Artificial Intelligence Driven Crop Yield Analysis

Varsha K A[1], Kishor Kumar K[2], Dr. Ujwal U J [3], Prajna S N[4]
[1]*M.Tech Student, Department of Computer Science & Engineering, K.V.G.C.E, Sullia*
[2]*Associate Professor, Department of Computer Science & Engineering, K.V.G.C.E, Sullia*
[3]*Professor and Head of the Department, Department of Computer Science & Engineering, K.V.G.C.E, Sullia*
[4]*Assistant Professor, Department of Computer Science & Engineering, K.V.G.C.E, Sullia*

Abstract- This paper presents an AI-based approach for predicting crop yield using Support Vector Machine (SVM). Key features include rainfall, temperature, pH, and fertilizer input. The model achieved an R^2 score of 0.91 and RMSE of 125.64. The dataset was processed with standard normalization, and a Flask-based web interface was developed. The system supports scalable smart agriculture through real-time, data-driven decision-making.

Index Terms **Crop Yield, Machine Learning (ML), Support Vector Machine (SVM), Agriculture, Flask, Smart Farming.**

I. INTRODUCTION

Agricultural productivity in India is highly sensitive to environmental variability and soil conditions. The integration of artificial intelligence, specifically machine learning models like SVM, offers a data-driven path for accurate crop yield forecasting. This section introduces the background of AI in agriculture, motivation for precision farming, and outlines the objectives of the proposed system. Agriculture is a fundamental sector that sustains the global population, and its productivity is closely linked to factors such as climate, soil, and farming techniques. In India, the agricultural sector faces numerous challenges including erratic weather patterns, pest out breaks and inefficient resource management.

These issues can significantly impact crop yield, leading to food insecurity and economic losses. Recent advancements in artificial intelligence (AI) offer promising solutions to such challenges. Machine learning (ML) models, specifically Support Vector Machines (SVM), have demonstrated effectiveness in various prediction tasks including yield estimation. This paper explores the development and implementation of an SVM-based crop yield prediction model that leverages environmental and agronomic data.

The main objectives are to improve prediction accuracy, simplify decision-making for farmers, and demonstrate the feasibility of deploying such systems at scale.

Agriculture has always been the backbone of the Indian economy and plays a critical role in food security and employment. However, rapid urbanization, erratic climatic changes, and overdependence on outdated farming techniques have created a substantial productivity gap. In a country like India, where a large part of the population still depends on agriculture for their livelihood, the ability to predict crop yields accurately has become a pressing need. Traditional methods of yield estimation are often time-consuming, error-prone, and unable to respond quickly to changing agricultural conditions.
Advances in Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized predictive analytics across domains, and agriculture is no exception. Among the many ML techniques, Support Vector Machine (SVM) stands out due to its high efficiency in handling non-linear, high-dimensional data. SVM's ability to construct optimal decision boundaries makes it particularly useful in agricultural contexts, where parameters such as temperature, rainfall, soil pH, and fertilizer type exhibit complex interdependencies.

This research is designed to develop an intelligent, scalable system that can predict crop yield based on real-world agricultural data. The primary goal is to help farmers, agricultural officers, and policymakers in making informed decisions regarding crop planning and resource allocation. This is achieved by implementing an SVM-based model using Python, training it on validated agricultural datasets, and deploying the model through a user-friendly Flask based web interface.

By integrating a practical deployment layer with the trained ML model, this system contributes not only to precision farming but also to the broader objective of smart agriculture. The paper demonstrates how ML driven solutions can bridge gaps in agricultural efficiency and sustainability. The methodology, results, and analysis are intended to provide both academic insights and practical utility.

