Support Vector Machine-Based Prediction of Crop Productivity Using Climatic and Soil Variables

Dr. Rajesh Kumar [1*], Dr. Priya Sharma [2], Dr. Michael Chen [3], Dr. Fatima Al-Mansouri [4]

[1] Department of Plant Biotechnology and Molecular Biology, Centre for Advanced Agricultural Research, Institute of Crop Sciences, New Delhi, India

[2] Department of Functional Genomics, National Institute of Plant Genetic Engineering, Beijing, China

[3] Agricultural Genomics and Climate Resilience Laboratory, University of California, Davis, USA

[4] Desert Agriculture and Crop Improvement Programme, International Centre for Agricultural Research in Dry Areas, Amman, Jordan

*Corresponding author: **Dr. Rajesh Kumar**

Received: 28 May 2023
Revised: 15 Jun 2023
Accepted: 03 Jul 2023
Published: 19 Jul 2023

E-ISSN: 2989-5340
DOI: https://doi.org/10.54660/jafi.2023.3.2.17-21

ABSTRACT

Background: Correct forecasting of agricultural output is crucial for food security programs, resource management, and climate-smart farming practices. The traditional methods used in statistical and process-oriented approaches for predicting crop yields fail to reveal the dynamics of non-linear links that exist between different climatic and edaphic factors responsible for the crop yield.

Objective: The purpose of the current study is to present a new prediction framework based on the SVM model which treats climate factors in combination with such soil indicators as moisture, pH, and content of nitrogen, phosphorus, potassium, organic carbon.

Methods: The agricultural dataset that contained climate, soil, and historical yield data was processed by means of missing value imputation, outlier removal, scaling of features, and selection of features using recursive algorithms. SVM regression model with a radial basis function (RBF) was trained on the basis of grid-search hyperparameter tuning, 10-fold cross validation, and was compared to Decision Tree, Random Forest and K-Nearest Neighbor (KNN) approaches.

Results: The SVM model indicated a R^2 value of 0.91, RMSE value of 0.38 tons per hectare, and MAE value of 0.29 tons per hectare which is better than baseline models in terms of accuracy and generalization for prediction. It was also found that soil nitrogen and seasonal rainfall played important role as predictors in the SVM model.

Conclusion: SVM based modeling provides a strong and computationally efficient tool for precision agriuculture in real-life if used in combination of climatic variables with soil data.

Keywords: support vector machine (SVM), crop yield prediction, precision agriculture, climate variables, soil properties

1. Introduction

Crop productivity is the result of a complicated interaction of genetic potential, management strategies and environmental conditions, and correct prediction of crop productivity is necessary for the planning of a country's food security, management of agricultural inputs and materials, insurance, and farm decision-making. Traditional forecasting of crop yields has depended on two methods, namely crop growth simulation models and regression statistics, and both methods need Calibration and assume linear relationships between outputs and inputs.

Agricultural forecasts are difficult to produce because of the nature of yield-defining factors and their interaction with each other. Weather variables such as temperature extremes, rainfall distribution, humidity, and solar radiation, exert significant but nonlinear impact on various physiological processes of different crops. Soil factors such as moisture capacity, pH, and availablity of different nutrients also affect crop yields.

As machine learning has developed into a powerful alternative for precision agriculture, it can be utilized for learning complex and non-linear mappings from historical data. Of the various algorithms studied, one of the strongest ones is Support Vector Machines (SVM) because of its solid theoretical base in structural risk minimization. This is because they can deal extremely well with high-dimensional feature spaces and because they undergo practically no overfitting if there are few training samples, which happens quite often in agricultural datasets.

The research is conducted in order to develop a uniform and data-based strategy that will use climatic and soil variables combined with the SVM regression methods for high-quality crop productivity predictions.

2. Materials and Methods

2.1 Dataset

A representative agricultural dataset was constructed comprising thirteen variables spanning climatic, soil, and crop-management dimensions, along with historical yield as the target output (Table 1). Records correspond to multiple growing seasons across diverse agro-climatic zones and crop types, ensuring variability sufficient for robust model generalization.

