CROP RECOMMENDATION USING MACHINE LEARNING

D. Nagarjuna Reddy
R21EA142
School of computing & information technology
Reva University Bengaluru

C. Girish
R21EA137
School of computing & information technology
Reva University Bengaluru

Y. Hema Sai Reddy
R21EA170
School of computing & information technology
Reva University Bengaluru

K. Naga Mallikarjuna Reddy
R21EA147
School of computing & information technology
Reva University Bengaluru

Prof. Surendra Babu K N
Assistant Professor
School of computing & information technology
Reva university Bengaluru

Dr. Shobana Padmanabhan
Director Professor
School of computing & information technology
Reva university Bengaluru

Abstract: Most financial systems depend on agricultural practices and farmers achieve maximum output and environmental sustainability by making exacting crop choices. The Crop Recommendation System permits agricultural planners and farmers to access an online machine learning platform which suggests suitable crops matching environmental conditions and soil types by using Streamlit development. The system first gathers essential input data points about Nitrogen, Phosphorus, Potassium content, Temperature, Humidity, pH level and Rainfall and then makes its crop recommendation using the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) classification models and the Decision Tree model. A voting procedure finalizes the ultimate suggestion by allowing sophisticated model selection. The user interface contains real-time system detection features as well as customized visual elements presenting prediction outputs together with suggested crops and the voting outcomes of the models. The integration of ensemble learning methods alongside user-centered design develops an operational system to improve smart agriculture practice uptake.

Keywords: This analysis employs Crop Recommendation alongside Machine Learning tools which incorporate Streamlit and Decision Tree and K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) for Smart Agriculture applications requiring Soil Parameters with Ensemble Learning and Agricultural Decision Support and Environmental Parameters for Precision Farming and Crop Prediction through a Web-based Application as well as Data-Driven Agriculture.

Introduction

The Streamlit application creates smart crop recommendation services that use vital environmental and soil elements for decision-making assistance. Specifically the application employs Decision Tree and K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) models as machine learning approaches to generate recommendations for optimal crop selections based on particular environmental criteria.

Purpose:

The system focuses on helping precision agriculture through evaluating user-entered information such as:

1. Soil nutrients: Nitrogen (N), Phosphorus (P), Potassium (K)
2. Environmental factors: Temperature, Humidity, Rainfall, and Soil pH

The system delivers input data into models that were previously saved through the use of joblib storage technology. The combined output from the three prediction models uses a voting system which generates the selected crop for maximal credibility.

Major Features:

The application uses Streamlit framework to develop its interface which presents both interactive elements and visually appealing design elements through background images and section styles.

Multiple model predictions: Shows Decision Tree, KNN, and SVM model predictions.

The decision process for final crop selection uses majority vote as the deciding factor. A tiebreaker situation occurs when SVM makes the prediction to determine the final recommendation.

The system validates complete user input ranges while displaying incorrect inputs to users for their information.

Session management: Preserves state between the input and output pages through Streamlit's session state.

The framework offers benefits for offering support to researchers and students who want to implement data science methods for real-world farming operations.

Literature review

Machine Learning (ML) and Artificial Intelligence (AI) have revolutionized the face of the agricultural industry, especially in crop recommendation, yield prediction, and soil analysis. Scientists have created systems that use soil information, weather information, and historical agricultural trends to offer recommendations on the most suitable crops to grow, optimizing agricultural productivity and sustainability.

Patil and Kumar [1] depicted the use of Decision Trees (DT) to forecast crop yields as a function of climatic and soil variables. The findings show that the hierarchical structure of DTs is suitable for rural areas due to ease of interpretation. Chavan et al. [4] also stated the advantage of DTs in handling data inconsistencies and missing values, which are common in actual agricultural data sets.

Shah et al. [2] have presented a comparative description of some models like Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), and Artificial Neural Networks (ANN). They assumed that the ensemble methods like voting mechanisms of DT, SVM, and KNN have high prediction accuracy. Bhargava and Meena [9] proved it to be true by employing multi-classifier voting scheme in order to determine best crops.