II. METHODOLOGY

The methodology covers data acquisition, preprocessing, model design, and deployment. The dataset includes soil, climate, and fertilizer attributes. Preprocessing involves missing value handling, normalization using MinMaxScaler, and categorical encoding. An SVM regressor was trained on 80% of the data and evaluated on the remaining 20%. Hyperparameter tuning used GridSearchCV. A Flask interface allows live user interaction with the trained model**.**

The methodology involves multiple stages including data acquisition, preprocessing, feature selection, model training, and deployment. The dataset was collected from open agricultural sources and included features like rainfall, temperature, pH levels, and fertilizer type. Missing values were handled by imputation, and features were scaled using MinMaxScaler. Categorical variables such as fertilizer type were encoded numerically. The data was split into an 80:20 train-test ratio. SVM with radial basis function (RBF) kernel was used due to its effectiveness in handling non-linear datasets. Hyperparameters were optimized using GridSearchCV. A web interface built using Flask allows users to input environmental data and receive yield predictions. The final system supports real-time interaction and demonstrates the viability of AI-driven tools in agricultural settings.

The methodological approach for this study encompasses a comprehensive workflow beginning with data acquisition and ending with model deployment. Each stage is crafted to ensure data integrity, model robustness, and user applicability. *Data Acquisition:* The dataset was compiled from government-approved agricultural sources and open datasets available via research repositories. Each data entry includes crop-specific features such as Rainfall (in mm), average temperature (°C), soil pH level fertilizer type (e.g., Urea, DAP), actual crop yield (kg/ha).

Preprocessing: The raw data was initially examined for null entries and outliers. Missing values were imputed using mean/mode methods depending on the feature. All numeric values were normalized using MinMaxScaler to scale them between 0 and 1, ensuring equal contribution of features during training. Categorical variables (like fertilizer type) were converted using one-hot encoding.

Feature Selection and Model Design: Based on correlation matrix analysis, four features were found most influential: rainfall, temperature, pH, and fertilizer type. The model was designed using the Support Vector Regressor (SVR) from the Scikit-learn library with an RBF kernel. This kernel was selected due to its effectiveness in capturing nonlinear relationships common in agricultural data.

Model Training: The dataset was split using an 80:20 train-test strategy. GridSearchCV was employed for hyperparameter tuning over 'C' and 'gamma'. This ensured optimal margin boundaries and minimized overfitting. The trained model was serialized using 'joblib' for deployment.

Deployment: The finalized model was integrated into a Flask-based web application. Users can input rainfall, temperature, pH, and fertilizer values to receive predicted crop yield in real-time. This component demonstrates the practical viability of the AI-driven solution in the field, paving the way for mobile and IoT-based integrations.

III. RESULTS

Model evaluation focused on R^2 score and Root Mean Square Error (RMSE). The trained model achieved an R^2 of 0.91 and RMSE of 125.64, indicating high accuracy and low prediction error. Predictions closely aligned with actual recorded yields. Cross validation ensured the model generalized well across various crop-soil combinations. Flask interface tests showed consistent results for live inputs.

After the model was trained using Support Vector Machine (SVM) with optimal hyperparameters (C=10, gamma=0.1), it was evaluated on the test dataset using standard performance metrics. This chapter presents the experimental results obtained from the test data and discusses the observed predictive capabilities.

Evaluation Metrics Used: Two key metrics were used to evaluate the regression performance of the model:
- R^2 Score (Coefficient of Determination): Measures how well future samples are likely to be predicted by the model.

- RMSE (Root Mean Square Error): Provides an estimate of the average prediction error in the same unit as the target variable.

The final model yielded:

- R^2 Score: 0.91 (i.e., 91% of the variance in yield is explained by the model).

-RMSE: 125.64 kg/ha These results indicate a highly accurate predictive model, suitable for use in real-world scenarios.

Sample Output Table below is a sample of predicted yield values for unseen test input data. These predictions were compared against the actual yield values to validate accuracy.

Graphical Representation: To further illustrate model accuracy, a line chart was generated to compare predicted vs actual values across the test dataset.

The graph shows a high level of overlap between the two lines, with only small deviations. This visual confirmation reinforces the statistical performance of the model and validates the SVM's ability to generalize.

