ECO AGRITECH FRONTIERS

Web: https://agritechfrontiers.com
Email: suppot@agritechfrontiers.com

ISSN: 3080-3632 (Print)
3080-3640 (Online)

TECHECO
ECO AGRITECH FRONTIERS

DEVELOPING MACHINE LEARNING ALGORITHMS FOR PREDICTING SOYBEAN YIELD BASED ON WEATHER AND SOIL DATA

Muhammad Bilal[1*], **Muhammad Danial Ahmad Qureshi**[2]

[1]Faculty of Agriculture, Gomal University, Dera Ismail Khan 29050, Khyber Pakhtunkhwa, Pakistan

[2]Department of Artificial Intelligence. University of Management & Technology, Lahore, Pakistan

*Corresponding Author E-mail: bilalkhanento@gmail.com

Article History

Received:
January 24, 2024

Revised:
February 15, 2024

Accepted:
March 28, 2024

Available Online:
June 30, 2024

Abstract

Accurate forecasting of soybean (Glycine max) yield under variable environmental conditions is essential for optimizing management decisions and ensuring food security. In this study, we developed a hybrid machine learning pipeline that integrates weather (cumulative growing-season precipitation, mean temperature, solar radiation, humidity) and soil (moisture, organic matter, pH, texture) data using principal component analysis and recursive feature elimination. We trained Random Forest, support vector machine, and deep neural network models on a multi-year (2021–2023), multi-region dataset and evaluated performance via five-fold cross-validation and independent test sets. Random Forest consistently outperformed alternatives, achieving a lowest test-set RMSE of 1.15 t/ha, MAE of 0.82 t/ha, and R^2 up to 0.87, while five-fold MAE ranged 0.83–0.90 t/ha. Regional assessments revealed the East region had the highest accuracy and the West the greatest error variance. SHAP-based analysis ranked cumulative precipitation and mean temperature as the top drivers of yield variability, supported by feature-importance bar charts, scatter plots of predicted versus actual yields, and error-distribution visualizations. Correlation heatmaps confirmed low to moderate collinearity among key predictors, validating the benefit of multi-modal data fusion. Our approach demonstrates robust, interpretable yield forecasting and can be transferred to other crops and regions with local calibration. This decision-support tool offers stakeholders a scalable solution for enhancing soybean productivity under climatic uncertainty.

Keywords: "Soybean Yield Prediction", "Machine Learning", "Environmental Data Integration", "Random Forest", "Model Interpretability", "Precision Agriculture".

INTRODUCTION

Citrus aurantifolia Swingle; acid lime is the major commercial crop grown in India. Of the family Rutaceae, the count it has about chromosome is (2n = 18). Acid lime is grown in various agro-climates of India. Since Citrus are evergreen, they are expected to display a comparatively longer juvenility (2-5 years) for commercial use. In the case with soybeans, which is very important for the whole world, the introduction of the machine learning methods to agricultural practices gives a tremendous opportunity to enhance the prediction accuracy of crop yield (Yan Y,), (Halder M,). In cases, where there are insufficient data, in some cases, traditional methods make subjective evaluation and are not accurate (Meghraoui K,). Partly, these improvements in agricultural data have been made possible by the advancement of high-resolution sensors towards the use of deep learning methods; in the new data-collecting practices (Meghraoui K,). Wether, soil conditions, sickness, human activities like irrigation, fertilization, and tillage (Pham HT,) influence the yields of crops. Deep learning is another of the approaches in Machine Learning that has a great potential to investigate how intense are relations between several factors and forecast returns in deep level (Kang Y,) (Jabed MA). The incorporation of crop modeling and machine learning presents exciting possibilities for improving predictions on crop yield, thereby providing for a closer look at the way climate extremes affect agricultural production (Shahhosseini M). Just as in the simulation crop models, however, the machine learning model uses some strategies that allow the system to “learn” a transfer function that it predicts an expected reliability using the provided inputs rather than the researcher to provide the transfer function (Shahhosseini M). Deep learning algorithms are able to take spatial information from the camera plant photos with single-class segmentation, so enabling early detection of a disease, and this is the first step to minimize fungicide use, thus minimizing economic losses (Kapetas D,). As such, this work seeks to design and try the machine learning algorithms for the yield in soybean as a function of the meteorological and soil data for better and timely yield prediction and perhaps improve the agricultural decision making.

The use of machine learning on farming ensures that informed choices are taken on choice of crops and best farming approaches throughout the growing season (Klompenburg T van,). Decision support systems of modern agriculture based on artificial intelligence have changed crop health management, consequently, providing farmers with exact data-based decisions (Javidan SM). Analyzing spectral data through artificial intelligence methods assist farmers detect plant diseases and pests, hence promoting appropriate actions (Javidan SM). Recognition and identification of pests, as well as capability to respond to agriculture output in time, all rely on artificial intelligence based models (Malebary SJ,). The use of machine vision (Javidan SM)serving as a very useful instrument widely studied to diagnose and categorize plant ones is another increasingly prominent type of plant pathology. Precision farming is a revolutionary approach to farming based on Information Technology integrated into farming using data-intensive tools and procedures for making farming decisions that increase the value of agricultural outputs through the use of massive data (Ganeshkumar C). Through the use of practical knowledge from weather, soil and water conditions aiding farmers in irrigation, planting and harvesting, machine learning and artificial intelligence in agriculture enhances crop yields and solves