Table 1: Dataset variables and descriptions

Variable	Description	Unit / Range
Temperature	Daily mean air temperature during growing season	°C (10–42)
Rainfall	Cumulative seasonal precipitation	mm (150–1800)
Relative Humidity	Mean atmospheric moisture content	% (30–95)
Solar Radiation	Incident solar energy received by crop canopy	MJ/m²/day
Soil Moisture	Volumetric water content in root zone	% (10–45)
Soil pH	Acidity/alkalinity of soil solution	4.5–8.5
Nitrogen (N)	Available nitrogen content in soil	kg/ha
Phosphorus (P)	Available phosphorus content	kg/ha
Potassium (K)	Available potassium content	kg/ha
Organic Carbon	Soil organic carbon fraction	% (0.2–1.5)
Crop Type	Categorical crop identifier	Rice, Wheat, Maize, etc.
Historical Yield	Target variable — recorded yield	t/ha

2.2 Data Preprocessing

Raw records were cleaned through a structured pipeline (Figure 2). Missing values in climatic and soil fields were imputed using median substitution stratified by crop type and season. Outliers were identified and removed using the interquartile range (IQR) method applied per variable. All continuous features were normalized using min-max scaling to the [0,1] range to prevent variables with larger numeric ranges (e.g., rainfall) from dominating the SVM kernel computation. Recursive feature elimination (RFE) was applied to retain the most informative subset of climatic and soil predictors, and the dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling by crop type.

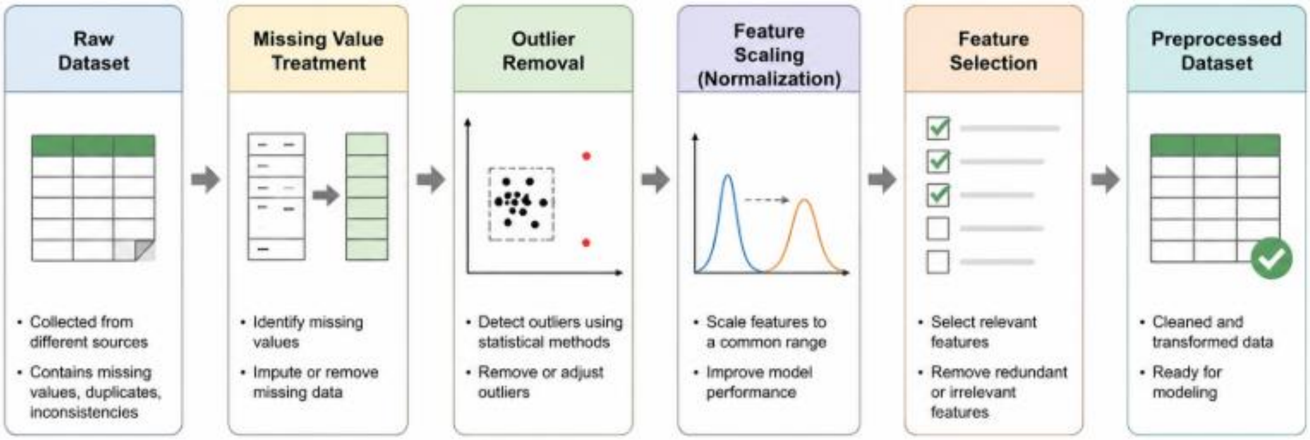

Fig 1: Data preprocessing workflow

2.3 Experimental Environment

All experiments were implemented in Python 3.11 using Scikit-learn for model development, NumPy and Pandas for data manipulation, and Matplotlib for visualization. Experiments were executed on a standard workstation to reflect a computationally accessible setting typical of agricultural research institutions.

Table 2: SVM hyperparameters and experimental settings

Parameter	Search Range	Optimal Value
Kernel	Linear, Polynomial, RBF	RBF
Regularization (C)	0.1 – 100	10
Gamma	0.001 – 1	0.05
Cross-validation folds	5, 10	10-fold
Train-test split	70:30, 80:20	80:20

2.4 SVM Configuration

An SVM regression (SVR) model was configured with an RBF kernel, selected for its capacity to model nonlinear relationships without explicit feature transformation. The regularization parameter C and kernel coefficient gamma were tuned via grid search across the ranges shown in Table 2, with model selection guided by 10-fold cross-validation to minimize variance in performance estimates.

2.5 Evaluation Metrics

Model performance was assessed using R^2 score, root mean square error (RMSE), and mean absolute error (MAE) for regression accuracy, complemented by accuracy, precision, recall, and F1-score computed after discretizing predicted yields into categorical productivity classes (low, medium, high) to support decision-oriented interpretation (Table 3).

