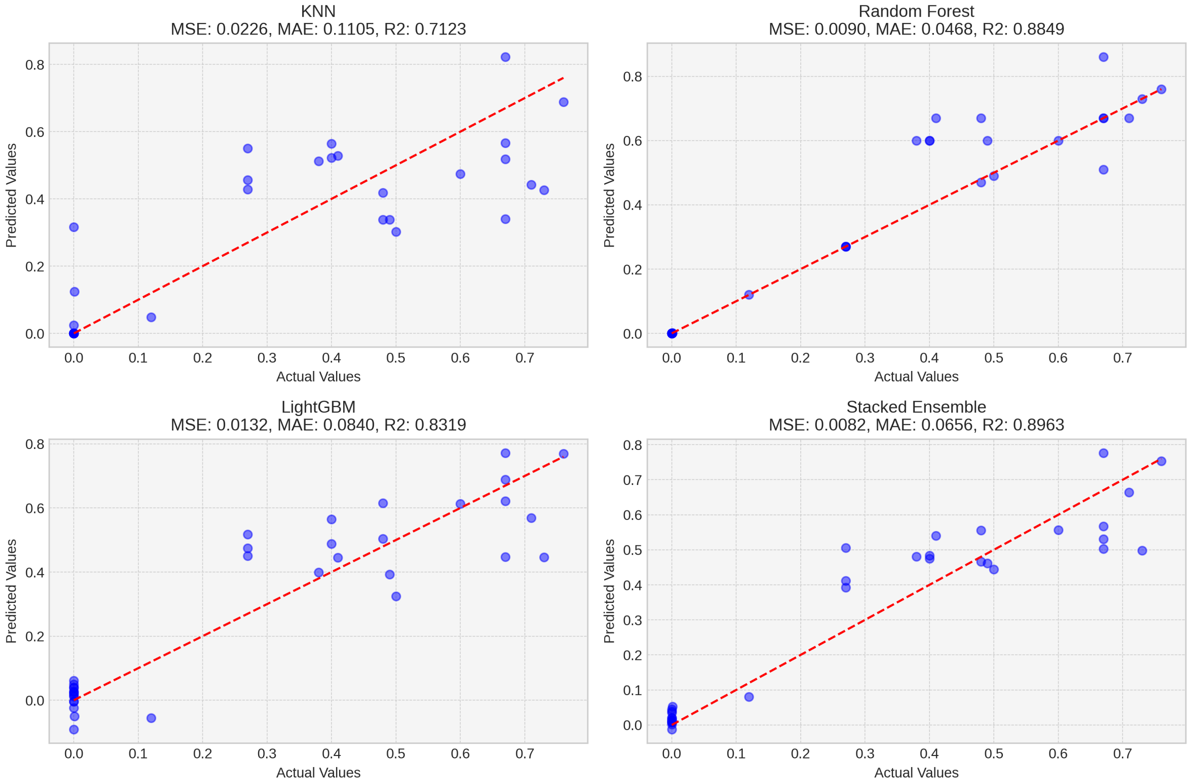


Figure 5.5: Predicted versus observed yields for the evaluated models.

and strengthen economic returns for growers.

5.4 Conclusion

This chapter presented a Stacked Ensemble Learning Model for predicting crop yields in greenhouse environments, with tomato production serving as a case study. The next chapter turns to a key question in data-driven agriculture: *“How can the data supporting these decisions remain secure, reliable, and trustworthy?”* To address this, we examine blockchain-based approaches for safeguarding data integrity and enhancing security in smart agriculture.

Chapter 6

Contribution 3: Blockchain-Based Approach to Securing Data in Smart Agriculture

6.1 Introduction

The previous chapters of the thesis have addressed two central aspects of precision agriculture: an interpretable crop selection system and a data-driven crop yield prediction in greenhouse environments. The first contribution addressed the question of *"what to plant?"* by developing an interpretable crop selection system that combines strong predictive accuracy with clear explanations of the factors influencing each recommendation. This transparency enables farmers and agronomists to understand why a particular crop is suggested, fostering confidence and supporting real-world adoption. The second contribution addressed the question of *"how much to expect?"* by applying machine learning to predict tomato yield as a case study, focusing on greenhouse production where rich and structured data are available.

Having answered these two questions, a third and equally critical challenge now emerges at the heart of data-driven agriculture: *"How can the data that supports these decisions remain secure, reliable, and trustworthy?"*

Modern smart farming generates and exchanges massive volumes of data, from sensor measurements to operational records, across networks of farms, research institutions, and service providers. The value of intelligent decision-support systems built on this information depends on its integrity and security. Without reliable safeguards to ensure accuracy, privacy, and controlled access, even advanced predictive models risk producing unreliable results, as their outputs rely on data that may be incomplete, altered, or inaccessible.

To address these challenges, this chapter introduces a blockchain-based aproach for securing and managing agricultural data. Building on the IoT-driven infrastructure, the proposed system combines edge computing, blockchain technology, and distributed file storage (IPFS) to deliver a secure and transparent approach to agricultural data management. Through the use of cryptographic methods and smart contracts, the system ensures that all transactions remain immutable and auditable, while also supporting data privacy and controlled access.

The remainder of this chapter is structured as follows: Section 6.2 presents the proposed approach and its overall architecture. Section 6.3 describes the implementation process, covering development tools, smart contract deployment, data encryption, secure storage, and performance evaluation. Section 6.4 concludes the chapter with a summary of the main contributions and results.

6.2 Proposed Solution

Digitalisation has become a key driver of economic growth in many sectors, including agriculture. In Algeria, agriculture remains central to both social and economic development, and the government has placed strong attention on digital transformation programs to improve efficiency and sustainability.

The proposed system is designed to connect all agricultural sites (AS) under a single secure platform managed by the government institution (GI). This framework enables a unified process for collecting, storing, and sharing agricultural data between the sites and the GI. To guarantee data security and integrity, blockchain technology is used as the foundation of the data management system.

As shown in Fig. 6.1, the architecture places the GI as the main authority supervising data exchanges. Raw data gathered from greenhouses is first stored at the edge to preserve authenticity. After encryption, the data is shared across the network, where access is limited to authorized users. This process protects ownership and confidentiality while preventing unauthorized use or alteration of the information.