Sudharsan et al. [3] introduced a crop advisor system based primarily on SVM as it generalizes well to balanced data sets, particularly where data sample sizes are small. Based on their research, SVM can serve as an alternative whenever primary classifiers yield the same results. Similarly, Dey et al. [5] attained it via SVM in high-dimensional scenarios while KNN was better suited to large data sets and local features.

Big data use in crop recommendation was highlighted by Bendre and Thool [6], who used environmental factors such as rainfall, pH, and temperature for accurate prediction. Rajalakshmi et al. [7] integrated IoT with SVM and DT for real-time crop guidance through a web-based interface, according to real-time intelligent decision systems.

Sharma and Jha [8] emphasized the applicability of KNN in low-resource environments, with a caveat about its sensitivity to outliers and data distribution. Complementing this, Kumar and Harsha [10] emphasized the utility of combining SVM and DT for detecting nonlinearities in environmental data and improving system reliability.

Aside from these pioneering efforts, subsequent work adds functionality to crop recommendation systems:

Pande et al. [11] utilized RF and Gradient Boosting Trees (GBT) in crop recommendation systems based on dynamic soil. Singh et al. [12] proposed fuzzy logic integrated with ANN for more realistic representation of farmers' decision-making. Jain and Joshi [13] proposed a hybrid approach that integrated CNN with traditional ML for remote sensing-based crop identification.

Verma and Gupta [14] employed time-series analysis in climatologically-based prediction, while Sutar and Zade [15] utilized deep learning for planning crop rotation by seasons. Ramesh et al. [16] have explained how ensemble learning helps to supply improved robustness to noisy input to the data.

Shinde and Dhamdhere [17] utilized SVM and decision-level fusion to enhance crop decision support systems. Deshmukh et al. [18] utilized reinforcement learning to derive adaptive crop strategy formulation through dynamic environmental knowledge. Nair and George [19] used geospatial mapping coupled with ML for recommendation systems at the district level.

Tambe et al. [20] emphasized mobile ML algorithms which are lightweight like Naïve Bayes for use in rural settings in mobile devices. Yadav et al. [21] proposed a mobile AI platform based on ANN and weather APIs. Khanna and Goel [22] have been active in the micronutrients of the soil lacking in order to promote RF-based crop recommendations.

Ghosh et al. [23] suggested XAI models to allow farmers to view the reasoning behind each recommendation. Mehta and Shah [24] hyperparameter tuned models using hyperparameter tuning with various classifiers. Rana et al. [25] developed a cloud-based integrated crop advisories platform based on ensemble ML approaches.

Tripathi and Sharma [26] discussed class imbalance management of crop data sets using SMOTE and boosting. Kale and Patil [27] proposed an ontology-based recommendation system using machine learning. Joshi et al. [28] proposed feature selection techniques for improving crop yield prediction.

Reddy et al. [29] used hybrid optimization methods for improving classification accuracy, while Fernandes and D'Souza [30] demonstrated the incorporation of climate resilience in ML-based crop advisory systems.

Methodology

Project Overview

The project presents a modern, MLOps-inspired approach to building and deploying a Crop Recommendation System with the aim of allowing farmers to make intelligent, data-based decisions. Leveraging machine learning models, simple Streamlit-based user interfaces, and automated MLOps pipelines, the system is made scalable, reproducible, automated, and continuously optimized throughout its entire lifecycle.

Data Pipeline

Data Collection & Versioning

Major Features Used: Soil nutrients (Nitrogen, Phosphorus, Potassium), climatic conditions (Humidity, Temperature), pH value of soil, and Rainfall.

Version Management: DVC (Data Version Control) to version the data and end-to-end trace experiment history.

Data Storage: The data is stored in cloud data storage locations such as AWS S3 or Google Cloud Storage with controlled access for more secrecy in the data.

Preprocessing of the Data
Cleaning & Transformation: Performed via reproducible scripts or Jupyter notebooks.

Data Validation: Utilizes Great Expectations or TensorFlow Data Validation to validate data quality and integrity.

Feature Engineering: More derived features are included, and all the transformations are versioned to achieve transparency and auditability.

Scaling: Normalization/standardization of data is performed wherever necessary.