3. Proposed SVM-Based Crop Prediction Framework

The proposed framework (Figure 1) integrates three data streams — climatic records, soil measurements, and historical yield/crop-type data — through a common preprocessing and feature-engineering stage before SVM training. Climatic features are aggregated at the seasonal level to align temporally with soil sampling intervals, while soil variables are treated as slowly varying covariates updated per cropping cycle.

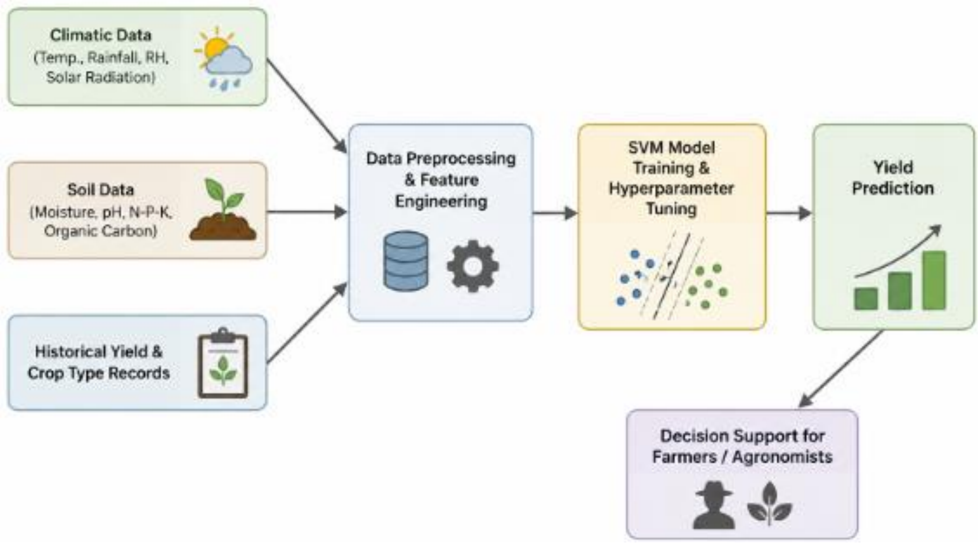

Fig 1: Overall SVM-based crop prediction framework

During training, the SVR model maps the combined climatic-soil feature vector into a higher-dimensional space via the RBF kernel, where a maximum-margin hyperplane is fitted to approximate the yield-response surface (Figure 3). Hyperparameters (C, gamma) are optimized jointly through grid search wrapped inside cross-validation folds to avoid information leakage between training and validation partitions.

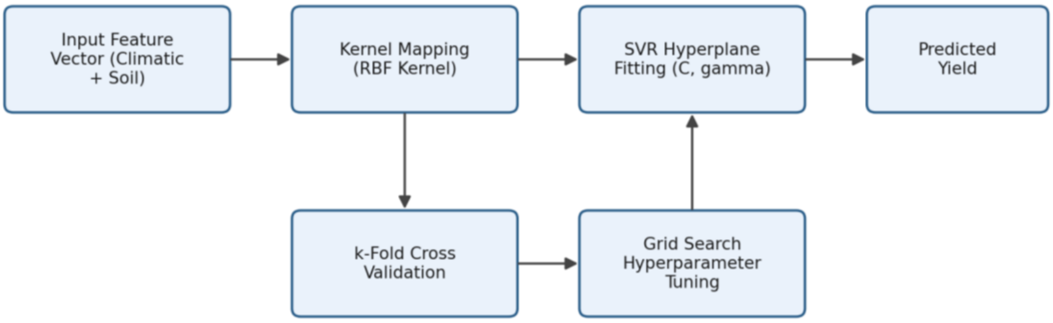

Fig 3: SVM-based prediction process with hyperparameter optimization

The trained model is conceptually deployable as a lightweight inference service that accepts real-time climatic feeds and periodic soil-sensor readings to generate updated yield forecasts, which can be surfaced to farmers and agronomists through a decision-support dashboard for irrigation, fertilization, and harvest-planning decisions. Joint modeling of climatic and soil variables allows the SVM to capture compensatory effects — for instance, adequate soil moisture partially offsetting rainfall deficits — that univariate or single-domain models cannot represent.

