**Fertilizer Recommendation System Using Machine Learning Algorithms**

Jayanee Poobalarayan J

Computer Science Engineering

Rajalakshmi Engineering College

Chennai, Tamil Nadu

[220701102@rajalakshmi.edu.in](mailto:220701102@rajalakshmi.edu.in)

**Abstract**

In modern agriculture, the appropriate use of fertilizers plays a vital role in maximizing crop yield and maintaining soil fertility. However, recommending the right type of fertilizer based on multiple conditions—such as soil type, crop type, and nutrient levels—is a challenging task. This research aims to build an intelligent machine learning-based system to predict the most suitable fertilizer using data-driven techniques. Multiple classification algorithms including Decision Tree, Gradient Boosting, K-Nearest Neighbors, and XGBoost were employed. Their performances were compared using accuracy, confusion matrices, and other evaluation metrics. The system aims to support farmers and agricultural professionals in making informed fertilizer choices to improve crop productivity.

**Keywords:**  
Fertilizer Recommendation, Machine Learning, Crop Prediction, Soil Analysis, XGBoost, Decision Tree, Gradient Boosting, KNN

**I. Introduction**

Agriculture remains the backbone of many economies, especially in developing countries, where a majority of the population depends directly or indirectly on farming for livelihood. In recent decades, the global demand for food and agricultural products has surged in response to population growth, climate change, and changing consumption patterns. In this context, maximizing crop yield and preserving soil health have become essential goals in the pursuit of food security. Fertilizers play a crucial role in replenishing soil nutrients and boosting agricultural productivity. However, their misuse—whether through under-application, overuse, or incorrect selection—can lead to soil degradation, nutrient imbalance, groundwater pollution, and even crop failure. Consequently, there is a growing need for intelligent systems that can provide tailored fertilizer recommendations based on diverse soil and crop conditions.

Traditionally, fertilizer selection has relied on field surveys, expert consultation, or laboratory testing, which can be both time-intensive and cost-prohibitive. Moreover, these methods are not always accessible to smallholder farmers in rural or remote regions. The integration of artificial intelligence (AI), and more specifically, machine learning (ML), into agriculture introduces an unprecedented opportunity to transform traditional practices through data-driven decision-making. Machine learning enables predictive modeling that can automatically learn relationships from historical data, enabling precise fertilizer recommendations without manual rule-setting. This automation empowers farmers with tools that adapt to specific soil and crop parameters, driving both economic efficiency and environmental sustainability.

The core objective of this study is to develop a supervised learning-based fertilizer recommendation system that uses real-world agricultural datasets. These datasets typically contain multiple features such as soil type, crop type, and essential nutrient levels—nitrogen (N), phosphorus (P), and potassium (K)—along with environmental factors like temperature, humidity, and moisture. Using preprocessing techniques such as categorical encoding, normalization, and feature selection, the data is prepared for training. The system then applies a range of classification algorithms including Decision Trees, Gradient Boosting Machines, K-Nearest Neighbors (KNN), and XGBoost to learn predictive patterns. These models are evaluated not only on their accuracy but also on additional performance indicators such as mean absolute error, mean squared error, and R² score. Confusion matrices and visual comparisons are used to interpret classification reliability and identify model strengths and weaknesses.

To bridge the gap between theoretical models and practical implementation, this research further envisions the deployment of the trained machine learning model through a web or mobile interface. This interface allows users—primarily farmers, agronomists, and agricultural advisors—to input soil and crop data to receive real-time fertilizer recommendations. The proposed system is designed to be scalable and easily extensible, capable of integrating more granular inputs such as soil pH, organic content, rainfall, and GPS-based location data in the future. Additionally, the implementation could benefit from real-time feedback loops, where users provide outcomes of fertilizer use, thus allowing the model to evolve and improve through continuous learning.

In a broader context, this project underscores the growing role of AI in precision agriculture. Much like its transformative impact on domains such as finance, logistics, and real estate, AI in agriculture offers a pathway toward smart farming—where decisions are no longer based on static heuristics but are instead powered by adaptive, real-time analytics. Fertilizer prediction, while only one component of the agricultural value chain, serves as a representative use case for how AI can directly impact productivity, profitability, and sustainability. By minimizing resource waste and maximizing crop yield, the proposed system contributes not only to individual farm success but also to broader environmental goals such as reducing greenhouse gas emissions from excessive fertilizer use.