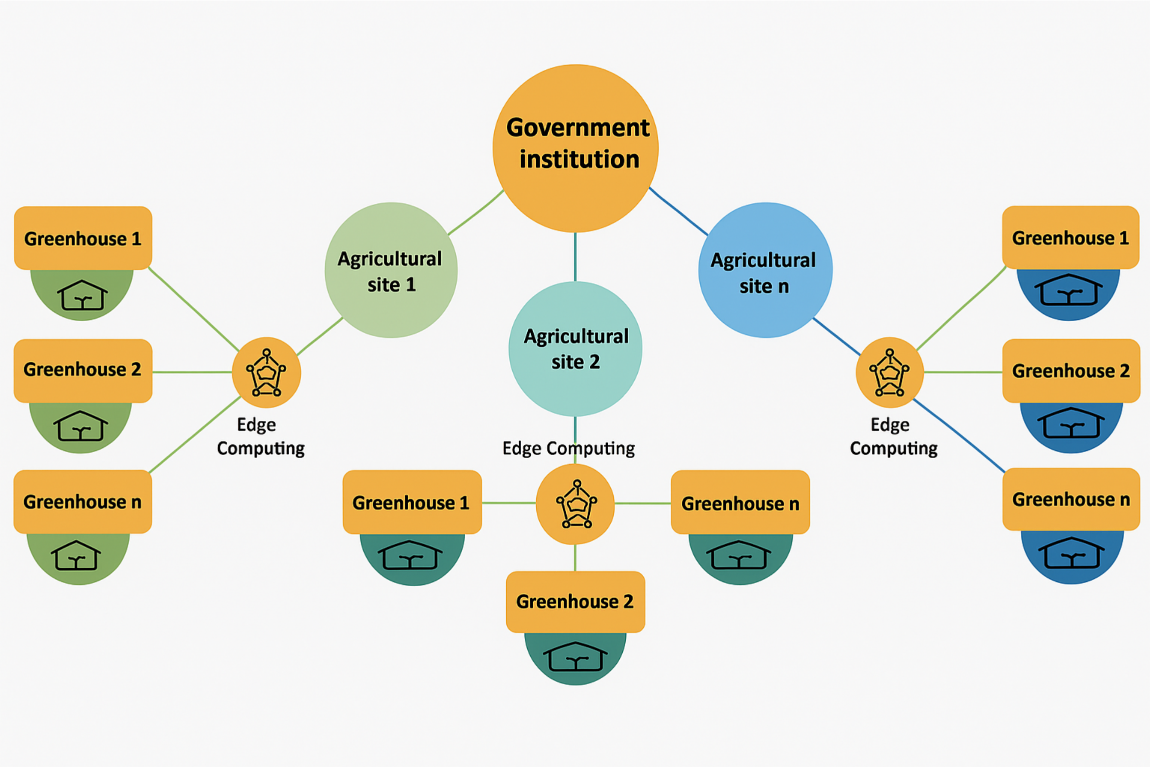


Figure 6.1: Hierarchy of the proposed system

6.2.1 Architecture of the Proposed Approach

The proposed approach involves two main actors: the agricultural sites (AS) and the government institution (GI). Each site is formally registered with the GI and is given a unique address. This address allows the site to access the platform securely and to carry out authorized operations within the system.

The current design concentrates on handling raw data produced by IoT devices installed in greenhouses. At the same time, the architecture has been built with flexibility in mind, making it possible to expand to other forms of information such as farmer records, crop production logs, and weather conditions when needed.

To safeguard data throughout its entire lifecycle, the architecture integrates blockchain technology, smart contracts, the InterPlanetary File System (IPFS), and strong encryption

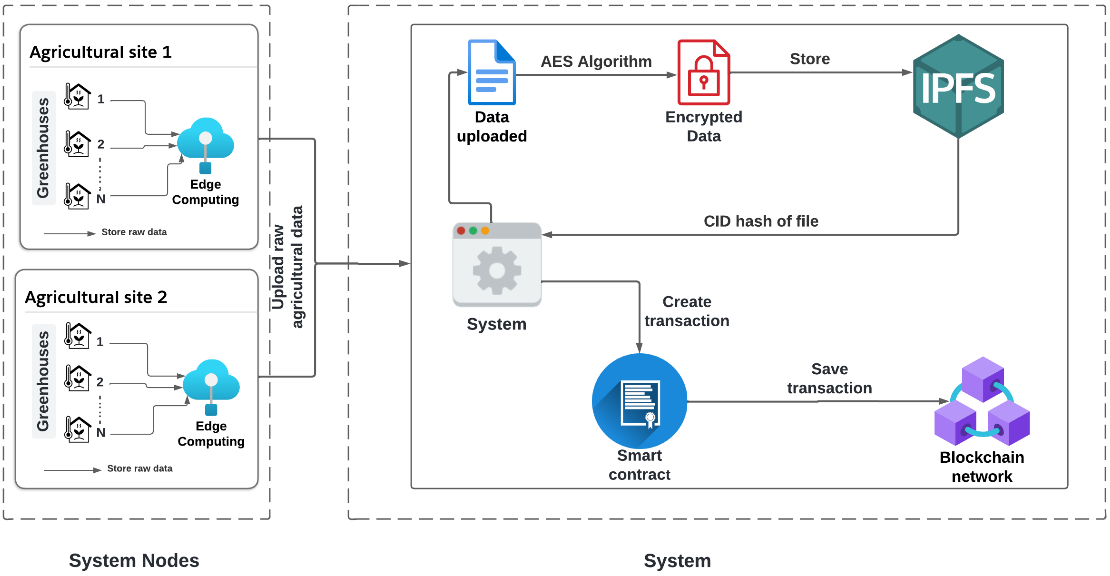


Figure 6.2: General architecture of the proposed system

methods. These components work together to provide security, transparency, and integrity in the management of agricultural data. The complete system architecture is presented in Fig. 6.2.

The workflow for data processing starts with the collection of raw data from each greenhouse, which is first stored locally at the edge. At this point, the data remains unchanged to preserve its original form. To maintain integrity, a SHA256 hash is calculated for the collected data. After this step, the data is encrypted using the AES algorithm, as shown in Fig. 6.3.

The encrypted files are then uploaded to the InterPlanetary File System (IPFS) for decentralized storage. Each file stored in IPFS is assigned a unique Content Identifier (CID), which acts as a cryptographic hash to ensure accurate retrieval and verification. Once this is done, a smart contract records a transaction that contains the key metadata. This transaction is written to the blockchain, creating a permanent and verifiable record of the data submission process.

6.2.2 Structure of Blocks and Transactions

In the proposed approach, each block is divided into two main parts: the header and the body. The header stores the key metadata of the block, which includes the block index,

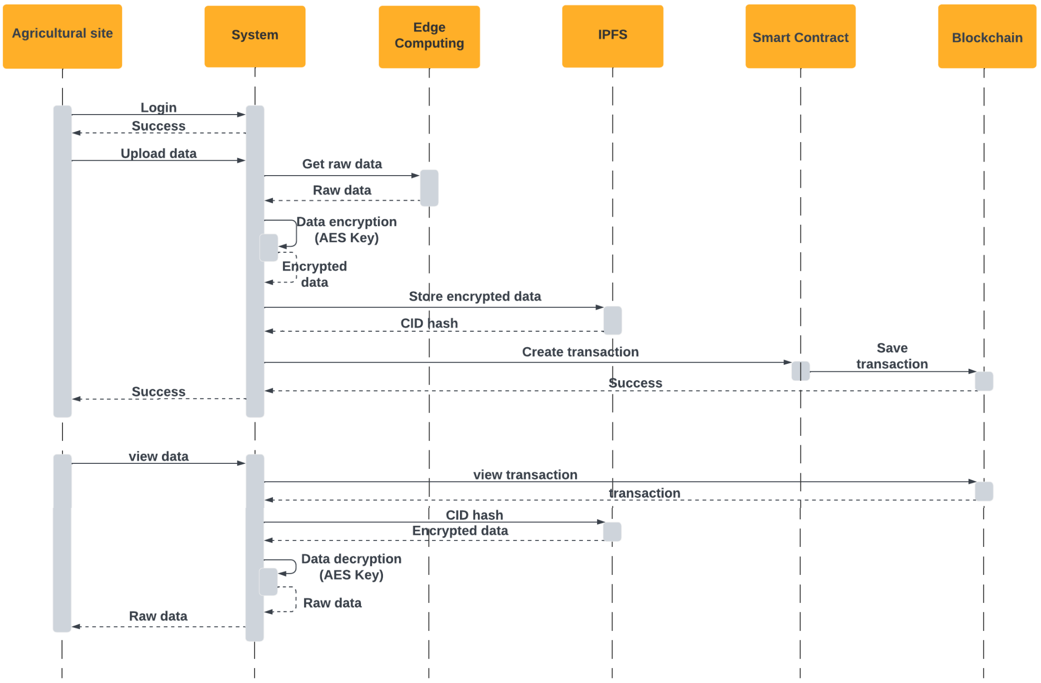


Figure 6.3: Sequence diagram of the proposed system.

timestamp, nonce (a value used only once), the hash of the previous block, and the hash of the current block. This arrangement guarantees the link between consecutive blocks and preserves the immutability of the entire chain.