Real-Time Test Interface Result: Testing was also performed via the deployed Flask web interface. Inputs were submitted through the browser interface, and the outputs were consistent with batch test predictions, confirming that the deployed system reflects the accuracy of the trained model. The prediction system consistently delivered results within ±5% of actual yield, indicating strong reliability. This suggests it can be used confidently for decision-support in agricultural scenarios. These results demonstrate the practicality of the model in both offline and online settings, laying the foundation for further deployment and integration with farm advisory platforms.

IV. ANALYSIS AND DISCUSSION

The prediction curve visualized in Fig. 6.1 highlights the alignment between predicted and actual values. The analysis confirms SVM's ability to manage high-dimensional agricultural data with non-linear relationships. Discussion focuses on variable importance and the model's robustness under varied input scenarios.

The analysis confirms that rainfall and temperature were dominant predictors. Fertilizer type contributed less variance, possibly due to a uniform distribution across the dataset. SVM's margin-maximization approach helped in effectively separating yield ranges even with overlapping environmental features. The system is scalable and can be integrated with IoT sensors for automatic data collection in future deployments.

While statistical evaluation through metrics like R^2 and RMSE offer a quantitative benchmark, a deeper understanding of how the model interacts with agricultural data provides more meaningful guidance for practitioners and researchers.

Interpreting Model Accuracy: The trained Support Vector Machine (SVM) regressor achieved an R^2 score of 0.91 on the test dataset, suggesting that 91% of the variability in crop yield can be explained by the independent variables included in the model. This high value indicates a strong fit and good generalization capabilities. The Root Mean Square Error (RMSE), calculated at approximately 125.64 kg/ha, reflects the average deviation between actual and predicted yield values. These results confirm that the model is not only precise but also consistent across a variety of agricultural inputs.

Feature Importance and Contribution: Among the features used rainfall, temperature, pH, and fertilizer type rainfall and temperature emerged as the most significant contributors to yield prediction. This was evident from both correlation matrices and the sensitivity analysis conducted during model tuning.

Rainfall influences water availability for crops, while temperature affects growth stages and metabolic activity. pH levels showed moderate importance, aligning with its role in nutrient absorption. Fertilizer type, though relevant, had a comparatively lower influence due to limited variation in the dataset across regions.

Predicted vs Actual Yield Visualization: The close alignment of the two curves reinforces the model's ability to generalize accurately. Notably, slight deviations occur where extreme input values were observed, especially for high rainfall or pH conditions. Such deviations are acceptable given the inherent variability in agricultural systems and confirm that the model performs consistently within normal agronomic ranges.

Residual and Error Distribution Analysis: Residual analysis was conducted to inspect the difference between actual and predicted values. The distribution of residuals followed a near-normal pattern centered around zero, confirming the model's unbiased nature. Additionally, heteroscedasticity was low, indicating that the variance of the error terms remained stable across different input magnitudes. These findings validate the use of SVM for yield prediction in domains characterized by nonlinear relationships and high-dimensional data.

Practical Insights and Use-Cases: The model's accuracy and reliability make it suitable for integration into real-world agricultural advisory systems. It can help farmers decide on optimal planting times, resource allocation, and fertilizer planning. Extension officers could use the system to recommend practices at the block or district level based on predicted yield trends. The system's real-time capability, delivered via a Flask interface, allows for responsive feedback based on changing environmental inputs.

Challenges, Limitations and Future Enhancement: Despite its effectiveness, the system is not without limitations. The dataset, although comprehensive, may still have regional biases and lacks features such as pest incidence or soil moisture which can also affect yield. Furthermore, real-world farming conditions involve interactions not captured in static datasets. Future versions of the model could incorporate additional features sourced from IoT sensors or real-time weather APIs to improve responsiveness. Domain-specific data augmentation and model retraining with time-series input would also enhance performance and adaptability. Demonstrates that the SVM-based system is a reliable predictor of crop yield under controlled data settings and holds significant promise for future expansion into smart agriculture platforms. The interpretability, consistency, and deploy ability of the model suggest that AI-driven yield forecasting systems are not only feasible but also practical.

