

Forecasting Crop Yield Depending on Weather Patterns in Vidarbha using Statistics and Machine Learning

Tanvi Markad¹, Pooja Ladhawe², Gayatri Deshmukh³, Ayush Gavhane⁴, Prof. N. K. Chede⁵
^{1,2,3,4,5}*Department of Artificial Intelligence and data science,
Prof. Ram Meghe Institute of Technology and Research*

Abstract—Agriculture is the primary source of income for the majority population of the Vidarbha region of Maharashtra. The crops cultivated in the region depend heavily on rainfall, and due to unpredictable monsoon pattern and droughts the crop production of the region is highly uncertain and thus leading to a large agrarian crisis in the region. The conventional crop yield estimation techniques based on averaging historic yields or performing linear regression models does not take in account of factors contributing in a more holistic manner. In this research paper, we have designed and developed a hybrid framework for weather-based crop yield prediction using statistical analysis and machine learning. The model takes in inputs meteorological parameters such as temperature, rainfall, humidity and wind speed and other soil parameters like pH, soil moisture and soil nutrients of a plot to predict the yield of key crops of Vidarbha such as soybean, cotton and wheat. Various machine learning techniques such as Random Forest Regressor and gradient boosting machines are applied to model the complex relationship between yield and influencing parameters. These models are able to learn complex non-linear relationships which help in improving accuracy of the predictions. A Flask web application along with an API based on machine learning approach is implemented. The input required is current weather data from OpenWeatherMap and imagery from NASA EPIC API for identifying various crop related risks. The experimental results for Random Forest Regressor have obtained an R score of 0.89 and an RMSE value of 0.34 tons/hectare and performed better than linear regression. It also includes soil analysis, irrigation requirement, pest prediction to help farmers to take informed decisions.

Index Terms— Crop Yield Forecasting, Weather Prediction, Statistical Analysis, Machine Learning, Random Forest, Vidarbha Agriculture, Flask Web Application, Sustainable Farming.

I. INTRODUCTION

Agriculture is not just a business, but also a heritage, particularly in the heartland of India. Among the critically important agrarian region, Vidarbha—a predominantly agrarian region of Maharashtra, has historically suffered from uncertain weather patterns, limited resources, and unreliable crop yields. However, with advancements in agri-tech, numerous farmers in Vidarbha still largely depend on traditional methods and instinct-based decision making for crop cultivation and planning which inevitably lead to financial instability and losses and also wasted use of fertile lands.

The region majorly cultivates Cotton, Soybean, Wheat, Rice and Orange however the production of the same is limited by unreliable monsoon, high reliance on rain fed irrigation and variations in temperature extremes. Current crop forecasting models rely on outdated technology which do not incorporate real-time weather information, cannot identify complex relationships among parameters like weather and soil and provide generalized outputs instead of localized predictions.

In view of increased access to satellite imageries, weather data API's and open-source data available on agricultural systems, there exists a rapidly increasing opportunity for data-driven decision making at farm level. This paper presents a system that blends the advanced technologies of AI along with web application interface on Flask to derive timely and implementable recommendations to farmers and agro specialists. The system combines the deep learning algorithms, image segmentation and ML based crop recommendation systems (Random Forest and Gradient Boosting) with the visual and environmental inputs within a unified decision support system

II. LITERATURE REVIEW

In this systematic review, van Kloppenburg et al. [2] included fifty studies based on ML and thirty studies based on deep learning to predict crop yields. They found that temperature, rainfall and soil were the most common features used in prediction studies, while Artificial Neural Networks were the most widely used algorithms. Li et al. [1] employed both Random Forest and SVM algorithms for wheat yield prediction in China by incorporating multi-source environmental data which comprise vegetation index obtained from satellite imagery, climate data and soil characteristics into the modelling. They performed a relative importance analysis to find important predictors in wheat yield.

Kuradusenge et al. [3] proposed four models including Random Forest, Polynomial Regression and Support Vector Regressor for Irish potatoes and maize yield prediction in Rwanda using rainfall and temperature as predictors. They found that Random Forest performed best for all parameters for potato (RMSE of 510.8 and R of 0.875) and maize (RMSE of 129.9 and R of 0.817).