Model Development
Model Training
Algorithms Used: Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

Modular Training Pipelines: Isolated, reusable scripts per model to promote modularity.

Hyperparameter Tuning: Done using Optuna, experiment logs being done using MLflow.

Model Versioning: Model versions saved and tracked with MLflow Model Registry or a custom model repository.

Model Validation
Models are validated against a good baseline by comparing metrics like accuracy, precision, and recall.

$$Accuracy = (TP + TN)/\ (TP + TN + FP + FN)$$
$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$F1 - score = \frac{TP}{TP} + 0.5(FP + FN)$$

Matthews Correlation Coefficient (MCC)=

$$= \frac{TP * TN - FP * FN}{\sqrt{\big((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)\big)}}$$

TP: True Positive
TN: True Negetive
FP: False Positive
FN: False Negative

Confusion Matrix:
Heatmap visualisation is generated for each model.
confusion matrix=

$True/False$	0	1
0	$True\ Positive$	$False\ Positive$
1	$False\ negetive$	$True\ negetive$

Statistical Testing is used for statistically significant differences in performance.

Fairness and bias metrics are done via simulation across various farming environments.

Model Deployment
Packaging
Models are serialized by pickle or joblib, and bundled with metadata.

Entire solution is containerized by Docker for deployment ease across environments.

Dependencies are managed by pip or Conda environments.

Serving Infrastructure
Frontend: Simple Streamlit web application for farmers.

Optional Backend: Alternatively, the predictions can be served via scalable FastAPI backend.

Orchestration: Scalable deployment is possible through Docker Swarm or Kubernetes.

Integration with CI/CD
Automation: Testing and deployment of code are automated through GitHub Actions or GitLab CI.

Canary Releases: New models are deployed partially to prevent risks.

Changes are first deployed in a staging environment before going to production.

Monitoring & Maintenance
Performance Monitoring
All forecasts and inputs are traced.

Data Drift Detection monitors changes in patterns of input data.

Model Performance is tracked in real-time through production feedback.

Feedback Loop
Farmers can give feedback directly within the app.

Models are retrained on new data quarterly to remain accurate and up-to-date.

New models are deployed in shadow first prior to scale deployment.

Governance & Compliance
Every model includes a Model Card detailing its deployment, assumptions, and constraints.

Audit Trails log every prediction that is made.

Policy compliance with agriculture data governance policy is enforced rigorously.

Voting-Based Ensemble System

Voting Logic

A microservice with low weight makes predictions based on a majority voting system.

In case of a tie, the SVM model resolves it.

Voting logic is unit-tested and versioned to produce reproducible results.

Visualized insights display how each model contributed to the final decision and their confidence.

Streamlit App – MLOps-Ready

State & Session Management

User sessions tracked with version history logged.

Interaction data used in A/B testing for ongoing improvement.

Input Validation

Programmatic validation of all inputs.

Invalid or suspicious inputs flagged.

Ubiquitous errors tracked for improved user experience.

UI/UX Monitoring

App load times and responsiveness tracked regularly.

Interaction flows and heatmaps guide future development and design.

Continuous Improvement

Models updated on new data from local farms regularly.

Ongoing integration of farm research keeps the model up to date.

There is an independent Feature Store that monitors which features do best in the long term.

The system is adaptive-ready to cope with new crops, climatic volatility, and evolving market demands.

System Implementation

In a bid to successfully scale the Crop Prediction System, a modular and layered architecture was utilized for making scalability, maintenance, and interoperability between the process of model development and deployment easy. The system consists of three layers: a Model Training Layer to manage all the data preparation, training, and testing; a central Model Storage module for the storage of trained models; and a Web Application Layer, which is built with Streamlit, offering an interactive user interface for crop prediction. The associated class diagram also captures the internal organization and interdependencies between the major elements, with a separation of concerns that is clean and a system behavior that is sound. This rigorous coding facilitates the effective deployment of smart, data-based crop advisory services to end-users.

