Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 27 (2) :227 -229 (June - 2025)
https://doi.org/10.54386/jam.v27i2.2917
https://journal.agrimetassociation.org/index.php/jam

Short communication

Soybean yield prediction leveraging advanced ensemble machine learning models

RAM MANOHAR PATEL* and KAMAL BUNKAR

Institute of Computer Science, Vikram University, Ujjain, Madhya Pradesh, India
**Corresponding author: rammanoharpatel@gmail.com*

Agriculture is vital for economies, but faces challenges from climate change that impact crop yields. Soybean, a key protein-rich oilseed, helps combat protein-energy malnutrition (Patel *et al.,* 2019). It is grown across various agro-ecological zones of Madhya Pradesh (MP). Weather variables critically influence soybean growth at different stages, with temperature (Bhagat *et al.,* 2017) and soil moisture affecting yields (Dakhore *et al.,* 2024). Predicting yield needs understanding of impact of factors throughout the growth period. Recent advancements in yield prediction have shifted towards machine learning (ML) and deep learning (DL). Sridhara *et al.,* (2024) highlighted how ML and DL can optimize crop yields by recognizing patterns and mitigating adverse environmental impacts. Artificial neural networks (ANN) provides better sugarcane yield predictions compared to random forest regression (RFR), support vector machine (SVM) and simple multilinear regression (SMLR) using weather variables throughout different growth stages. Dhinakaran and Thangavel (2024) developed various predictive models, including SMLR, RFR, support vector regression (SVR), cat boost regression (CBR) and hybrid models with variance inflation factor (VIF), with the hybrid CBR-VIF model showing the best performance in predicting paddy yield based on temperature and rainfall. Sridhara *et al.,* (2023), reported that SVM, random forest, least absolute shrinkage and selection operator (LASSO) and elastic net outperformed SMLR in pigeon pea yield prediction. ML and DL techniques, utilizing extensive district-wise data from India over 24 years, have significantly improved agricultural yield estimation (Sharma *et al.,* 2023).

Prior studies focused on weather-related yield predictions neglecting multi-sourced data. This study employed district level multi-sourced data of remote-sensing derived gridded climatic, soil moisture and seasonal fertilizer use variables in comparing ensemble regressors like decision tree (DTR), gradient boosting (GBR), RFR and XGBoost regression (XGBR), with SMLR. Combining soil moisture with nutrient use, improves prediction accuracy and promotes sustainable crop management.

Data description and pre-processing

The soybean yield, weather, soil moisture, and

Table 1: Main analysis variables description

Variables name	Description	Unit/Value
Yield	Soybean yield	kg ha^{-1}
Tmax	Maximum temp. at 2 m	^{0}C
Tmin	Minimum temp. at 2 m	^{0}C
RH	Relative humidity at 2 m	%
SR	Solar radiance	kW-hr m^{-2} day^{-1}
SMR	Root zone soil moisture (Surface to 100 cm depth), **0:** Water-free soil, **1:** Saturated soil	0-1
SMT	Top surface soil moisture (Surface to 5 cm depth), **0:** Water-free soil, **1:** Saturated soil	0-1
RF	Rainfall	mm
N	Nitrogen	000 Tons
P	Phosphorus	000 Tons
K	Potassium	000 Tons
NPK	N+P+K	000 Tons

Article info - DOI: *https://doi.org/10.54386/jam.v27i2.2917*
Received: 4 February 2025; Accepted:15 April 2025; Published online : 1 June, 2025

"This work is licensed under Creative Common Attribution-Non Commercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) © Author (s)"

Table 2: Selected feature for analysis

Variable name	Features (regression coefficients)
RF	RF31 (-0.13), RF33 (0.21), RF34 (-0.21), RF35 (-0.26), RF37 (-0.18), RF38 (-0.18)
SMR	SMR26 (-0.30), SMR34 (-0.53)
SR	SR27 (0.16), SR29 (-0.18), SR31 (-0.16), SR32 (-0.23), SR33 (0.23), SR34 (-0.28), SR35 (-0.22), SR36 (0.44)
Tmax	Tmax38 (-0.41)
Tmin	Tmin25 (0.06), Tmin28 (-0.23), Tmin33 (-0.18), Tmin35 (0.38), Tmin38 (0.14)
RH	RH28 (0.58)
N	N (-0.27)
K	K (0.10)