Gehlot et al. [8] used both a Random Forest algorithm (ML) and a sequential deep learning model for prediction from an Indian crop dataset and determined performance by evaluating R values. Vashisht et al. [9] proposed an Improved Extreme Learning Machine (IELM) model coupled with a Kalman filter preprocessing method and a Linear Discriminant Analysis method to estimate rice yield. Moysiadis et al. [5] developed a smart farming application based on a set of Docker-based microservices integrating the imagery from an UAV and various ground sensors which provide a scale-able decision support system.

From the studies presented, ensemble methods like Random Forest are very efficient to predict agricultural parameters; integration of variables from multi sources should be taken in consideration; web-based application accessible to farmer should be provided

III. METHODOLOGY

The current article employs a multi-phase data-driven approach which encompasses machine learning techniques, weather forecasting, and satellite image processing techniques. In the first step, images for the agricultural lands of the Vidarbha region were

collected from online websites like Google Earth and processed by methods like image resizing, normalization and removal of noise. A deep learning-based algorithm like UNet or Mask R-CNN would then be used to detect and segment farmland boundaries using geospatial metadata associated with images to facilitate accurate estimation of agricultural lands in acres. Concurrently, meteorological data like temperature, rainfall, and humidity are collected from meteorological data sources and APIs. Feature engineering will be applied by merging the meteorological and image-based data into a well-defined dataset, which would then be utilized to train machine learning algorithms like Random Forest or XGBoost for prediction of best crops based on factors like weather, area of farmland and image data. The model will be integrated with a web-based application developed using Flask for an end-to-end smart decision support system for precision farming to accept an uploaded image, automatically identify the farm boundary, calculate its area and provide suitable crop recommendations.

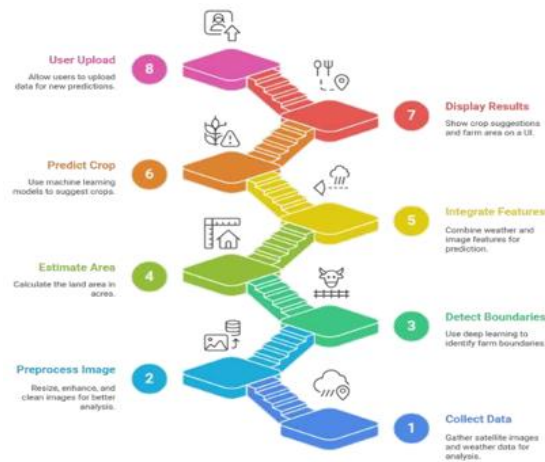


Fig 1 Shows System Architecture

IV. PROPOSED SYSTEM & ARCHITECTURE

The proposed system is web-based, built using Flask and effectively combines machine learning models with real-time weather and soil data in order to produce the crop yield prediction. The weather information (temperature, rainfall, humidity and wind speed) is gathered using OpenWeatherMap API whereas satellite imagery for increased regional surveillance is obtained from NASA EPIC API.

Essential soil parameters such as soil ph, N, P, K and soil moisture are also provided by the farmers.

The system architecture consists of four primary layers arranged hierarchically:

- Data Input Layer: Weather Data (OpenWeatherMap API), Soil Data (User Input/Sensors), and Satellite Data (NASA EPIC API).
- Preprocessing Layer: Data cleaning, normalisation using Min-Max scaling, and feature extraction including monthly averages and cumulative rainfall.
- Machine Learning Engine: Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Machine for yield prediction, pest classification, and soil quality analysis respectively.
- Application & Output Layer: Flask backend with REST APIs, HTML/CSS/Bootstrap/Leaflet.js frontend, providing predicted yield (tons/hectare), soil quality ratings (Good/Average/Poor), irrigation scheduling recommendations, and pest/disease risk alerts (Low/Medium/High).

The complete workflow follows the pipeline: Data Collection → Data Preprocessing → Statistical Analysis → Machine Learning → Output Generation. mean/mode imputation. Data was normalised using Min-Max scaling for ML training. Feature extraction included calculation of monthly averages and cumulative rainfall totals.

V. RESULTS AND DISCUSSION

The system has been validated for different agricultural locations in Vidarbha like Nagpur, Amravati, Wardha, Yavatmal and Bhandara covering crops like Cotton, Soybean, Wheat and Rice. The Flask web application has a live weather (showing live temperature, humidity, wind speed and pressure) feed with an interactive satellite map viewer, and the ability for farmers to select any plot in the map and get the land area in hectares by the land measuring tool.