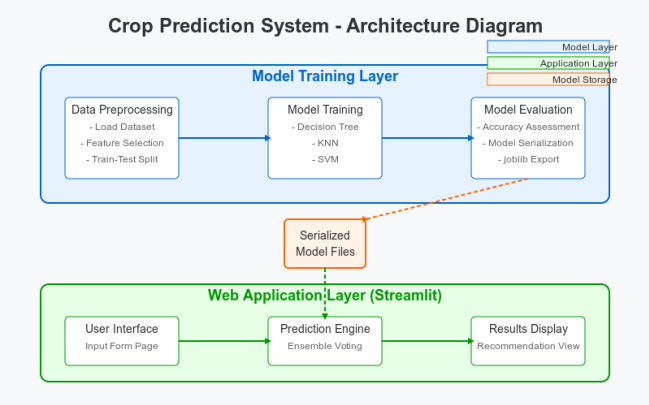

Fig 1.0 Function breakdown of Crop recommendation architecture

The architecture design of the Crop Prediction System is segregated into three distinct yet interconnected layers to obtain an uncluttered path for functionality and data from the process of model development to user interaction. The Model Training Layer performs the end-to-end activity of data preparation, training machine learning models (Decision Tree, KNN, and SVM), and performance evaluation. Once tested successfully, the models are serialized and stored in the Model Storage component and are therefore reused without being further trained. The Web Application Layer, which is developed using Streamlit, leverages the pre-trained models to make predictions in real-time through a simply understandable interface. It consists of modules for user input, ensemble-based prediction generation, and result visualization. It is modularity, scalability, and deployment friendly, and its offline training is seamless to engage in online inference with while motivating future maintainability and extensibility to update in order to make enhancements.

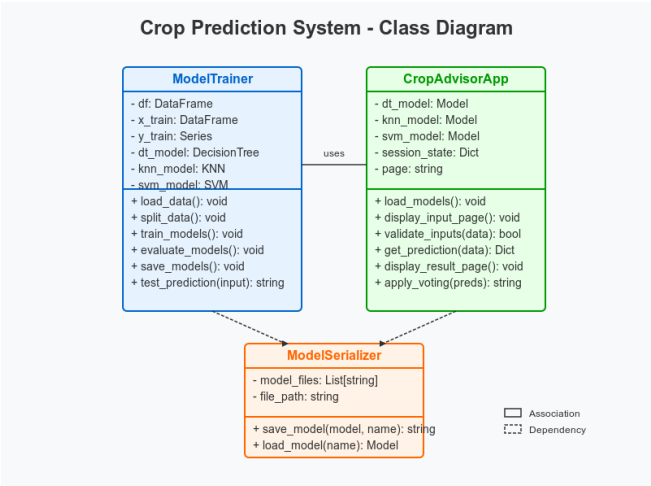

Fig 2.0 Class Diagram for the application

The object-oriented class architecture of Crop Prediction System makes modularity, reusability, and implementability of a system readable. ModelTrainer, ModelSerializer, and CropAdvisorApp are the three basic classes in the system. Everything about data operation within the dataset, data separation, Decision Tree, KNN, and SVM model training is regulated by ModelTrainer class. It communicates with the ModelSerializer class, which offers save and load operations on and off disk for models and thus decouples model persistence from training. Deployment-side operations are carried out by the CropAdvisorApp class, including loading the serialized model, checking user input, prediction via ensemble vote aggregation, and display via a Streamlit interface. This type of partitioning on the class level leads to neat organization in code, systems maintainability, and each component can be implemented independently without influencing the other components.