- Feature names followed by week (SMW) number and regression coefficients in brackets

fertilizer consumption data (Table 1) for MP, India from 1990 to 2022 have been compiled. Gridded daily climate (excluding rainfall) and soil moisture data were sourced from NASA POWER (NASA, 2024), while daily rainfall data from the Indian Meteorological Department (IMD, 2024). Key weather and soil variables include maximum and minimum temperatures, average relative humidity, solar radiation, and rainfall; root zone soil moisture (SMR) and surface soil moisture (SMT). Soybean area, production, and yield data were obtained from ICRISAT (2024) and the Directorate of Economics and Statistics (DES, 2024). Seasonal fertilizer consumption data for Nitrogen (N), Phosphorus (P) and Potassium (K) were acquired from the Fertilizer Association of India (FAI, 2024).

Daily data converted to weekly data based on standard meteorological weeks (SMWs) from 25th to 38th SMW, marking soybean sowing and harvesting periods. The study focused on 17 major soybean districts with over 75% soybean area relative to net cropped area, which collectively represented over 53% of soybean area and production in MP for 2022-23. Stepwise SMLR sequentially adds features at a 0.05 p-value and removes features at a 0.1 p-value at every step. Feature selection by both stepwise and VIF techniques at a 5% significance level identified 25 predictors from 102 weekly variables generated from eleven key features and the dataset was split into 70% training and 30% test sets for model calibration and validation (Arvind *et al.,* 2022).

Features having significantly positive impact on yield are rainfall of 33rd SMW; SR of 27th, 33rd and 36th SMWs; Tmin of 25th, 35th and 38th SMWs; RH of 28th SMW and K consumption; while, rest have significantly negative impact (Table 2). Among them RH of 28th SMW is most important and Tmin of 25th SMW is least important features.

Evaluation of models

All selected features significantly influence soybean yield at the 5% level, with VIF below 10 and a Durbin-Watson statistic of 1.937 indicating no multicollinearity (Setiya *et al.,* 2024). In training, DTR has the lowest MAE, RMSE, nRMSE and highest R^2; however, in testing, RFR performs best with lowest MAE, RMSE, nRMSE and highest R^2, delivering 0.165, 0.24, 0.056, and 93.5% precision. XGBR, with 500 estimators, achieves 93.3% precision, while MLR explains 72% variability. RFR followed by XGBR provide optimal precision (Table 3).

The study has performed feature selection and comparative assessment of the ensemble learning models' predictability for soybean yield, in Madhya Pradesh, India. The stepwise feature selection with VIF methodology has been applied using SMLR to select the best set of predictors for developing prediction models. Rainfall during the 33rd SMW; solar radiation of 27, 33 and 36 SMWs; Tmin of 25, 35 and 38 SMWs; RH of 28th SMW and K use have a significantly positive influence on yield. Among the models, RFR followed by XGBR outperformed the other models. Hence, it is concluded that RFR can reliably be used to forecast soybean yield before harvesting.

ACKNOWLEDGEMENT

The authors are gratefully acknowledge all the institutions for providing necessary secondary data.

Funding: No external funding is involved

Table 3: Evaluation matrices for model performance during training and testing

Models\Matrices	Training				Testing			
	MAE	RMSE	nRMSE	R^2	MAE	RMSE	nRMSE	R^2
Simple multilinear regression (SMLR)	0.404	0.498	0.114	0.750	0.395	0.499	0.117	0.720
Decision tree (DTR)	0.000	0.000	0.000	1.000	0.254	0.504	0.118	0.715
Gradient boosting (GBR)	0.156	0.199	0.046	0.960	0.224	0.294	0.069	0.903
XGBoost regression (XGBR)	0.067	0.091	0.021	0.992	0.174	0.245	0.057	0.933
Random forest regression (RFR)	0.064	0.089	0.020	0.992	0.165	0.240	0.056	0.935

Conflict of Interests: Authors declare that there is no conflict of interest related to this article.