The body of the block contains the transactions recorded at that point. Depending on the circumstances, it may hold a single transaction or a group of several transactions bundled together.

Each transaction in the system is described by the following fields:

- **Transaction:** A unique hash that identifies the transaction.
- **From:** The address of the agricultural data owner.
- **To:** The address of the government institution.
- **Data:** Details about the shared data, including the IPFS hash (Content Identifier), the file name, and the file hash.

All transactions and data transfers are encrypted, which ensures that only the legitimate data owner can decrypt and access the original information. To maintain integrity, the system allows nodes to display past transactions and their related data, providing a clear record for auditing. When data is decrypted, a verification step is carried out by comparing it with the locally stored hash. This guarantees that the data remains genuine and unchanged throughout its entire lifecycle.

6.3 Implementation

This section explains how the proposed system was implemented in practice. It presents the development environment and outlines the main technologies used to build the system.

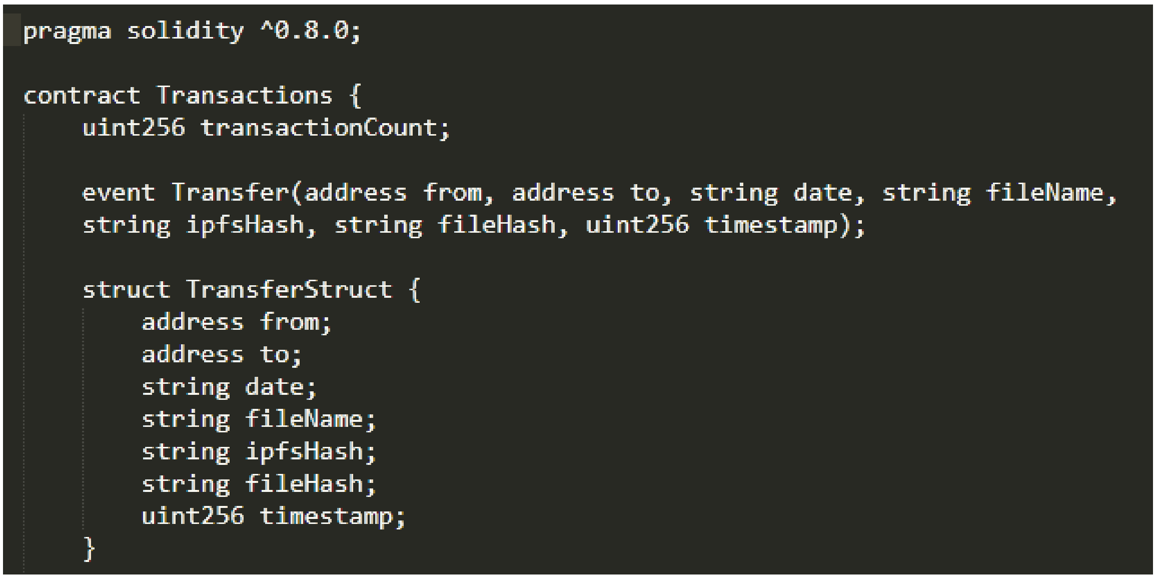
6.3.1 Development Environment and Tools

The implementation of the system relied on a set of tools and platforms chosen for their suitability in building secure and distributed applications:

- **Ethereum Blockchain:** Used to support smart contracts and decentralized operations. All transactions and contract logic were deployed and tested on the Goerli network, which provides a safe environment for development and experimentation.
- **Solidity:** The programming language employed to design and implement the smart contract logic.
- **IPFS:** Adopted for decentralized storage of encrypted agricultural data, ensuring integrity and protection against tampering.

6.3.2 Smart Contract Deployment and Data Transactions

Several Ethereum test networks are available, such as Ropsten, Rinkeby, and Goerli. In this work, the Goerli network was selected, and four nodes were simulated to represent the participants and simplify the operational setup. As shown in Fig. 6.4, the deployed smart contract records essential metadata for every transaction, including the sender, recipient, date, timestamp, IPFS hash, file name, and file hash. The transaction hash is automatically generated by the network, ensuring that each transaction is uniquely identifiable and permanently stored.

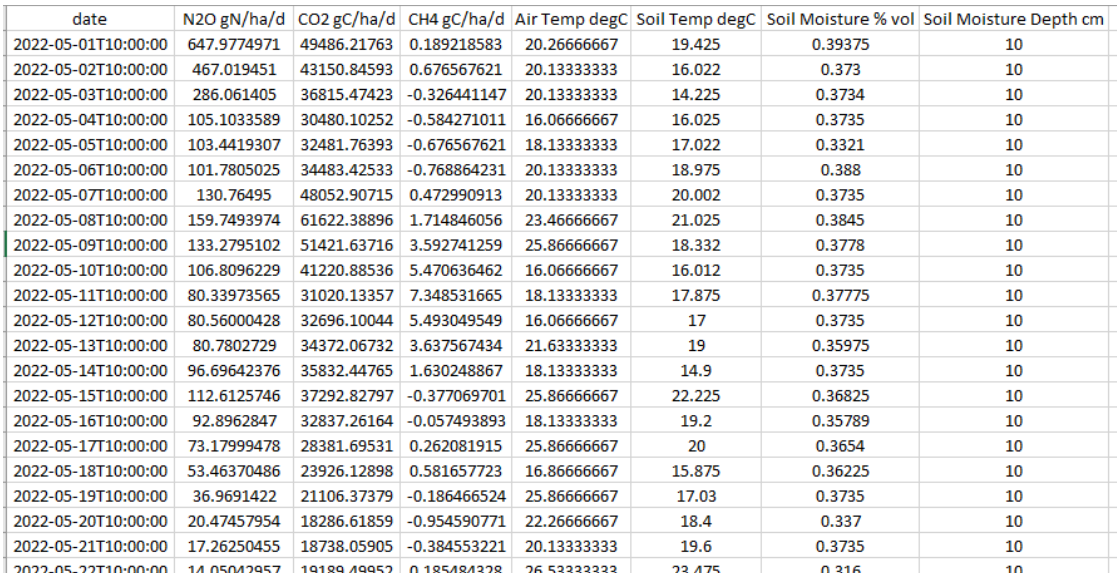


```
pragma solidity ^0.8.0;

contract Transactions {
    uint256 transactionCount;

    event Transfer(address from, address to, string date, string fileName,
    string ipfsHash, string fileHash, uint256 timestamp);

    struct TransferStruct {
        address from;
        address to;
        string date;
        string fileName;
        string ipfsHash;
        string fileHash;
        uint256 timestamp;
    }
```

Figure 6.4: Excerpt from the smart contract.