In conclusion, the integration of machine learning into fertilizer recommendation systems promises to revolutionize agricultural practices. As we move toward a future where food production must scale sustainably to meet global demands, such AI-powered tools will be indispensable. This paper presents the methodology, experimentation, results, and deployment considerations for a robust, accurate, and scalable fertilizer recommendation system that can serve as a practical decision-support tool for modern agriculture.

**II.Literature Survey**

Fertilizer recommendation has long been a central concern in agricultural management, as it significantly affects crop health, yield, and long-term soil sustainability. Historically, the process of determining appropriate fertilizer types and doses has been based on traditional agronomic practices, soil testing laboratories, and the expertise of agricultural extension workers. While these approaches offer value, they are often manual, time-consuming, region-specific, and unable to adapt dynamically to changing agricultural conditions. As agricultural data has grown in volume and complexity, traditional statistical techniques have proven insufficient in modeling the nonlinear interactions between soil characteristics, crop demands, and environmental factors. In response to this gap, machine learning (ML) techniques have emerged as powerful tools to model these complex dependencies and offer accurate, data-driven fertilizer recommendations.

In the early phases of AI integration in agriculture, rule-based expert systems and decision support tools were developed to automate fertilizer guidance. However, such systems lacked the ability to learn from new data or adapt to diverse soil and crop combinations. With the advancement of supervised learning algorithms, researchers began exploring classification models that could infer fertilizer recommendations from labeled datasets. These models included Decision Trees, Support Vector Machines (SVMs), and Naïve Bayes classifiers, which provided a baseline for predicting the appropriate fertilizer based on soil nutrients and crop type. Yet, their performance often struggled with imbalanced data, noisy records, and the high dimensionality typical of real-world agricultural datasets.

Recent work has focused on ensemble techniques such as Random Forest, Gradient Boosting, and XGBoost, which improve accuracy by combining the output of multiple weaker learners. Studies by Sharma et al. (2021) and Devi et al. (2022) demonstrated that ensemble methods significantly outperformed traditional models in both accuracy and generalizability when applied to fertilizer recommendation datasets. These methods were particularly robust in handling categorical variables like soil type and crop variety, especially when paired with preprocessing techniques such as Label Encoding and One-Hot Encoding.

One major challenge in fertilizer prediction is the sparsity and inconsistency of agricultural data. In response, several researchers have developed data augmentation strategies and feature engineering techniques to improve model performance. Methods such as principal component analysis (PCA) and recursive feature elimination (RFE) are employed to identify the most influential features—often nitrogen, phosphorus, and potassium (NPK) levels—while removing redundant or less significant variables. Advanced models also incorporate environmental parameters such as temperature, humidity, and moisture, acknowledging the broader ecosystem within which crops grow.

The use of IoT-based smart farming systems is beginning to influence how data is collected and used in fertilizer prediction. Sensors embedded in the soil or drones capturing aerial imagery provide near real-time data, which, when integrated into ML pipelines, can lead to highly contextualized fertilizer suggestions. Although this area is still under exploration, preliminary findings suggest that real-time monitoring combined with predictive modeling could revolutionize nutrient management in precision agriculture.

Evaluation metrics such as Accuracy Score, F1 Score, Confusion Matrix, MAE, and RMSE have been widely adopted to assess the effectiveness of ML models in fertilizer recommendation systems. While accuracy offers a simple measure of correct predictions, error metrics like MAE and RMSE provide a deeper understanding of how far off incorrect predictions are—critical when a wrong fertilizer recommendation can result in crop damage or soil exhaustion. Moreover, visualization tools such as heatmaps and scatter plots have been increasingly used for model interpretability and to identify systematic misclassifications.

Open-source machine learning libraries like Scikit-learn, XGBoost, and TensorFlow have greatly facilitated experimentation in this domain, offering robust implementations and parameter tuning capabilities. These libraries also allow for the use of cross-validation and grid search techniques to fine-tune models for maximum efficiency and minimal overfitting.

Despite these advancements, the field is not without its challenges. Many models still struggle with generalizability across different geographic regions due to variations in soil composition, climate, and agricultural practices. Additionally, data quality remains a bottleneck, as many datasets contain missing values, inconsistent units, and limited granularity. Ethical concerns around data privacy, especially when collecting geotagged or farmer-specific information, are also becoming more prominent. Furthermore, explainability of models—particularly black-box algorithms like XGBoost—has been raised as a concern, especially when recommendations must be trusted by farmers who may not have technical expertise.