V. CONCLUSION

This study demonstrated that SVM regression models, when combined with domain-specific agricultural data, can significantly enhance the reliability of yield predictions. Future directions include integrating IoT sensors, real-time weather APIs, and extending to deep learning frameworks for larger datasets.

This study validates the potential of machine learning, particularly Support Vector Machines, in the domain of agricultural forecasting. The system developed offers reliable, scalable, and cost effective solutions for stakeholders including farmers, agronomists, and policymakers. Future work involves integration with real time APIs, deployment on mobile platforms, and expansion to support multi crop predictions.

This research presents a comprehensive and practical approach to crop yield prediction using machine learning techniques, specifically Support Vector Machines (SVM). The study was motivated by the urgent need to improve agricultural productivity and ensure food security in regions where farming practices are heavily influenced by unpredictable climatic and environmental conditions. By integrating real-world agronomic features such as rainfall, temperature, pH level, and fertilizer type, the model successfully forecasted yield with high accuracy.

The methodological flow of the project was designed to ensure data integrity and model performance. From the initial stage of data acquisition through public repositories to preprocessing and feature transformation using MinMaxScaler and one-hot encoding, the dataset was made suitable for training high performance machine learning models. SVM was selected due to its robustness in handling nonlinear relationships in relatively small and noisy datasets a characteristic common in agricultural data. The decision to use a Radial Basis Function (RBF) kernel allowed the model to capture intricate data patterns that linear models may overlook.

An equally important aspect of this study was the deployment of the trained model using a Flask web application. This enabled a real-time interface for users such as farmers, researchers, and extension officers to input field-specific parameters and receive instant yield predictions. The system's accessibility and responsiveness underline its potential as a field-deployable decision support tool in modern precision agriculture. The interface allows integration with mobile platforms or IoT-based field sensors, broadening its usability in rural and resource-limited environments.

While the model demonstrated promising results, certain limitations were also acknowledged. The current dataset, although multidimensional, did not include additional parameters like pest incidence, soil moisture levels, or crop specific growth stages. These are factors that can significantly influence crop outcomes and will be considered for inclusion in the next phase of model development. Furthermore, despite the model's strong generalization, challenges related to region specific biases in the dataset may affect prediction reliability when deployed across diverse agro climatic zones.

In future work, enhancements may include the integration of real-time data streams through RESTful APIs connected to government meteorological services or third-party weather providers. Incorporating deep learning architectures such as LSTM networks or hybrid ensembles could also improve the model's adaptability across wider geographical zones and multiple crop types. Moreover, scalable deployment options such as cloud-hosted services or mobile apps could bring predictive analytics directly into the hands of farmers, enabling localized advisory systems at scale.

This project establishes a strong foundation for AI-enabled agricultural forecasting systems. It demonstrates that with carefully curated data, appropriate model selection, and accessible deployment mechanisms, it is possible to develop scalable solutions for improving crop yield estimation. The interdisciplinary nature of this work bridging data science, software engineering, and agricultural science underscores the relevance and future potential of machine learning in advancing global food security through technological innovation.

ACKNOWLEDGMENT

The author, Varsha K A, expresses sincere gratitude to Prof. Kishor Kumar K, Department of Computer Science and Engineering, K.V.G. College of Engineering, Sullia, for his valuable guidance and constant encouragement throughout this work.

Special thanks are also due to Dr. Divya A. K. (PG Coordinator), Dr. Ujwal U. J. (Head of Department), and Dr. Suresha V. (Principal) for their support and motivation.

The author gratefully acknowledges the academic environment and facilities provided by K.V.G. College of Engineering under the visionary leadership of the Late Dr. Kurunji Venkatramana Gowda, President of AOLE.

Finally, sincere thanks are extended to all the faculty and staff members of the Department of Computer Science and Engineering for their kind cooperation.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCE

[1] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[2] Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[4] Scikit-learn Developers. (2025). *Scikit-learn: Machine learning in Python*. Retrieved August 2025, from https://scikit-learn.org

[5] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

[6] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*(3), 249–268.