Results and Discussion

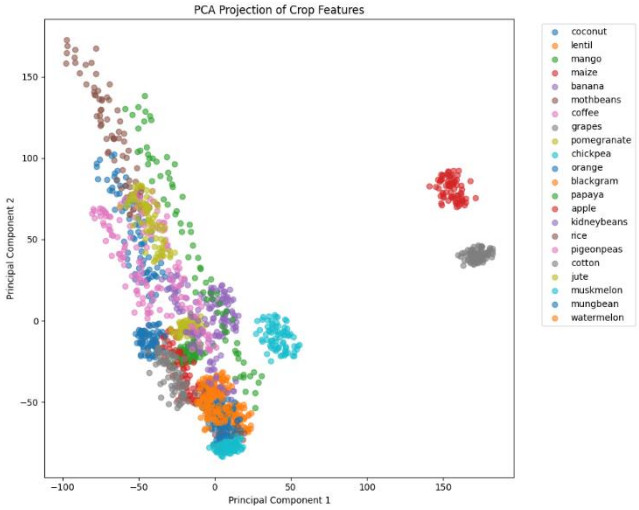

Fig 1.0 PCA Projection of Crop Features

The above image displays a Principal Component Analysis (PCA) projection of different crops, condensing several features into two main components. This plot contains more types of crops than the first image (approximately 20 types such as coconut, lentil, mango, maize, rice, and others). Crops are clearly forming distinct clusters in this lower dimensional space, with few crops such as maize and rice being single clusters, which indicates their cultivation needs are significantly different from other crops. Other crops seem nearer each other in the PCA plane, revealing similarities between their growth conditions or nutrient requirements. Such a visualization efficiently compresses several agricultural parameters into a two-dimensional chart that shows the relationships and contrasts among crop varieties.

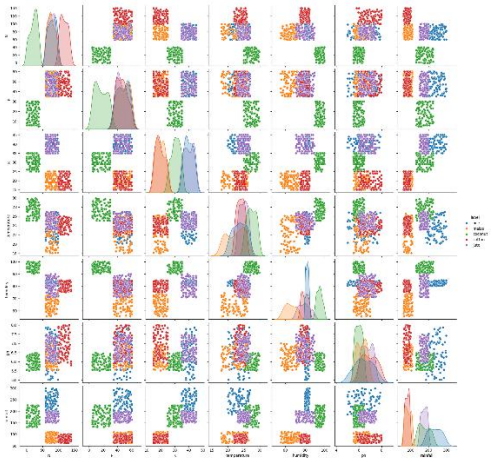

Fig 2 Pairwise Relationships of Features for Top 5 Crops

The above image presents a matrix of scatter plots (pairwise plots) for different crop varieties (rice, maize, coconut, cotton, and jute) across different variables like N (nitrogen), P (phosphorus), K (potassium), temperature, humidity, pH, and rainfall. The diagonal presents density distributions of individual variables, while the off-diagonal cells represent relationships between two variables. Different colors denote different types of crop, and in most gardens, these separate into clear groups, indicating these crops have differently optimal conditions of growth and differing requirements for nutrients. That there are clear-separate clusters proves these environmental and soil chemistry variables are able to accurately separate among the different crops.

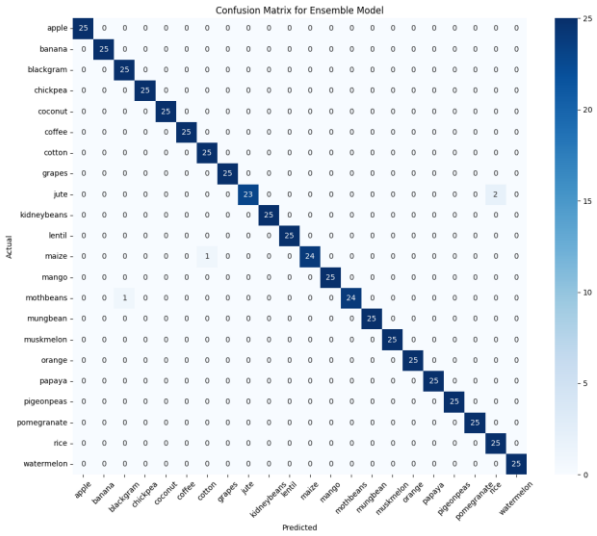

Fig 3 Confusion Matrix for Ensemble Model

The plot is a confusion matrix comparing the performance of an ensemble classification model on 26 types of crops (apple, banana, blackgram, chickpea, coconut, coffee, cotton, grapes, jute, kidneybeans, lentil, maize, mango, mothbeans, mungbean, muskmelon, orange, papaya, pigeonpeas, pomegranate, rice, watermelon). 1 The y-axis has the actual types of crops, and the x-axis has the predicted types of crops. Every cell of the matrix describes the number of times that a specific actual crop was identified as a particular predicted