6.3.3 Data Encryption and Secure Storage

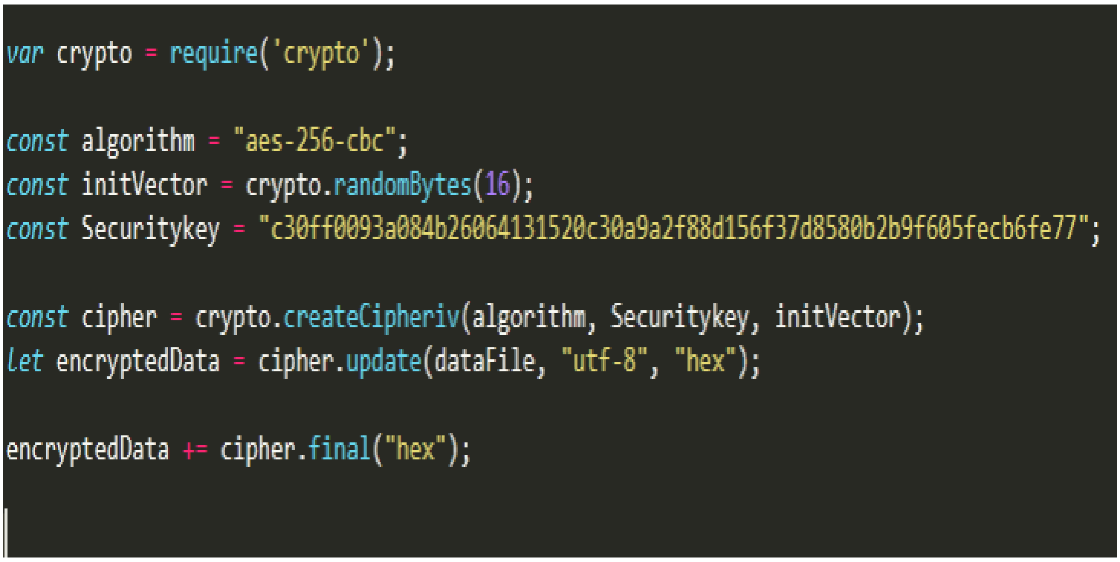
Raw agricultural data, an example of which is shown in Fig. 6.5, is first collected from greenhouse IoT devices. Before storage, the data is converted into a standardized text format and encrypted using the AES algorithm (Fig. 6.6). This process ensures that only authorized users can access the information, thereby maintaining confidentiality and privacy.



date	N2O gN/ha/d	CO2 gC/ha/d	CH4 gC/ha/d	Air Temp degC	Soil Temp degC	Soil Moisture % vol	Soil Moisture Depth cm
2022-05-01T10:00:00	647.9774971	49486.21763	0.189218583	20.26666667	19.425	0.39375	10
2022-05-02T10:00:00	467.019451	43150.84593	0.676567621	20.13333333	16.022	0.373	10
2022-05-03T10:00:00	286.061405	36815.47423	-0.326441147	20.13333333	14.225	0.3734	10
2022-05-04T10:00:00	105.1033589	30480.10252	-0.584271011	16.06666667	16.025	0.3735	10
2022-05-05T10:00:00	103.4419307	32481.76393	-0.676567621	18.13333333	17.022	0.3321	10
2022-05-06T10:00:00	101.7805025	34483.42533	-0.768864231	20.13333333	18.975	0.388	10
2022-05-07T10:00:00	130.76495	48052.90715	0.472990913	20.13333333	20.002	0.3735	10
2022-05-08T10:00:00	159.7493974	61622.38896	1.714846056	23.46666667	21.025	0.3845	10
2022-05-09T10:00:00	133.2795102	51421.63716	3.592741259	25.86666667	18.332	0.3778	10
2022-05-10T10:00:00	106.8096229	41220.88536	5.470636462	16.06666667	16.012	0.3735	10
2022-05-11T10:00:00	80.33973565	31020.13357	7.348531665	18.13333333	17.875	0.37775	10
2022-05-12T10:00:00	80.56000428	32696.10044	5.493049549	16.06666667	17	0.3735	10
2022-05-13T10:00:00	80.7802729	34372.06732	3.637567434	21.63333333	19	0.35975	10
2022-05-14T10:00:00	96.69642376	35832.44765	1.630248867	18.13333333	14.9	0.3735	10
2022-05-15T10:00:00	112.6125746	37292.82797	-0.377069701	25.86666667	22.225	0.36825	10
2022-05-16T10:00:00	92.8962847	32837.26164	-0.057493893	18.13333333	19.2	0.35789	10
2022-05-17T10:00:00	73.17999478	28381.69531	0.262081915	25.86666667	20	0.3654	10
2022-05-18T10:00:00	53.46370486	23926.12898	0.581657723	16.86666667	15.875	0.36225	10
2022-05-19T10:00:00	36.9691422	21106.37379	-0.186466524	25.86666667	17.03	0.3735	10
2022-05-20T10:00:00	20.47457954	18286.61859	-0.954590771	22.26666667	18.4	0.337	10
2022-05-21T10:00:00	17.26250455	18738.05905	-0.384553221	20.13333333	19.6	0.3735	10
2022-05-22T10:00:00	14.05042957	19189.49952	0.185484328	26.53333333	23.475	0.316	10

Figure 6.5: Example of raw agricultural data.

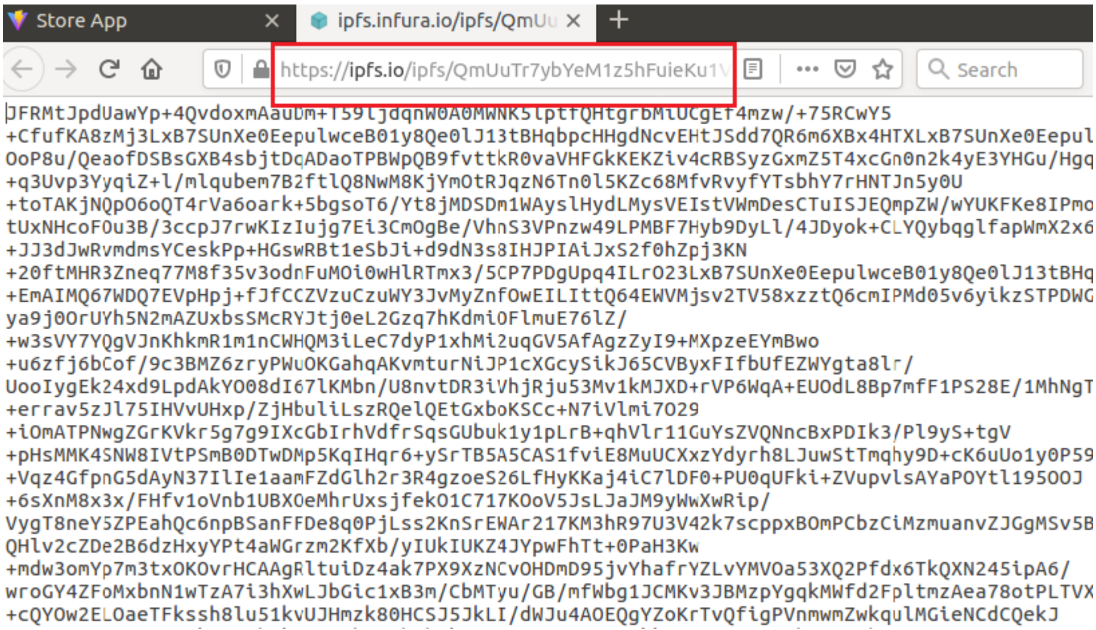
After encryption, the files are uploaded to IPFS, which assigns a unique Content Identifier



```
var crypto = require('crypto');

const algorithm = "aes-256-cbc";
const initVector = crypto.randomBytes(16);
const Securitykey = "c30ff0093a084b26064131520c30a9a2f88d156f37d8580b2b9f605fecb6fe77";

const cipher = crypto.createCipheriv(algorithm, Securitykey, initVector);
let encryptedData = cipher.update(dataFile, "utf-8", "hex");

encryptedData += cipher.final("hex");
```

Figure 6.6: AES encryption process.