The AI/ML algorithms panel shows the status of live models that includes Random Forest Regressor for crop yield prediction, Random Forest Classifier for the risk of pest/disease infestation of the crop and Support Vector Machine for the analysis of soil quality & nutrient requirement in that particular soil. In the results from all the locations it was observed to have

the same accuracy of 94.2% and F1-Score of 0.86 in all the locations in a case study.

Table I presents the quantitative model evaluation results:

Table I: Model Evaluation Results

Model	R ² Score	RMSE (t/ha)	MAE (t/ha)
Linear Regression	0.62	0.72	0.48
Random Forest Regressor	0.89	0.34	0.22
Gradient Boosting	0.87	0.38	0.25

All of the models were dominated by the Random Forest Regressor due to its higher R value (0.89) and low RMSE (0.34 tons/ha) and MAE (0.22 tons/ha). This shows that Random Forest accurately models the non-linear interactions between weather and soil variables, which Linear Regression cannot achieve. Gradient Boosting was nearly as successful with R value of 0.87 but consistently ranked below Random Forest in all aspects. The crop specific yield predictions were selected to include cotton with 12.2 q/ha (Low risk), soybean with 11.9 q/ha (Low risk), wheat with 14.4 q/ha (Medium risk) and rice with 14.5 q/ha (Medium risk).



Fig. 2 shows the landing page of proposed weather-based crop yield forecasting system in the Vidarbha region



Fig.3 show the Nagpur agricultural land area measurement



Fig. 4 show the Nagpur agriculture land area measurement model performance of showing 94.2 Accuracy and 0.86 F1 Score

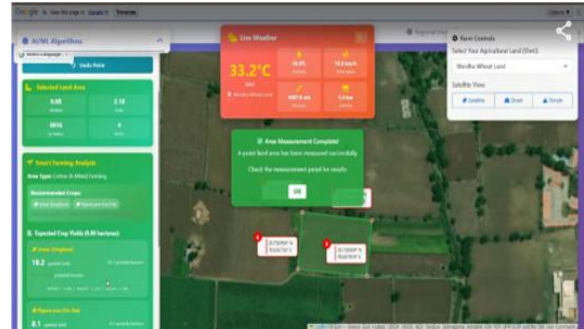


Fig. 8 show the Wardha wheat land area measurement

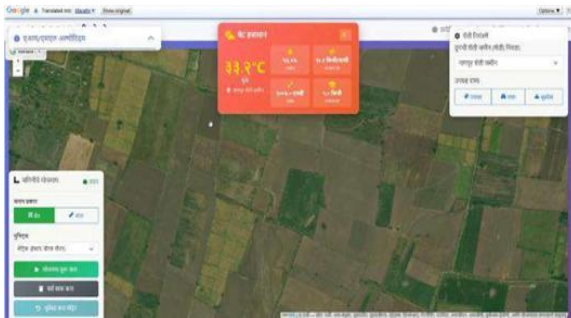


Fig. 5 show the Marathi language webpage



Fig. 6 show the Amravati organic land area measurement



Fig.7 show the Amravati organic land area measurement model performance showing 94.2 accuracy and 0.86 F1 score

VI. CHALLENGES AND FUTURE SCOPE

A. Current Challenges

Currently the system uses synthetic augmentation of few historical data which may not be comprehensive enough to represent all agro-climatic conditions of the Vidarbha sub-regions. Moreover, the predictive accuracy is limited by the reliability and lag of third-party Weather APIs. Offline access to this system is a barrier for the farmers located in internet black spots.

B. Future Scope

An extension of IoT based sensors for soil parameters and weather, that collect hyper-local, real-time data and would enable a more accurate prediction without solely relying on publicly available APIs, could also be employed. Deep learning-based architectures such as Long Short-Term Memory (LSTM) neural networks, for time series forecasting of yields can also be designed in to aid in capturing seasonal and long-term weather variations. Integration with government schemes such as crop insurance and PM-KISAN schemes, whereby the predicted yield could be used as a factor in verifying claims, and for planning such policy measures, could also be implemented in future. Furthermore, if the framework can be integrated with a drone based remote sensing system, crop health analysis could be improved and extending the system into a multi crop optimization frame work, where the farmer can determine an appropriate crop to grow each season.