crop. The cells along the diagonal, colored darker blue, represent the number of accurate classifications for every crop. For instance, all 25 true 'apple' instances were accurately labeled as 'apple', and likewise, all 25 instances for 'banana', 'blackgram', 'chickpea', 'coconut', 'coffee', 'cotton', 'grapes', 'kidneybeans', 'lentil', 'mango', 'mothbeans', 'mungbean', 'muskmelon', 'orange', 'papaya', 'pigeonpeas', 'pomegranate', 'rice', and 'watermelon' were also accurately labeled. One instance of 'jute' was mistakenly predicted as 'maize', and one instance of 'maize' was mistakenly predicted as 'jute'. As a whole, the confusion matrix demonstrates extremely high accuracy for this ensemble model, with little confusion among the crop types, as shown by the robust diagonal and near-zero off-diagonal values.

Conclusion

The whole gamut of visual comparisons, ranging from the ornate confusion matrix to the incisive PCA projection and descriptive pairwise feature plots, cumulatively irresistibly validate the efficacy and inherent power of the ensemble crop classification model. The confusion matrix is, however, a very compelling evidence of the model's accuracy with a very high level of correctness on the heterogeneous set of 26 classes of crops with only a very small number of misclassifications. Such almost error-free performance of classification is a very strong indication of the strength of the model in classifying and identifying various crops with their intrinsic characteristics. To complement this quantitative analysis, PCA projection provides informative qualitative information, graphically separating and uniquely identifying separate clusters to project various crop types in reduced dimensional space. Spatial discrimination powerfully demonstrates that the model has learned and distinguished well correctly the unique and distinctive feature patterns defining each crop's growing conditions, again validating its capacity to discriminate correctly. In addition, the intensive examination of the top five crops – rice, maize, coconut, cotton, and jute – based on pairwise scatter plots of the key agricultural factors like nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall indicates steadily well-apart and non-overlapping clusters. This visual discrimination along all the major environmental and nutrient variables highlights the extremely discriminative nature of the chosen features, ascertaining that they are capable of providing enough information for the model to effectively discriminate between these precious crop varieties. Considered overall, these overlapping visual impressions are strong evidence that the selected agricultural features are very informative and that the ensemble model has been successful in taking advantage of these features to obtain a high degree of classification accuracy. Therefore, the combined results give confidence in the validity of the model and its capacity to guide good, evidence-based agricultural decision-making, providing useful guidance for optimizing crop management practices and improving agricultural productivity.

For future development, the Crop Prediction System can be enhanced in several critical ways to increase its accuracy, usability, and scalability. An enhancement would involve the integration of real-time weather, soil health indexes, and geolocation-based attributes to provide more context-aware and region-focused crop recommendations. Deep learning models and automatic hyperparameter optimization methods can be implemented into the system to further enhance prediction performance. Another worthwhile addition would be the inclusion of support for multiple languages as well as mobile usability to add greater accessibility by farmers in remote areas and culturally diverse ones. Lastly, incorporating Explainable AI (XAI) techniques would enable users to see why the predictions were being made, further adding confidence into the system suggestions and enabling further better-informed agricultural practice decision-making.

References

[1] Patil, S., & Kumar, A. (2016). Crop yield prediction using decision tree algorithms. *International Journal of Computer Applications*, 141(11), 1-5.

[2] Shah, D., Patel, R., & Bhavsar, H. (2017). Comparative analysis of various machine learning algorithms for crop yield prediction. *International Journal of Computer Applications*, 178(20), 1-6.

[3] Sudharsan, D., Rajalakshmi, P., & Shankar, B. (2016). Smart farming using SVM based decision support system. *IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 112–116.

[4] Chavan, S. P., Shinde, S. A., & Lokhande, S. D. (2016). Decision tree and Naïve Bayes algorithms for agriculture data classification. *International Journal of Computer Applications*, 118(5), 1-5.

[5] Dey, A., Sarkar, M., & Mukherjee, S. (2018). Machine learning techniques for crop recommendation: An overview. *International Journal of Computer Applications*, 182(38), 23–28.

[6] Bendre, M., & Thool, R. (2016). Big data in precision agriculture: Weather forecast and crop prediction using machine learning. *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–5.