(CID) to each entry. The CID serves as a cryptographic fingerprint, enabling both retrieval and verification of the stored data. Fig. 6.7 provides an example of encrypted agricultural data stored and accessed through the IPFS network.



```
JFRMtJpdUawYp+4QvdoxmAauDm+T59ljdqnW0A0MWNK5lptfQHtgrbMiUCgEf4mzw/+75RCwY5
+CfufKA8zMj3LxB7SUnXe0EepulwceB01y8Qe0lJ13tBHqbpcHHgdNcvEHtJSdd7QR6m6XBx4HTXLxB7SUnXe0Eepul
OoP8u/QeaofDSBsGXB4sbjtDqADaoTPBWpQB9fvttkR0vaVHFGkKEKZiv4cRBSyzGxmZ5T4xcGn0n2k4yE3YHGu/Hgq
+q3Uvp3YyqiZ+l/mlqubem7B2ftlQ8NwM8KjYmOtRJqzN6Tn0l5KZc68MfvRvyfYTsbhY7rHNTJn5y0U
+toTAKjNQpO6oQT4rVa6oark+5bgsoT6/Yt8jMDSDm1WAyslHydLMysVEIstVWmDesCTuISJEQmpZW/wYUKFKe8IPmo
tUxNHcoF0u3B/3ccpJ7rwKIzIujg7Ei3CmOgBe/VhnS3VPnzw49LPMBF7Hyb9DyLl/4JDyok+CLYQybqglfapWmX2x6
+JJ3dJwRvmdmsYCeskPp+HGswRBt1eSbJi+d9dN3s8IHJPIAiJxS2f0hZpj3KN
+20ftMHR3Zneq77M8f35v3odnFuMOi0wHlRTmx3/5CP7PDgUpq4ILrO23LxB7SUnXe0EepulwceB01y8Qe0lJ13tBHq
+EmAIMQ67WDQ7EVpHpj+fJfCCZVzuCzuWY3JvMyZnfOwEILIttQ64EWVMjsv2TV58xzztQ6cmIPMd05v6yikzSTPDWG
ya9j0OrUYh5N2mAZUxbsSMcRYJtj0eL2Gzq7hKdmiOFlmuE76lZ/
+w3sVY7YQgVJnKhkmR1m1nCWHQM3iLeC7dyP1xhMi2uqGV5AfAgzZyI9+MXpzeEYmBwo
+u6zfj6bCof/9c3BMZ6zryPWuOKGahqAKvmturNiJP1cXGcySikJ65CVByxFIfbUfEZWYgta8lr/
UooIygEk24xd9LpdAkYO08dI67lKMbn/U8nvtDR3iVhjRju53Mv1kMJXD+rVP6WqA+EUOdL8Bp7mfF1PS28E/1MhNgT
+errav5zJl75IHVvUHxp/ZjHbuliLszRQelQEtGxboKSCc+N7iVlmi7O29
+iOmATPNwgZGrKVkr5g7g9IXcGbIrhVdfrSqsGUbuk1y1pLrB+qhVlr11GuYsZVQNncBxPDIk3/Pl9yS+tgV
+pHsMMK4SNW8IVtPSmB0DTwDMp5KqIHqr6+ySrTB5A5CAS1fviE8MuUCXxzYdyrh8LJuwStTmqhy9D+cK6uUo1y0P59
+Vqz4GfpnG5dAyN37IlIe1aamFZdGlh2r3R4gzoeS26LfHyKKaj4iC7lDF0+PU0qUFki+ZVupvlsAYaPOYtl195OOJ
+6sXnM8x3x/FHfv1oVnb1UBX0eMhrUxsjfekO1C717KOoV5JsLJaJM9yWwXwRip/
VygT8neY5ZPEahQc6npBSanFFDe8q0PjLss2KnSrEWAr217KM3hR97U3V42k7scppxBOmPCbzCiMzmuanvZJGgMSv5B
QHlv2cZDe2B6dzHxyYPt4aWGrzm2KfXb/yIUkIUKZ4JYpwFhTt+0PaH3Kw
+mdw3omYp7m3txOKOvrHCAAgRltuiDz4ak7PX9XzNCvOHDmD95jvYhafrYZLvYMVOa53XQ2Pfdx6TkQXN245ipA6/
wroGY4ZFoMxbnN1wTzA7i3hXwLJbGic1xB3m/CbMTyu/GB/mfWbg1JCMKv3JBMzpYgqkMWfd2FpltmzAea78otPLTVX
+cQYOw2ELOaeTFkssh8lu51kvUJHmzk80HCSJ5JkLI/dWJu4AOEQgYZoKrTvQfigPVnmwmZwkqulMGieNCdCQekJ
```

Figure 6.7: Encrypted agricultural data stored on IPFS.

6.3.4 Performance and Security Analysis

The main goal of the system is to manage agricultural data collected by IoT devices in a secure and reliable manner. This includes not only safe storage but also protection of data integrity and controlled sharing between agricultural sites (AS) and the government institution (GI).

Security and integrity are achieved through two complementary layers:

- Large files are stored off-chain on IPFS, while key references such as the IPFS hash and file hash are stored on-chain. This approach improves efficiency while keeping verification straightforward.
- The blockchain ledger guarantees immutability. Once data is written, it cannot be changed or removed. Even in the case of network attacks or attempts to alter records, cryptographic validation and distributed storage ensure that the data remains intact and accessible.

During development and testing, the system showed reliable performance in recording, encrypting, storing, and retrieving agricultural data. Smart contracts executed transactions securely and in a transparent way, while the use of IPFS made it possible to retrieve and verify files using their unique CID. These results demonstrate the feasibility of the proposed architecture and its suitability for deployment in real agricultural environments.

6.4 Conclusion

This chapter presented a blockchain-based approach to secure agricultural data, ensuring integrity, confidentiality, and controlled sharing. By integrating edge computing, encryption, IPFS storage, and a private blockchain, the system protects agricultural data throughout its lifecycle and fosters trust among farmers and stakeholders.

Chapter 7

General Conclusion and Perspectives

This chapter presents the main contributions of this thesis and points to several research directions that deserve additional investigation in the future.

7.1 Summary

The research conducted in this thesis produced three principal contributions that address the questions introduced at the beginning of the work. These contributions are outlined below:

- **Interpretable crop selection (*“what to plant”*)** The first contribution proposes an interpretable and accurate system for crop selection. The proposed CS-AdaRF-SHAP system combines two key elements. First, an Adaptive Boosting of Random Forest (AdaRF) ensemble iteratively reweights misclassified instances to improve separation between crops with similar characteristics and to achieve stable predictive performance under different agricultural conditions. Second, SHapley Additive exPlanations (SHAP) provide both global and local interpretability by measuring the influence of each feature—such as soil nutrients, pH, temperature, humidity, and rainfall—on individual predictions. This allows farmers and agronomists to understand the reasoning behind each recommendation.

  Experiments showed that the CS-AdaRF-SHAP system reached a test accuracy of 99.77%, with precision, recall, and F1-score all close to 100%. The combination of strong predictive ability and transparent decision-making supports trust in AI-based agricultural decision systems and encourages their practical adoption.

- **Data-Driven Crop Yield Prediction (*"how much to expect"*)** Building on the crop selection framework, the second contribution focuses on yield prediction using tomato production in greenhouse conditions as a case study. The work relied on greenhouse data consisting of multivariate time-series measurements of key environmental variables.

  A stacked ensemble learning architecture was developed by combining Gradient Boosting Regressors, Random Forests, and Support Vector Regression within a meta-learner to capture nonlinear interactions and improve generalization. This design produced higher predictive accuracy than standard regression methods and enabled reliable daily yield forecasts, which support the planning of storage, labor, and marketing activities in controlled agricultural settings.

- **Blockchain-Based Approach to Securing Data in Smart Agriculture (*"How can the data that supports these decisions remain secure, reliable, and trustworthy?"*)** The third contribution addresses the protection and reliability of agricultural data by introducing a blockchain-based management framework. The proposed system integrates blockchain technology, smart contracts, edge computing, and the InterPlanetary File System (IPFS) into a unified architecture to guarantee data integrity and secure sharing among stakeholders. All transactional metadata, including information on data ownership, access permissions, and file hashes, is recorded on the blockchain through custom smart contracts, providing immutability, transparency, and auditable access for all participants.