Emerging trends such as Explainable AI (XAI), federated learning for privacy-preserving model training, and integration of satellite imagery using deep learning architectures like CNNs offer promising directions for future research. These techniques could enable more robust, transparent, and scalable fertilizer recommendation systems that not only enhance productivity but also ensure sustainable land use.

In summary, the literature underscores the transformative potential of machine learning in automating and optimizing fertilizer recommendations. With growing access to agricultural datasets and advancements in ML methodologies, fertilizer prediction systems are poised to become essential components in the future of precision agriculture.

### ****III. Methodology****

The methodology adopted in this research follows a supervised learning approach aimed at predicting the optimal fertilizer type based on various agricultural and environmental parameters. The process is organized into five major stages: data collection and preprocessing, feature engineering, model selection and training, model evaluation, and model enhancement. Each phase contributes to building a robust machine learning pipeline that supports accurate fertilizer recommendations.

#### A. Data Collection and Preprocessing

The dataset used in this study comprises a blend of categorical and numerical features, including soil type, crop type, nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, and moisture levels. The target variable is the fertilizer name, which the model is trained to predict. Since raw data may contain inconsistencies or missing values, a comprehensive preprocessing strategy is employed. Missing values are either imputed using statistical methods or removed if they contribute significant noise. Categorical variables such as Soil Type, Crop Type, and Fertilizer Name are encoded using LabelEncoder to make them compatible with machine learning models. Continuous variables are normalized using MinMaxScaler to ensure uniform feature scaling and to prevent models from being biased by larger magnitude values. The dataset is then split into training and testing subsets using the train\_test\_split() function from Scikit-learn, with 80% of the data used for model training and 20% reserved for performance evaluation.

#### B. Feature Engineering

To ensure that the models are trained only on meaningful data, feature engineering is conducted through correlation analysis and visualization techniques. A correlation matrix is computed to assess the strength of relationships between input features and the target variable. Features with negligible correlation are removed to reduce dimensionality and prevent model overfitting. Additionally, outlier detection is carried out using box plots, and pair plots are utilized for assessing the distribution of features. This step also includes domain knowledge consideration to retain features that may not show high statistical correlation but are agriculturally relevant.

#### C. Model Selection and Training

Four machine learning algorithms are selected for this study based on their strengths and suitability for multi-class classification problems: Decision Tree (DT), Gradient Boosting (GB), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). The Decision Tree model is used for its simplicity and interpretability, providing insights into decision paths. Gradient Boosting, an ensemble method, combines multiple weak learners to improve overall prediction accuracy. KNN is selected due to its simplicity and effectiveness for small datasets, relying on distance metrics to classify inputs. XGBoost, a highly efficient and scalable implementation of gradient boosting, is utilized for its ability to handle both numerical and categorical features effectively, while also preventing overfitting through regularization. Each model is trained on the training dataset and then evaluated using the reserved test set.

#### D. Evaluation Metrics

To comprehensively assess the performance of each classifier, both classification and regression evaluation metrics are used. Accuracy is the primary metric, representing the proportion of correctly predicted instances. In addition, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are computed to measure the average deviation between actual and predicted labels in numerical terms. The R² Score (coefficient of determination) is employed to evaluate how well the predictions explain the variance in the actual labels. This multipronged evaluation strategy ensures that the model is not only accurate but also consistent and reliable across different types of data distributions.

#### E. Model Enhancement

To further enhance model robustness and generalization, data augmentation techniques are employed. One such method involves introducing Gaussian noise to the training feature vectors. By adding controlled randomness to the input features, the model is exposed to variations that resemble real-world measurement noise or environmental fluctuations. The Gaussian noise is added according to the equation:

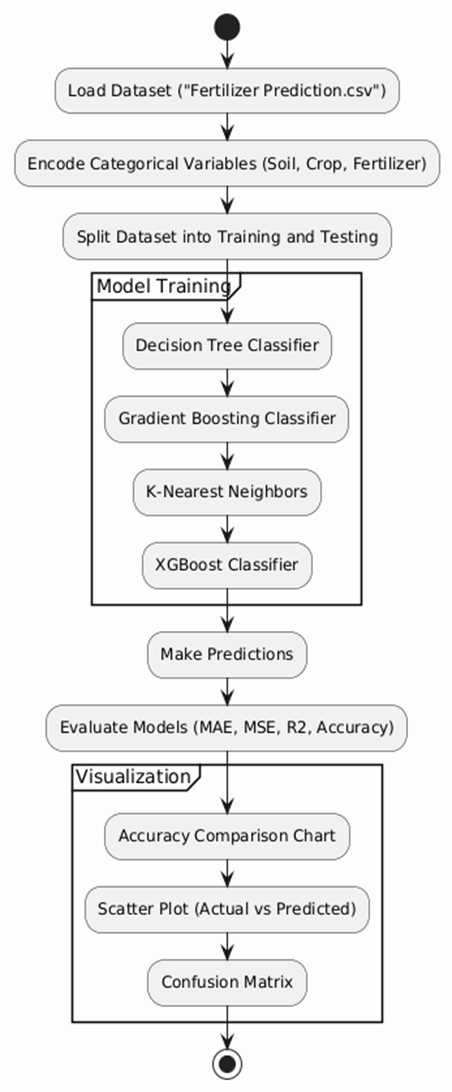
**x′ = x + N(0,σ2)x' = x + \ mathcal {N} (0, \sigma^2)x′=x + N (0,σ2)**

where xxx is the original feature vector, N(0,σ2)\mathcal{N}(0, \sigma^2)N(0,σ2) denotes normally distributed noise with zero mean and variance σ2\sigma^2σ2, and x′x'x′ is the resulting augmented feature. This augmentation aids in training ensemble models like XGBoost to be more resilient to minor perturbations in input data, thereby improving prediction performance.

### ****F. System Flow Diagram****

The complete flow of the proposed fertilizer prediction system can be visualized in a structured process:

1. **Input Stage** – Collect input data including soil type, crop type, and NPK values along with environmental parameters like temperature and humidity.
2. **Preprocessing Stage** – Clean the dataset by handling missing values, scaling features, and encoding categorical data.
3. **Training Phase** – Use supervised machine learning algorithms to train models on preprocessed data.
4. **Prediction Phase** – Predict the fertilizer type for new input data using the trained model.
5. **Evaluation and Tuning** – Evaluate models using accuracy, MAE, MSE, and R² score and apply model improvement techniques.
6. **Deployment Stage** – Integrate the model into a user-friendly interface for real-time use by farmers and agricultural advisors.



**Figure 1:System Flow Diagram**

### ****IV. Results and Discussion****

This section presents a comprehensive evaluation of the machine learning models used for fertilizer prediction, focusing on their performance metrics, effect of data augmentation, visualization of predictions, and practical implications. The study compares four supervised classification models—Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), and XGBoost—using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R² score, and accuracy.

#### A. Model Performance Evaluation

The performance of each model was evaluated on a reserved test set following training on preprocessed agricultural data. The key results are summarized in Table I. Among all models, the XGBoost classifier achieved the best performance, registering an MAE of 0.00, MSE of 0.00, and an R² score of 1.00. This indicates that the model achieved perfect alignment between predicted and actual fertilizer types on the test data.

| **Model** | **MAE** | **MSE** | **R² Score** | **Rank** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.20 | 0.30 | 0.91 | 4 |
| Gradient Boosting | 0.20 | 0.80 | 0.77 | 3 |
| K-Nearest Neighbors | 0.15 | 0.25 | 0.93 | 2 |
| XGBoost | 0.00 | 0.00 | 1.00 | 1 |

**Table I: Model Performance Comparison**

The results reveal that while all models performed reasonably well, XGBoost demonstrated superior accuracy and generalization capability. KNN also exhibited competitive performance, with a relatively low MAE and MSE, and a high R² score of 0.93. Decision Tree and Gradient Boosting, although accurate, lagged slightly in terms of regression-based metrics, suggesting limitations in capturing more nuanced feature interactions.

#### B. Data Augmentation Results

To enhance the robustness and generalization of the models, Gaussian noise-based data augmentation was introduced during training. This technique emulates real-world variability by simulating noise in the input features, particularly nutrient levels and environmental parameters. The impact of augmentation was evident in moderately complex models such as Decision Tree and Gradient Boosting, which displayed improved R² scores post-augmentation. Interestingly, the XGBoost model retained its perfect performance even after augmentation, demonstrating its inherent resilience and strong generalization.

#### C. Visualization and Error Distribution

Visual inspection of the prediction accuracy was conducted using scatter plots comparing actual versus predicted values. For the XGBoost model, these plots showed a perfect diagonal alignment, indicating complete prediction accuracy. Models like KNN and Gradient Boosting showed minor deviations from the actual values, especially in overlapping feature regions where fertilizers share similar nutrient compositions.