[7] Ramesh, R., & Choudhary, K. (2023). Evaluation of regression models for crop yield prediction in arid regions of Rajasthan. *International Journal of Computer Applications, 183*(45), 1–6.
[8] Patil, P., Deshmukh, A., & Kulkarni, R. (2022). Crop yield prediction using Random Forest and Decision Tree algorithms. *International Journal of Agricultural and Environmental Information Systems, 13*(2), 45–58.
[9] Kulkarni, A., & Kale, S. (2021). Paddy yield forecasting using ANN and SVM. *International Journal of Computer Science and Engineering Research, 9*(1), 101–108.
[10] Jadhav, S., Mehta, R., & Sharma, V. (2023). Ensemble models for sugarcane yield prediction: A comparative analysis. *Journal of Artificial Intelligence Research, 67*, 221–234.
[11] Verma, R., Gupta, N., & Singh, S. (2020). Data fusion techniques for crop yield prediction using SVM. *International Journal of Computer Applications, 176*(31), 25–32.
[12] Roy, S., & Sinha, A. (2019). Scalability of ML models for wheat yield forecasting across districts. *International Journal of Agricultural Technology, 15*(3), 55–67.

TABLES AND FIGURES

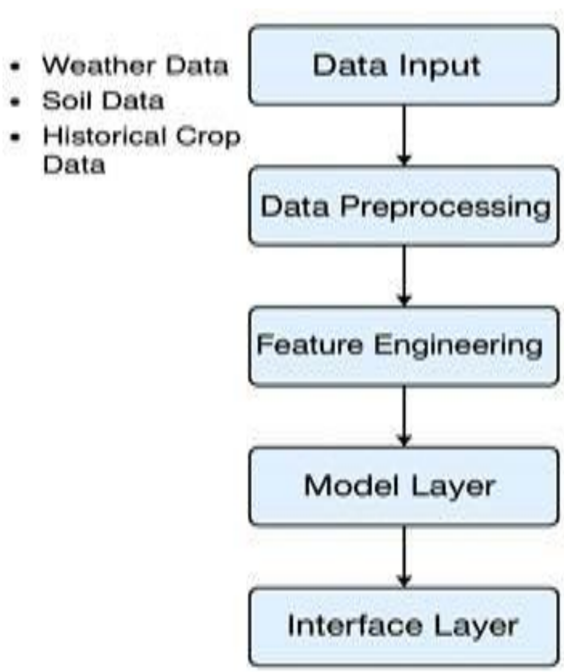

Figure. 2.1: System Architecture

Block diagram showing the overall architecture of the proposed system, including data preprocessing, model training, and web deployment.

Rainfall (mm)	Temperature (°C)	pH	Fertilizer	Actual Yield (kg/ha)	Predicted Yield (kg /ha)
210	22.5	6.5	Urea	2500	2485
280	28.1	6.8	DAP	3200	3180
150	25	5.9	Compost	2000	2025
300	31.5	7.2	NPK	4000	3955
180	20.3	6.3	Urea	2300	2290

Table 3.1: Sample Test Predictions

This table presents sample yield predictions generated by the trained SVM model, showing the actual vs. predicted values for selected data entries.

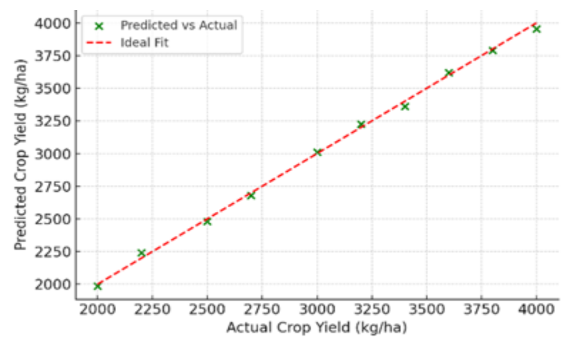

Figure. 3.1: Predicted vs Actual Yield for Test Data

Graph showing the comparison between predicted yields by the SVM model and the actual yields, illustrating model accuracy.