7.2 Perspectives

Several improvements could be made to the work done in this thesis, and research directions that require further investigations in the future. We listed some of them in the items below:

- The current crop selection framework was trained on a balanced dataset covering 22 crop types with 7 features. Future work could incorporate larger and more diverse datasets that include additional crops, regional soil profiles, and seasonal variations. Integrating satellite imagery and remote sensing indices would further enrich the feature space and allow the system to operate effectively across different geographic scales.

- The yield prediction model focused on tomato production in greenhouse environments. Future studies could extend the same methodology to other crops or to open-field cultivation, where external factors such as weather fluctuations and pest outbreaks introduce additional uncertainty. Combining ensemble learning with deep learning architectures, such as recurrent or transformer-based models, may improve the capacity to capture long-term temporal dependencies and produce more accurate forecasts.

- Another perspective is to extend the blockchain framework toward advanced data analytics and decision automation. By combining blockchain with machine learning modules deployed at the edge, the system could support on-chain analysis of sensor data for tasks such as anomaly detection, quality assessment, and predictive maintenance. This integration would provide verifiable analytical results directly on the blockchain, strengthening trust among stakeholders while enabling faster and more autonomous agricultural operations.

- Another promising direction is the creation of user-friendly decision support systems that unify crop selection, yield prediction, and secure data management within a single integrated platform. Such a system could take the form of a mobile or web application that delivers predictions, explanatory analyses, and blockchain-based verification through an intuitive interface, enabling farmers, agronomists, and policy makers to access reliable information and adopt these technologies with minimal technical effort.

List of Publications

Journal Paper

1. Mancer, M’hamed, Labib Sadek Terrissa, and Soheyb Ayad. ”Interpretable Crop Selection for Optimized Farming Decisions.” International Journal of Computing and Digital Systems 17, no. 1 (2025): 1-14.

Book Chapters

1. Mancer, M’hamed, Labib Sadek Terrissa, Soheyb Ayad, Hamed Laouz, and Noureddine Zerhouni. ”Advancing Crop Recommendation Systems Through Ensemble Learning Techniques.” In The Proceedings of the International Conference on Smart City Applications, pp. 45-54. Cham: Springer Nature Switzerland, 2023.

2. Laouz, Hamed, Soheyb Ayad, Labib Sadek Terrissa, and M’hamed Mancer. ”Water Amount Prediction for Smart Irrigation Based on Machine Learning Techniques.” In The Proceedings of the International Conference on Smart City Applications, pp. 21-30. Cham: Springer Nature Switzerland, 2023.

3. Mancer, M’hamed, Labib Sadek Terrissa, and Soheyb Ayad. ”Machine Learning-Based Prediction of Tomato Yield in Greenhouse Environments.” In International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024), pp. 117-128. Atlantis Press, 2024.

Conference Papers

1. Mancer, M'hamed, Labib Sadek Terrissa, Soheyb Ayad, and Hamed Laouz. ”A Blockchain-based approach to securing data in smart agriculture.” In 2022 International Symposium on iNnovative Informatics of Biskra (ISNIB), pp. 1-5. IEEE, 2022.

2. Mancer, M'hamed, Khelili Mohamed Akram, Ezedin Barka, Kazar Okba, Slatnia Sihem, Saad Harous, Belkacem Athamena, and Zina Houhamdi. ”Blockchain technology for secure shared medical data.” In 2022 International Arab Conference on Information Technology (ACIT), pp. 1-6. IEEE, 2022.

3. Mancer, M'hamed, Labib Sadek Terrissa, Soheyb Ayad, and Hamed Laouz. ”MLP-powered smart application to enhance efficiency and productivity in Algerian agriculture.” NCAIA'2023: 80.

4. Mancer, M'hamed, Labib Sadek Terrissa, Soheyb Ayad, and Hamed Laouz. ”Tomato Crop Forecasting: A Comparative Analysis of Regression Models.” In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), pp. 649-653. IEEE, 2024.

Bibliography

[1] World Bank. Employment in agriculture (% of total employment), 2024. Accessed: 2025-06-10.

[2] World Bank. Agriculture, forestry, and fishing, value added (% of gdp), 2024. Accessed: 2025-06-10.

[3] DevelopmentEducation.ie. Is the world on track to achieve zero hunger by 2030? five takeaways from 2 key reports, 2020. Accessed: 2025-06-10.

[4] Virginia Tech. Global agricultural productivity index [image], 2019. Accessed: 2025-06-10.

[5] Facts & Factors Research. Global smart agriculture market share is likely to reach at a cagr value of around 9.30% by 2028, September 2022. Published September 19, 2022; Accessed: 2025-06-21.

[6] World Bank. World development indicators: Employment in agriculture (% of total employment). `https://databank.worldbank.org/source/world-development-indicators`, 2023. Accessed: 2025-05-28.

[7] Food and Agriculture Organization of the United Nations. The state of food and agriculture 2023. `https://www.fao.org/publications/sofa/2023/en/`, 2023. Accessed: 2025-05-01.

[8] Benjamin Kisliuk, Jan Christoph Krause, Hendrik Meemken, Juan Carlos Saborío Morales, Henning Müller, and Joachim Hertzberg. Ai in current and future agriculture: An introductory overview. *KI - Künstliche Intelligenz*, 37(1):1–14, 2023.

[9] Garima Gupta and Sudhir Kumar Pal. Applications of ai in precision agriculture. *Discover Agriculture*, 3(61):1–14, 2025.

[10] Ersin Elbasi, Nour Mostafa, Zakwan AlArnaout, Aymen I Zreikat, Elda Cina, Greeshma Varghese, Ahmed Shdefat, Ahmet E Topcu, Wiem Abdelbaki, Shinu Mathew, et al. Artificial intelligence technology in the agricultural sector: A systematic literature review. *IEEE access*, 11:171–202, 2022.

[11] Uwaga Monica Adanma and Emmanuel Olurotimi Ogunbiyi. A comparative review of global environmental policies for promoting sustainable development and economic growth. *International Journal of Applied Research in Social Sciences*, 6(5):954–977, 2024.

[12] Zhaoyu Zhai, José Fernán Martínez, Victoria Beltran, and Néstor Lucas Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.

[13] Rodrigue Kongne Nde, Jean Louis Ebongue Kedieng Fendji, Blaise Omer Yenke, and Julius Schöning. Crop selection: A survey on factors and techniques. *Smart Agricultural Technology*, 9:100602, 2024.

[14] Jeremy Whish, Lindsay Bell, and Peter de Voil. How resilient is your farming system strategy for the long haul? long term simulations of risk and sustainability of various farming systems experiments using apsim. *GRAINS RESEARCH UPDATE*, page 23, 2022.

[15] University of Minnesota Extension. Crop and field planning tools for vegetable farmers: Step 1 – decide what to grow. `https://extension.umn.edu/vegetable-growing-guides-farmers/crop-and-field-planning-tools-vegetable-farmers`. Accessed: 2025-09-24.

[16] TS Stombaugh, TG Mueller, SA Shearer, CR Dillon, and GT Henson. Guidelines for adopting precision agriculture practices. *Publ. No. PA-2. Lexington, Kentucky. University of Kentucky Cooperative Extension Service. Assistant Professor Biosystems and Agricultural Engineering University of Kentucky*, 128:40546–0276, 2001.

[17] Thi Ha Lien Le, Paul Kristiansen, Brenda Vo, Jonathan Moss, and Mitchell Welch. Understanding factors influencing farmers' crop choice and agricultural transfor-

mation in the upper vietnamese mekong delta. *Agricultural Systems*, 216(February):103899, 2024.

[18] Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3):1–42, 2017.