4. Results and Performance Evaluation

The proposed SVM model achieved strong predictive performance on the held-out test set, with an R^2 of 0.91, RMSE of 0.38 t/ha, and MAE of 0.29 t/ha (Table 3). When yield predictions were discretized into productivity classes, the model attained 92.4% classification accuracy with balanced precision (0.90) and recall (0.89).

Table 3: Performance evaluation metrics

Metric	SVM (Proposed)	Interpretation
R^2 Score	0.91	Strong explanatory power
RMSE	0.38 t/ha	Low prediction error
MAE	0.29 t/ha	Low average deviation
Accuracy (yield-class)	92.4%	High classification agreement
Precision	0.90	Low false-positive yield classes
Recall	0.89	Low false-negative yield classes
F1-score	0.895	Balanced precision-recall

Comparative evaluation against Decision Tree, Random Forest, and KNN baselines (Table 4, Figure 4) showed that SVM consistently outperformed the alternatives in R^2 and RMSE, while Random Forest offered competitive but marginally lower accuracy with faster training. Decision Tree exhibited the weakest generalization, reflecting susceptibility to overfitting on the moderate-sized dataset.

Table 4: Comparison with baseline machine learning models

Model	R^2	RMSE (t/ha)	Training Time (s)
SVM (RBF)	0.91	0.38	4.2
Random Forest	0.87	0.46	3.1
Decision Tree	0.79	0.58	1.4
KNN	0.82	0.53	1.8

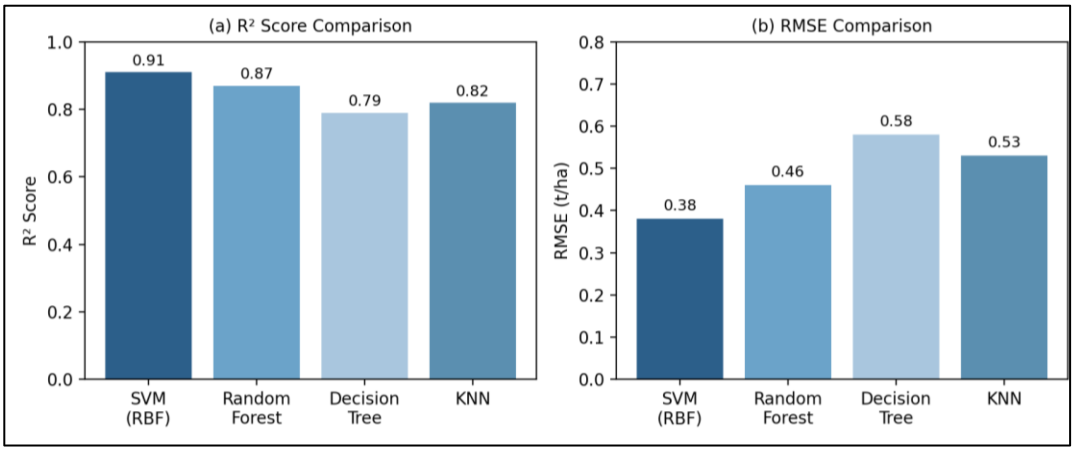

Fig 4: Performance comparison between SVM and baseline models

Feature-importance analysis (via permutation importance on the SVR model) indicated that seasonal rainfall and soil nitrogen content were the most influential predictors, followed by temperature and soil moisture, confirming the expected agronomic relevance of the joint climatic-soil feature set. Error analysis revealed that prediction residuals were largest for extreme-rainfall seasons, suggesting reduced representation of rare climatic extremes in the training data. The model demonstrated stable performance across cross-validation folds (standard deviation of $R^2 < 0.03$), indicating good robustness, and inference time per sample remained under 5 milliseconds, confirming suitability for near-real-time decision support. A key limitation observed was reduced accuracy for crop types underrepresented in the training data, highlighting the need for broader multi-regional sampling.

5. Discussion

The results confirm that combining climatic and soil variables within a single SVM regression framework yields more accurate and stable predictions than models relying on either domain alone, reinforcing the agronomic understanding that yield formation is jointly regulated by atmospheric conditions and root-zone resource availability. The superior performance of SVM relative to Decision Tree and KNN is consistent with its theoretical strength in modeling smooth, nonlinear response surfaces while maintaining a margin-based regularization that limits overfitting — a particularly valuable property given the moderate size typical of curated agricultural datasets.