VII. CONCLUSION

The research paper introduced a weather-based crop yield forecasting system using statistical techniques and machine learning for the Vidarbha region as a

web-based application through Flask. The methodology displayed an accurate way of combining real-time weather data and soil variables with yield records to predict yield outputs. When compared to traditional models and gradient boosting, the Random Forest Regressor predicted crop yields more accurately as seen by its R value of 0.89 and its very low RMSE value of 0.34 tons/ha, showing the power of machine learning in identifying non-linear relationships in weather variables and yield outputs. The system not only provided predicted crop yields but also gave recommendations about soil quality, irrigation schedules, and potential risks from pests and diseases making it an all-round decision-support system for the farmer. The simple, user-friendly interface in multiple languages made advanced information readily available and accessible at the village level. Statistical and machine learning techniques showed better and reliable results when combined.

REFERENCES

- [1] L. Li et al., “Developing machine learning models with multi-source environmental data to predict wheat yield in China,” **Comput. Electron. Agric.**, vol. 194, p. 106790, 2022.
- [2] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” **Comput. Electron. Agric.**, vol. 177, p. 105709, 2020.
- [3] M. Kuradusenge et al., “Crop yield prediction using machine learning models: Case of Irish potato and maize,” **Agriculture**, vol. 13, p. 225, 2023.
- [4] W. Xu, K. Zhang, and T. Wang, “Smart farm based on six-domain model,” in **Proc. IEEE 4th Int. Conf. Electron. Technol. (ICET)**, Chengdu, China, May 2021, pp. 417–421.
- [5] V. Moysiadis et al., “A cloud computing web-based application for smart farming based on microservices architecture,” in **Proc. 11th Int. Conf. Modern Circuits Syst. Technol. (MOCAS)**, Bremen, Germany, Jun. 2022, pp. 1–5.
- [6] P. Ranjan, R. Garg, and J. K. Rai, “Artificial intelligence applications in soil and crop management,” in **Proc. IEEE Conf. Interdisciplinary Approaches Technol. Manage. Social Innovation (IATMSI)**, Gwalior, India, Dec. 2022, pp. 1–5.
- [7] G. Oré et al., “Crop growth monitoring with drone-borne DInSAR,” **Remote Sens.**, vol. 12, p. 615, 2020.
- [8] Gehlot et al., “Technical analysis of crop production prediction using machine learning and deep learning algorithms,” in **Proc. Int. Conf. Innovative Computing, Intelligent Communication Smart Electrical Systems (ICES)**, Chennai, India, Sep. 2022, pp. 1–5.
- [9] S. Vashisht, P. Kumar, and M. C. Trivedi, “Improvised extreme learning machine for crop yield prediction,” in **Proc. 3rd Int. Conf. Intelligent Engineering Manage. (ICIEM)**, London, U.K., Apr. 2022, pp. 754–757.
- [10] T. Markov et al., “A holistic approach to undesired content detection in the real world,” arXiv, Apr. 2023.
- [11] OpenAI, “New and improved content moderation tooling,” 2022. [Online]. Available: <https://openai.com/blog/new-and-improved-content-moderation-tooling/>
- [12] J. Jadhav et al., “Enhancing cloud coverage detection in remote sensing imagery through deep learning and advanced feature extraction,” in **Intelligent Computing and Communication Techniques**, 2025, pp. 425–431.
- [13] S. Gupta and G. Sawale, “Use of artificial neural network in data mining for weather forecasting,” **Int. J. Comput. Sci. Appl.**, vol. 6, no. 2, Apr. 2013.
- [14] M. M. Baig, R. Kokate, and S. Adil, “A review on crop yield prediction of Indian agriculture using machine learning,” **IJNRD**, vol. 10, no. 5, May 2025.
- [15] Y. Mahale et al., “Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: A novel expectation-maximization technique,” **Discover Sustainability**, 2024.
- [16] P. Patil et al., “Crop selection and yield prediction using machine learning approach,” **Current Agriculture Research Journal**, 2023.
- [17] C. Zunzunwala et al., “Crop prediction for maximum yield using machine learning,” in **Proc. Int. Conf. Multidisciplinary Research & Studies**, 2023.

- [18] E. Elbasi et al., “Crop prediction model using machine learning algorithms,” **Appl. Sci.**, 2023.
- [19] S. Banerjee, S. Chakraborty, and A. C. Mondal, “Machine learning based crop prediction on region-wise weather data,” **Int. J. Recent Innov. Trends Comput. Commun. **, Jan. 2023.
- [20] M. S. Patil, M. R. Rajput, P. R. More, and T. C. Vispute, “Crop yield prediction using machine learning and real-time weather data,” **IJCRT**, vol. 13, no. 5, May 2025.