[19] Karamveer Singh Sidhu, Ramandeep Singh, Snehdeep Singh, and Gunjot Singh. Data science and analytics in agricultural development. *Environment Conservation Journal*, 22(SE):9–19, 2021.

[20] Christian Haertel, Matthias Pohl, Abdulrahman Nahhas, Daniel Staegemann, and Klaus Turowski. Toward a lifecycle for data science: A literature review of data science process models. In *PACIS 2022 Proceedings*, page 242, 2022.

[21] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. Big data preprocessing: methods and prospects. *Big data analytics*, 1:1–22, 2016.

[22] Suad A Alasadi and Wesam S Bhaya. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107, 2017.

[23] Yajie Ma, Jin Jin, Qihui Huang, and Feng Dan. Data preprocessing of agricultural iot based on time series analysis. In *Intelligent Computing Theories and Application: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part I 14*, pages 219–230. Springer, 2018.

[24] C Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 1999.

[25] MIT Critical Data, Matthieu Komorowski, Dominic C Marshall, Justin D Salciccioli, and Yves Crutain. Exploratory data analysis. *Secondary analysis of electronic health records*, pages 185–203, 2016.

[26] John Wilder Tukey et al. *Exploratory data analysis*, volume 2. Springer, 1977.

[27] Yingsen Mao et al. *Data visualization in exploratory data analysis: An overview of methods and technologies*. PhD thesis, 2015.

[28] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018.

[29] Vishal Meshram, Kailas Patil, Vidula Meshram, Dinesh Hanchate, and S.D. Ramkteke. Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, 1:100010, 2021.

[30] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2010.

[31] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[32] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[33] Frank Emmert-Streib and Matthias Dehmer. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1439, 2022.

[34] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021.

[35] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

[36] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[37] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[38] Daniel N Moriasi, Jeffrey G Arnold, Mark W Van Liew, Ronald L Bingner, R Daren Harmel, and Tamie L Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.

[39] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[40] Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2019.

[41] Tomohiro Ohgushi et al. Anomaly detection for agricultural vehicles using autoencoders. *Sensors*, 22(10):3608, 2022.

[42] Ebenezer Olaniyi, Dong Chen, Yuzhen Lu, and Yanbo Huang. Generative adversarial networks for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 196:106892, 2022.

[43] Zhen Li et al. Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems. *Plants*, 12(5):972, 2023.

[44] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[45] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[46] Ozlem Turgut, Ibrahim Kok, and Suat Ozdemir. Agroxai: Explainable ai-driven crop recommendation system for agriculture 4.0. *arXiv preprint arXiv:2412.16196*, 2024.

[47] Christoph Molnar. *Interpretable Machine Learning.* Lulu. com, 2020.

[48] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404, 2021.

[49] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[50] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[51] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[53] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[54] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[55] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[56] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):841–887, 2017.

[57] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[58] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of shap explanations. *Journal of Artificial Intelligence Research*, 74:851–886, 2022.

[59] Ana Victoria Ponce-Bobadilla, Vanessa Schmitt, Corinna S Maier, Sven Mensing, and Sven Stodtmann. Practical guide to shap analysis: explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17(11):e70056, 2024.

[60] Zibin Zheng, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang. An overview of blockchain technology: Architecture, consensus, and future trends. In *2017 IEEE international congress on big data (BigData congress)*, pages 557–564. Ieee, 2017.

[61] Ghassan Karame and Srdjan Capkun. Blockchain security and privacy. *IEEE Security & Privacy*, 16(04):11–12, 2018.

[62] World Bank. Agriculture and food: Overview. `https://www.worldbank.org/en/topic/agriculture/overview`, 2023. Accessed: 2025-05-01.

[63] World Bank. Agricultural employment by region: Africa and south asia (2020–2025). `https://databank.worldbank.org/source/world-development-indicators`, 2023. Accessed: 2025-05-28.

[64] Food and Agriculture Organization of the United Nations. The state of food and agriculture 2023: Realizing the potential of agriculture for poverty reduction. `https://www.fao.org/publications/sofa/2023/en/`, 2023. Accessed: 2025-05-28.

[65] Food and Agriculture Organization of the United Nations. Agricultural resilience in the face of covid-19. `https://www.fao.org/documents/card/en/c/CB3181EN`, 2021. Accessed: 2025-05-28.

[66] Concern Worldwide and Welthungerhilfe. 2020 global hunger index: One decade to zero hunger. `https://www.globalhungerindex.org/`, 2020. Accessed: 2025-05-28.

[67] R. et al. Mendoza. The role of soils in sustainability, climate change, and biodiversity. *Environments*, 4(3):36, 2024.

[68] A. Sánchez and D. Gómez. Sustainable development strategies and good agricultural practices: A focus on agroforestry. *Sustainability*, 16(22):9878, 2024.

[69] M. et al. Lefebvre. Shaping the landscape: Agricultural policies and local biodiversity effects in the eu. *Land Use Policy*, 26(3):545–552, 2009.

[70] Rattan Lal. *Ecosystem Services Linked to Soil Carbon in Forest and Agricultural Ecosystems*. Elsevier, 2018.

[71] N. H. Batjes. Soil carbon storage. https://www.nature.com/scitable/knowledge/library/soil-carbon-storage-84223790/, 2020. Accessed 2025-05-28.

[72] M. Qaim et al. Sustainable transformation of agriculture requires addressing productivity gaps. *Heliyon*, 9(6):e084232, 2024.

[73] Per Pinstrup-Andersen and Jessica Fanzo. Connecting the food and agriculture sector to nutrition outcomes: a global review. *Food Security*, 15(1):45–62, 2022.

[74] Roseline Remans, Jessica Fanzo, and Cheryl Palm. Measuring nutritional quality of agricultural production systems: Concepts, indicators, and frameworks. *Global Food Security*, 1(3):180–194, 2017.

[75] Impacts of global climate change on agricultural production. *MDPI Agronomy*, 2024.

[76] Climate change impacts on agriculture and food supply. https://www.epa.gov/climateimpacts/climate-change-impacts-agriculture-and-food-supply, 2024.

[77] Assessing the impact of climate change on agricultural productivity. *ScienceDirect*, 2024.

[78] The impact of extreme weather events as a consequence of climate change. *ScienceDirect*, 2023.

[79] What is the impact of intensive agriculture on land degradation? https://www.thehappyturtlestraw.com/what-is-the-impact-of-intensive-agriculture-on-land-degradation/, 2024.

[80] Food and Agriculture Organization of the United Nations. General and food consumer price indices inflation rates, 2023. Accessed: 2025-05-28.

[81] Food and Agriculture Organization of the United Nations. Strengthening smallholder producers' skills and market access, 2023. Accessed: 2025-05-28.

[82] Food and Agriculture Organization of the United Nations. Digital technologies in agriculture and rural areas: Status report, 2022. Accessed: 2025-05-28.

[83] Food and Agriculture Organization of the United Nations. What factors shape small-scale farmers' and firms' adoption of new technologies?, 2023. Accessed: 2025-05-28.

[84] R. K. Srivastava, R. Singh, and S. Tripathi. Traditional agriculture: a climate-smart approach for sustainable food production. *Agricultural Research*, 6(4):383–396, 2017.

[85] Ken E. Giller, Thomas Delaune, João Vasco Silva, Mark van Wijk, James Hammond, Katrien Descheemaeker, Gerrie van de Ven, Antonius G. T. Schut, Godfrey Taulya, Regis Chikowo, and Jens A. Andersson. Small farms and development in sub-saharan africa: Farming for food, for income or for lack of better options? *Food Security*, 13:1431–1454, 2021.

[86] Andrea Albanese, Matteo Nardello, and Davide Brunelli. Automated pest detection with dnn on the edge for precision agriculture. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11(3):458–467, 2021.