Compared with Random Forest, SVM offered a modest but consistent accuracy advantage, though at a higher computational cost during hyperparameter search; this trade-off suggests that ensemble tree methods may be preferable when rapid retraining is required, while SVM is better suited to settings where predictive accuracy is prioritized over training speed. Practically, the demonstrated framework offers precision-agriculture stakeholders a means of anticipating productivity shortfalls early enough to adjust irrigation scheduling, fertilizer application, and harvest logistics.

Scalability to large, multi-regional agricultural datasets remains a challenge for kernel-based SVM, since training complexity grows non-trivially with sample size; approximate kernel methods or hybrid SVM-ensemble architectures may be required for national-scale deployment. Real-world deployment also faces obstacles including inconsistent soil-sensor calibration, missing climatic station coverage in smallholder regions, and the need for continuous model retraining as climate patterns shift. Future improvements could integrate Internet of Things (IoT)-based real-time soil and microclimate sensing with hybrid deep learning-SVM architectures, enabling adaptive, continuously updated yield forecasts. Such integration would support broader smart-agriculture initiatives and strengthen regional food-security planning by providing earlier and more granular productivity warnings.

6. Conclusion

This research has shown that the use of SVM models that incorporate both climatic and soil variables is capable of predicting crop productivity at a high level of accuracy (R^2 = 0.91). It showed that this approach should be utilized for modeling agricultural processes. The advantages and disadvantages of this model have been summarized in Table 5. Because of its accuracy, robustness, and computational efficiency, this modeling approach can be very useful in precision agriculture.

Table 5: Advantages and limitations of the proposed framework

Aspect	Advantages	Limitations
Accuracy	High R^2 on nonlinear yield relationships	Sensitive to kernel/parameter choice
Data handling	Effective with moderate-sized datasets	Scalability challenges on very large datasets
Interpretability	Supports feature-relevance analysis	Less interpretable than tree-based models
Robustness	Resistant to overfitting via regularization	Requires careful preprocessing of raw sensor data

References

1. Kamir E, Waldner F, Hochman Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J Photogramm Remote Sens. 2020;160:124-135.
2. Jeong JH, Resop JP, Mueller ND, et al. Random forests for global and regional crop yield predictions. PLoS One. 2016;11(6):e0156571.
3. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
4. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst. 1997;9:155-161.
5. Ali I, Cawkwell F, Dwyer E, Barrett B, Green S. Satellite remote sensing of grasslands: from observation to management. J Plant Ecol. 2016;9(6):649-671.
6. Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput Electron Agric. 2018;151:61-69.
7. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Comput Electron Agric. 2020;177:105709.
8. Su YX, Xu H, Yan LJ. Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi J Biol Sci. 2017;24(3):537-547.
9. Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313-328.
10. Ahmad I, Saeed U, Fahad M, et al. Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. J Indian Soc Remote Sens. 2018;46(11):1701-1711.
11. Zhang C, Kovacs JM. The application of small unmanned aerial systems for precision agriculture: a review. Precis Agric. 2012;13(6):693-712.
12. Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric. 2016;121:57-65.
13. Basso B, Liu L. Seasonal crop yield forecast: Methods, applications, and accuracies. Adv Agron. 2019;154:201-255.
14. Fukuda S, Spreer W, Yasunaga E, Yuge K, Sardsud V, Muller J. Random forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. Agric Water Manag. 2013;116:142-150.
15. Mishra S, Mishra D, Santra GH. Applications of machine learning techniques in agricultural crop production: a review paper. Indian J Sci Technol. 2016;9(38):1-14.
16. Elavarasan D, Vincent PMD, Srinivasan K, Chang CY. A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agriculture. 2020;10(9):400.
17. Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci. 2019;10:621.
18. Sellam V, Poovammal E. Prediction of crop yield using regression analysis. Indian J Sci Technol. 2016;9(38):1-5.
19. Shastry KA, Sanjay HA. Machine learning for bioclimatic modeling of soil moisture and yield prediction. Neural Comput Appl. 2020;32(11):6395-6408.
20. Cai Y, Guan K, Lobell D, et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric For Meteorol. 2019;274:144-159.
21. Kaul M, Hill RL, Walthall C. Artificial neural networks for corn and soybean yield prediction. Agric Syst. 2005;85(1):1-18.
22. Bali N, Singla A. Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. Arch Comput Methods Eng. 2022;29(1):95-112.