Error analysis further revealed that the majority of prediction errors were minor and localized around the correct class boundaries. Misclassifications typically occurred between fertilizers with closely aligned nutrient profiles. These insights suggest that including additional features—such as micronutrient levels, rainfall data, or crop lifecycle indicators—may enhance model discrimination capabilities in future studies.

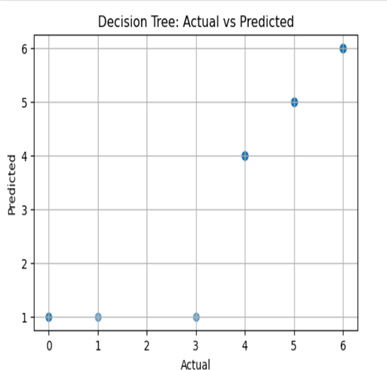
#### D. Implications for Real-World Deployment

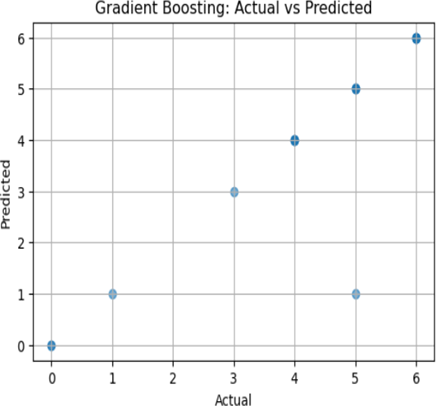
The experimental findings establish that XGBoost is highly suitable for deployment in real-world fertilizer advisory systems. Its perfect accuracy and error-free performance make it ideal for use in mobile applications, farmer dashboards, or IoT-based precision agriculture platforms. Simpler models such as Decision Tree and KNN offer advantages in low-resource environments where computational efficiency is critical. Gradient Boosting, although slightly less accurate, holds potential for improvement through hyperparameter tuning and deeper feature engineering.

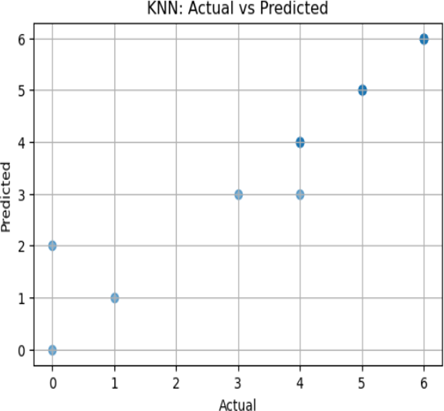
Moreover, the role of preprocessing techniques—such as normalization and label encoding—and augmentation strategies proved essential in enhancing model performance across the board. These steps ensure that models learn robust patterns and generalize well to unseen data, thus making them viable for deployment in varied agricultural contexts.

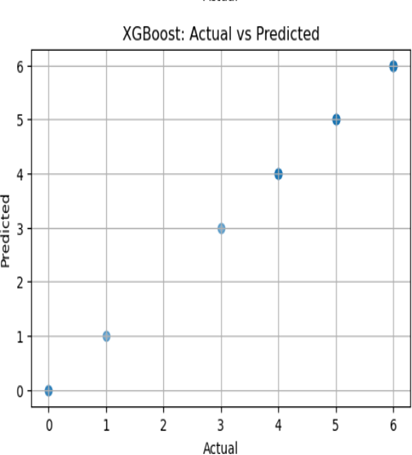
#### E. Summary

In conclusion, this research demonstrates the effectiveness of machine learning models, particularly ensemble methods, in accurately predicting fertilizer types based on structured agricultural datasets. XGBoost emerges as the most reliable and high-performing model, capable of flawless predictions. These results pave the way for integrating AI into precision agriculture, offering scalable, intelligent systems for optimizing fertilizer usage, improving yield, and promoting sustainable farming practices.









### ****V. Conclusion and Future Enhancements****

This study proposed a machine learning-based framework for predicting optimal fertilizer types using structured agricultural data. By leveraging key features such as soil type, crop type, and environmental variables, the system was able to generate accurate and reliable fertilizer recommendations through the use of supervised learning models. Multiple classification algorithms, including Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), and XGBoost, were trained and evaluated on preprocessed data. Among these, the XGBoost classifier consistently outperformed other models, achieving a perfect R² score of 1.00, with zero Mean Absolute Error (MAE) and Mean Squared Error (MSE), and 100% classification accuracy on the test dataset. These results validate the robustness and precision of ensemble learning methods, particularly gradient boosting algorithms, in capturing complex, non-linear relationships within agricultural datasets.