[87] Giulia Chiaraluce, Deborah Bentivoglio, Adele Finco, Mariantonietta Fiore, Francesco Contò, and Antonino Galati. Exploring the role of blockchain technology in modern high-value food supply chains: Global trends and future research directions. *Agricultural and Food Economics*, 12(1):6, 2024.

[88] Fatima Zahra Bassine, Terence Epule Epule, Ayoub Kechchour, and Abdelghani Chehbouni. Recent applications of machine learning, remote sensing, and iot approaches in yield prediction: a critical review. *arXiv preprint arXiv:2306.04566*, 2023.

[89] Leonidas Droukas, Zoe Doulgeri, Nikolaos L Tsakiridis, Dimitra Triantafyllou, Ioannis Kleitsiotis, Ioannis Mariolis, Dimitrios Giakoumis, Dimitrios Tzovaras, Dimitrios

Kateris, and Dionysis Bochtis. A survey of robotic harvesting systems and enabling technologies. *Journal of Intelligent & Robotic Systems*, 107(2):21, 2023.

[90] Michael Robertson, Peter Carberry, and Lisa Brennan. The economic benefits of precision agriculture: case studies from australian grain farms. *Crop Pasture Sci*, 60:2012, 2007.

[91] Nicoleta Tantalaki, Stavros Souravlas, and Manos Roumeliotis. Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of agricultural & food information*, 20(4):344–380, 2019.

[92] George Papadopoulos, Simone Arduini, Havva Uyar, Vasilis Psiroukis, Aikaterini Kasimati, and Spyros Fountas. Economic and environmental benefits of digital agricultural technologies in crop production: A review. *Smart Agricultural Technology*, page 100441, 2024.

[93] Nazish Aijaz, He Lan, Tausif Raza, Muhammad Yaqub, Rashid Iqbal, and Muhammad Salman Pathan. Artificial intelligence in agriculture: advancing crop productivity and sustainability. *Journal of Agriculture and Food Research*, page 101762, 2025.

[94] Batool Alsowaiq, Noura Almusaynid, Esra Albhnasawi, Wadha Alfenais, and Suresh Sankrayananarayanan. Crop recommendation assessment for arid land using machine learning. In *International Conference on ICT for Sustainable Development*, pages 323–332. Springer, 2023.

[95] Sudarshan Reddy Palle and Shital A Raut. Crops recommendation system model using weather attributes, soil properties, and crops prices. In *Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022*, pages 323–338. Springer, 2023.

[96] Surekha Janrao and Deven Shah. Return on investment framework for profitable crop recommendation system by using optimized multilayer perceptron regressor. *IAES International Journal of Artificial Intelligence*, 11(3):969, 2022.

[97] Raswitha Bandi, M Sai Surya Likhit, S Rajavardhan Reddy, Sathwik Raj Bodla, and Vempati Sai Venkat. Voting classifier-based crop recommendation. *SN Computer Science*, 4(5):516, 2023.

[98] Punith Kumar, H Varun Prabhu, and HN Champa. Crop recommendation using ensemble stacking machine learning approach. In *2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon)*, pages 1–6. IEEE, 2023.

[99] Behnaz Motamedi and Balázs Villányi. A predictive analytics model with bayesian-optimized ensemble decision trees for enhanced crop recommendation. *Decision Analytics Journal*, 12:100516, 2024.

[100] Sally Elghamrawy, Athanasios V Vasilakos, Ashraf Darwish, and Aboul Ella Hassanien. An intelligent crop recommendation model for the three strategic crops in egypt based on climate change data. In *The Power of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations*, pages 189–205. Springer, 2023.

[101] Sita Rani, Amit Kumar Mishra, Aman Kataria, Saurav Mallik, and Hong Qin. Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, 13(1):15997, 2023.

[102] Kalaiselvi Bakthavatchalam, Balaguru Karthik, Vijayan Thiruvengadam, Sriram Muthal, Deepa Jose, Ketan Kotecha, and Vijayakumar Varadarajan. Iot framework for measurement and precision agriculture: predicting the crop using machine learning algorithms. *Technologies*, 10(1):13, 2022.

[103] Alonica Villanueva, Asmadi Dorado, Erika Marie Gerarman, John Brian F Quebral, Maria Cecilia Venal, Daryl Neil L Valenzuela, and Menchie Rosales. Real-time best-fitted crops recommendation system based on agricultural soil health. In *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, pages 1–6. IEEE, 2022.

[104] Mohamed Omar Abdullahi, Abdukadir Dahir Jimale, Yahye Abukar Ahmed, and Abdulaziz Yasin Nageye. Revolutionizing somali agriculture: harnessing machine learning and iot for optimal crop recommendations. *Discover Applied Sciences*, 6(3):77, 2024.

[105] Anitha Palakshappa, Sowmya Kyathanahalli Nanjappa, Punitha Mahadevappa, and Sinchana Sinchana. Smart irrigation with crop recommendation using ma-

chine learning approach. *Bulletin of Electrical Engineering and Informatics*, 13(3):1952–1960, 2024.

[106] S Kiruthika and D Karthika. Iot-based professional crop recommendation system using a weight-based long-term memory approach. *Measurement: Sensors*, 27:100722, 2023.

[107] Yashashree Mahale, Nida Khan, Kunal Kulkarni, Shivali Amit Wagle, Preksha Pareek, Ketan Kotecha, Tanupriya Choudhury, and Ashutosh Sharma. Crop recommendation and forecasting system for maharashtra using machine learning with lstm: a novel expectation-maximization technique. *Discover Sustainability*, 5(1):134, 2024.

[108] Christine Musanase, Anthony Vodacek, Damien Hanyurwimfura, Alfred Uwitonze, and Innocent Kabandana. Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices. *Agriculture*, 13(11):2141, 2023.

[109] Tomato land & water. `https://www.fao.org/land-water/databases-and-software/crop-information/tomato/en/`. Accessed 07-Jul-2023.

[110] M. Mancer, L. Terrissa, S. Ayad, and H. Laouz. Tomato crop forecasting: A comparative analysis of regression models. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, pages 649–653, 2024.

[111] P. Muruganantham, S. Wibowo, S. Grandhi, N. Samrat, and N. Islam. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14:1990, 2022.

[112] M. Mancer, K. Akram, E. Barka, K. Okba, S. Sihem, S. Harous, B. Athamena, and Z. Houhamdi. Blockchain technology for secure shared medical data. In *2022 International Arab Conference on Information Technology (ACIT)*, pages 1–6, 2022.

[113] M. Mancer, L. Terrissa, S. Ayad, and H. Laouz. A blockchain-based approach to

securing data in smart agriculture. In *2022 International Symposium on INnovative Informatics of Biskra (ISNIB)*, pages 1–5, 2022.

[114] A. Raghuvanshi, U. Singh, G. Sajja, H. Pallathadka, E. Asenso, M. Kamal, A. Singh, and K. Phasinam. Intrusion detection using machine learning for risk mitigation in iot-enabled smart irrigation in smart farming. *Journal of Food Quality*, pages 1–8, 2022.

[115] M. Mancer, L. Terrissa, S. Ayad, H. Laouz, and N. Zerhouni. Advancing crop recommendation systems through ensemble learning techniques. In *The Proceedings of the International Conference on Smart City Applications*, pages 45–54, 2023.

[116] Atharva Ingle. Crop recommendation dataset. `https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset`, December 2020. Kaggle.

[117] S. Hemming, H. Zwart, A. Elings, A. Petropoulou, and I. Righini. Autonomous greenhouse challenge, second edition (2019). `https://data.4tu.nl/articles/_/12764777/2`, 2020. 4TU.ResearchData.

[118] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[119] Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.