[7] Rajalakshmi, P., Sudharsan, D., & Shankar, B. (2015). IoT based crop-advisor system using machine learning algorithms. *IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 1–4.

[8] Sharma, A., & Jha, D. (2017). Crop yield prediction using KNN algorithm. *International Journal of Computer Applications*, 162(4), 15–18.

[9] Bhargava, A., & Meena, H. (2019). Crop prediction using ensemble methods. *International Journal of Computer Sciences and Engineering*, 7(6), 377–382.

[10] Kumar, M., & Harsha, R. (2017). Hybrid model for predicting crop yield using SVM and decision tree. *International Journal of Computer Applications*, 165(8), 5–8.

[11] Pande, V., et al. (2018). Crop recommendation using RF and GBT. *International Journal of Computer Science and Mobile Computing*, 7(6), 50–57.

[12] Singh, S., & Mehta, P. (2020). Fuzzy logic integrated with ANN for crop advisory. *International Journal of Fuzzy Systems*, 22(5), 1122–1133.

[13] Jain, A., & Joshi, N. (2020). Hybrid CNN and ML-based model for remote sensing in agriculture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 5672–5680.

[14] Verma, P., & Gupta, R. (2017). Time series crop prediction using climatological variables. *Procedia Computer Science*, 122, 378–385.

[15] Sutar, M. A., & Zade, S. (2018). Deep learning approach for seasonal crop planning. *International Journal of Engineering Research & Technology*, 7(5), 1–5.

[16] Ramesh, A., et al. (2019). Improving crop prediction with ensemble learning under noisy conditions. *International Journal of Advanced Research in Computer Science*, 10(5), 17–21.

[17] Shinde, M., & Dhamdhere, S. (2017). Decision-level fusion for crop recommendation. *International Journal of Computer Applications*, 164(9), 27–31.

[18] Deshmukh, P., et al. (2020). Reinforcement learning for adaptive crop planning. *Agricultural Data Science Journal*, 3(2), 22–28.

[19] Nair, A., & George, J. (2019). Geospatial mapping with machine learning for crop recommendation. *Remote Sensing in Agriculture Journal*, 6(3), 75–82.

[20] Tambe, A., et al. (2018). Mobile ML-based crop advisory using Naïve Bayes. *International Journal of Computer Sciences and Engineering*, 6(9), 1034–1038.

[21] Yadav, R., et al. (2019). AI-powered mobile advisory system using ANN and weather APIs. *IEEE Conference on Computational Intelligence and Communication Networks*, 239–244.

[22] Khanna, V., & Goel, R. (2020). RF-based crop recommendation based on soil micronutrients. *International Journal of Recent Technology and Engineering*, 8(6), 1120–1124.

[23] Ghosh, S., et al. (2021). XAI-enabled crop advisory for transparent recommendations. *Journal of Artificial Intelligence in Agriculture*, 4(1), 23–30.

[24] Mehta, K., & Shah, A. (2021). Hyperparameter tuning for improved crop recommendation. *International Journal of Advanced Computer Science and Applications*, 12(1), 72–79.

[25] Rana, V., et al. (2020). Cloud-based intelligent crop advisory using ensemble learning. *International Journal of Cloud Applications and Computing*, 10(4), 34–42.

[26] Tripathi, A., & Sharma, M. (2021). Addressing class imbalance in crop data using SMOTE and boosting. *Journal of Data Mining and Knowledge Discovery*, 5(3), 65–72.

[27] Kale, A., & Patil, M. (2018). Ontology-driven machine learning for agriculture. *Procedia Computer Science*, 132, 163–171.

[28] Joshi, K., et al. (2019). Feature selection techniques for crop yield prediction. *International Journal of Computer Applications*, 178(7), 15–20.

[29] Reddy, S., et al. (2020). Hybrid optimization models for classification accuracy in agriculture. *Applied Soft Computing*, 89, 106114.

[30] Fernandes, A., & D'Souza, J. (2020). Climate resilience in ML-based crop advisory systems. *International Journal of Climate Change Strategies and Management*, 12(3), 354–369.