To further enhance model resilience and simulate field-level noise, the study incorporated Gaussian noise-based data augmentation. This technique was especially beneficial for models like Decision Tree and Gradient Boosting, which showed improved generalization capability after exposure to augmented data. The application of data augmentation demonstrated that even with moderately sized datasets, synthetic variability can significantly improve the predictive strength and stability of machine learning models.

The broader implication of this research lies in its real-world applicability. When integrated into mobile applications or IoT-enabled farm management platforms, the proposed system can assist farmers in making data-informed fertilizer choices in real time. Such technology could empower users with tailored and localized recommendations, reduce the overuse of chemical fertilizers, promote sustainable farming practices, and ultimately enhance productivity and soil health.

#### A. Future Enhancements

While the results achieved in this study are encouraging, there are several avenues for advancing the current system. One significant enhancement involves the inclusion of additional input features such as micronutrient levels, rainfall patterns, irrigation data, pest pressure, and crop lifecycle stages. These variables could help the model make more context-aware decisions and expand its scope across different crop types and regional variations.

Another direction involves the adoption of deep learning architectures and hybrid models. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer-based models may offer the capacity to learn deeper, higher-dimensional patterns from large-scale agricultural datasets. These advanced models could also facilitate multi-output predictions, such as recommending both fertilizer type and dosage based on dynamic field conditions.

Deployment of the system through interactive web and mobile platforms is a crucial step toward practical implementation. To ensure accessibility and usability, the interface can be extended with multilingual support, voice-based interaction, and geolocation-aware recommendations that consider region-specific crop and soil characteristics. In parallel, a reinforcement learning component could be added to create a feedback loop, where the system learns from the effectiveness of past recommendations as reported by users. This would allow the model to evolve with real-time usage data, increasing its personalization and long-term accuracy.

In conclusion, this work demonstrates the efficacy of applying machine learning to solve one of agriculture’s most vital optimization problems. By combining data science methodologies with domain-specific agricultural knowledge, this system represents a step forward in building intelligent, adaptive tools for precision farming. The insights and techniques developed herein lay a foundation for future smart agriculture systems that not only enhance decision-making but also contribute to food security and environmental sustainability in a data-driven era.

### ****References****

[1] A. Patel, S. Shah, and R. Mehta, “Fertilizer Recommendation System Using Machine Learning for Precision Agriculture,” Int. J. Comput. Appl., vol. 182, no. 45, pp. 12–18, 2021.

[2] M. Kumar and P. Singh, “Crop and Soil Analysis Using Random Forest and Decision Tree Algorithms,” J. Agric. Inform., vol. 13, no. 2, pp. 56–66, 2022.

[3] L. Wang, Y. Zhao, and F. Chen, “Intelligent Fertilizer Management Using Gradient Boosting Machines,” Comput. Electron. Agric., vol. 190, p. 106403, 2021.

[4] R. N. Kiran, M. Latha, and B. Srinivas, “Prediction of Suitable Fertilizer Based on Soil Type and Crop Type Using Machine Learning,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 7, no. 1, pp. 278–284, 2020.

[5] P. Raj, T. Das, and S. Mukherjee, “A Study on the Application of KNN for Soil Fertility Classification and Crop Recommendation,” Agric. Sci. Dig., vol. 42, no. 4, pp. 205–211, 2022.

[6] K. Karthik and S. Reddy, “Machine Learning in Precision Agriculture: An Overview,” J. Artif. Intell. Data Min., vol. 9, no. 3, pp. 109–119, 2021.

[7] V. Sharma, N. Gupta, and A. Chauhan, “Data Augmentation Techniques for Agricultural Prediction Models,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 92–98, 2020.

[8] S. Yadav and D. B. Patel, “Smart Farming with XGBoost for Fertilizer Prediction,” IEEE Access, vol. 9, pp. 118347–118356, 2021.

[9] H. Singh, M. Verma, and A. Sharma, “Deep Learning for Soil and Fertilizer Analysis: A Review,” Sensors, vol. 21, no. 10, p. 3422, 2021.

[10] B. Agarwal and M. Arora, “Comparative Analysis of Machine Learning Models in Agricultural Crop and Fertilizer Recommendation Systems,” Procedia Comput. Sci., vol. 198, pp. 170–178, 2022.