Data availability: Data used in the analysis is available freely and can be retrieved from the source.

Authors contribution: **R. M. Patel**: Conceptualization, data collection, model development and data analysis, manuscript drafting; **K. Bunkar**: Conceptualization, Manuscript finalization.

Disclaimer: The contents, opinions and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Arvind, K.S., Vashisth, A., Krishanan, P. and Das, B. (2022). Wheat yield prediction based on weather parameters using multiple linear, neural network and penalised regression models. *J. Agrometeorol.,* 24 (1): 18-25. https://doi.org/10.54386/jam.v24i1.1002

Bhagat, K.P., Bal, S.K., Singh, Y., Potekar, S., Saha, S., Ratnakumar, P., Wakchaure, G.C. and Minhas, P.S. (2017). Effect of reduced PAR on growth and photosynthetic efficiency of soybean genotype*s. J. Agrometeorol.,* 19(1):1-9. https://doi.org/10.54386/jam.v19i1.734

Dakhore, K.K., Kadam, Y.E., Kadam, D.R., Mane, R.B., Kapse, P.S. and Bal, S.K. (2024). Crop-weather relationship of soybean in Marathwada region of Maharashtra. *J. Agrometeorol.,* 26(2):163-167. https://doi.org/10.54386/jam.v26i2.2438

Dhinakaran, S., and Thangavel, C. (2024). Weather based paddy yield prediction using machine learning regression algorithms. *J. Agrometeorol.,* 26(3):344-348. https://doi.org/10.54386/jam.v26i3.2598

DES (2024). Directorate of Economics and Statistics (DES), Ministry of Agriculture and Farmers Welfare, Government of India. https://ea*nds.dacnet.nic.i*n

FAI (2024). Fertilizer Association of India (FCI). Fertilizer Statistics (1990-2022). New Delhi: Fertilizer Association of India. Accessed from the ICAR–National Soybean Research Institute Library, Indore.

ICRISAT (2024). District-level crop production statistics for India (1966-2022). International Crops Research Institute for the Semi-Arid Tropics. https://data.icrisat.org/dld/

IMD (2024). India Meteorological Department (IMD), Pune, Daily Gridded Rainfall Data (0.25°×0.25°) for India (1990-2022).

NASA (2024). National Aeronautics and Space Administration (NASA), Prediction of Worldwide Energy Resources (POWER) Project, NASA POWER agro-climatology data for 1990-2022. NASA Langley Research Center. https://power.larc.nasa.gov

Patel, R.M., Sharma, P. and Sharma, A.N. (2019). Prediction of girdle beetle (*Oberiopsis brevis*) infestation through pest-weather model in soybean. *J. Entom. Zool. Stud.,* 7(4):718-723

Setiya, P., Nain, A.S. and Satpathi, A. (2024). Comparative analysis of SMLR, ANN, Elastic net and LASSO based models for rice crop yield prediction in Uttarakhand. *Mausam*, 75(1):191-196**.**

Sharma, P., Dadheech, P., Aneja, N. and Aneja, S. (2023). Predicting agriculture yields based on machine learning using regression and deep learning. *IEEE Access*, 11:11255-111264.

Sridhara, S., Manoj, K.N., Gopakkali, P., Kashyap, G.R., Das, B., Singh, K.K. and Srivastava, A.K. (2023). Evaluation of machine learning approaches for predi*ction of pigeo*n pea yield based on weather parameters in India. *Int. J. Biometeorol.,* 67(1):165-180.

Sridhara, S., Soumya, B.R. and Kashyap, G.R. (2024). Multistage sugarcane yield prediction using machine learning algorithms. *J. Agrometeorol.,* 26 (1):37-44. https://doi.org/10.54386/jam.v26i1.2411