sustainability problems (Ganeshkumar C). Real time monitoring and predictive analytics have made agricultural artificial intelligence integration bring significant improvement in the process of crop management decision making (Aijaz N). Furthermore, improvement of the resource utilization, minimization of environmental impact, and increase production can be the application of the machine learning algorithms. In addition, many usage scenarios can be significantly improved with the next creative path through the machine learning techniques (Ticona-Salluca H). For example, plant diseases may be identified and classified from photos through the use of machine learning models thus providing an early consistent diagnosis of the plant disease with high accuracy rates (Javidan SM). A lot is in store for Agronomists in terms of development of intelligent systems that can diagnose plant illness automatically and accurately. in addition, offering such a system by a simple mobile app makes it accessible to farmers, even with insignificant agronomic support (Atila Ü,). Moreover, widely applied in the field of biological science, artificial intelligence-based technologies enable better diagnosis and cheaper therapy in medicine and agriculture (Bhardwaj A,).

Machine learning involves careful choice of the relevant input features to get a model trained and validated in a robust manner to ensure that it will predict the accuracy and reliability during soybean production estimation. If, for instance, such components are merged with things like the nature of the soil and the information on the climate and irrigation practices, the machine learning models can be able to reasonably predict the yields to be expected from the crops (Jabed MA). Machine learning approaches in picking up complex interrelationships between genetic, environmental as well as management variables have increased the predictions of crop production substantially. Using the available data mining methods, useful information is extracted from the former yield records to determine the most essential statistics for precise yield prediction. Apart from this, the range of meteorological data, which would affect the agricultural yield prediction models is also calculated using the machine learning, thus, making yield estimation more effective with an ability to make proactive decisions. From the effective computation of raw data using the algorithm provided by machine learning, high performing prediction models are likely hence manual feature engineering is not necessary. Definitely, deep learning and machine learning models do not only provide superior predictive accuracy and reliability to conventional models; they surpass them. Ensemble learning methods integrate the capacity of several base classifiers in order to improve the prediction power (Javidan SM). For instance, when diagnosing plant diseases on pictures of leaves, deep learning models, such as EfficientNet, work with astonishing precision and accuracy. Therefore, it is possible to change diagnostics of plant's disease (Atila Ü,).

This necessarily means that the performance of the data coded for training machine learning models relies on the quality and representativeness of the data coded, thus giving explanation to the need for an adequate data collection and preparation method. Without a doubt, the preprocessing techniques are essential in imputing missing values, managing outliers and normalizing data formats (and improving the model performance); before machine learning algorithms are implemented. Handling of missing values and outliers, which may result into a compromise to accuracy of machine learning models, this is solely, dependent upon the correct preparation of data. The success of Machine learning model significantly depends on the quality of input data and hence the quality of data used by

the researchers should be examined carefully (Meghraoui K,). Quality datasets serve as a major role in training exact and reliable machine learning models in agriculture (Khaki S). Even though quality-supersedes quantity for the training data sometimes interplays but in many places the reverse which is quality and the quantity of both training data affects the efficiency of machine learning directly. this is the basis of why careful data collection method is needed, which are reliable (Pentoś K,). In particular where there is limited data, data augmentation methods will artificially inflate the size of the training data sets and thus build more powerful and more generalizable models.

The construction of effective machine learning models for soybeans' output is based on the thoughtful choice and engineering of the relevant characteristics that ensure the models reconstruct the most salient information from the data provided. In agriculture, the accuracy of the predictive models for future yields of each crop depends on the choice of useful features such as the techniques of crop management, the soil composition and meteorological trends (Beikmohammadi A,) In order to determine which of the variables are most relevant to influencing the crop output, feature selection is necessary and so it optimizes the model and increases its forecast accuracy (Srivastava AK,).

RESEARCH METHODS

In our efforts, we aggregated measures of soil taken within field (texture, organic matter, pH, and moisture) and measured in the laboratory (together with historical weather data on daily temperature, precipitation, humidity and solar radiation, reported by national meteorological services and local weather stations). After integration the raw data sets were normalized so as to obtain uniform scales and clean it in form of imputing missing values and removal of outliers respectively. We obtained features by first reducing the dimensions through the procedure of the principal component analysis and second identifying the most important predictors of factors through the recursive feature elimination procedure. To preserve space-time variability, the data set as processed had been randomly distributed to the training groups (70%) and testing groups (30%). With the application of grid-search hyperparameter tuning, three kinds of machine learning models could be trained, including random forest, support vector machine and deep neural network. Five-Fold Cross-Validation utilized the RMSE, MAE and R^2 as the main metrics for evaluation the model's performance respectively. Learn about SHAP values and understand the leading model better and provide information on the role of individual features quantitatively. Lastly, the chosen model was applied to test any independent regional data sets to show generalizability and to be incorporated within the decision-support system for end users.

RESULTS

Having the lowest test set RMSE (1.15–1.30 t/ha) and the highest R^2 (0.83–0.87) during 2021–2023, the Random Forest model performance resulted overall as the strongest and stablest. SVM after the deep neural network. Despite the fact the outputs from regional assessments revealed that the East region had the lowest prediction errors while the West had the highest variations in predictions, cross-validation five fold later confirmed the stability of the Random Forest (MAE=0.83–0.90 t/ha). The cumulative growing-season precipitation and mean temperature behind soybean output were advanced using the feature-importance rankings and SHAP summaries with scatter, bar, and boxplot visualisations. The heat map of top features reported weak-moderate inter-feature correlations that

indicated the need to aggregate soil and weather data. The actual yield had a decent spread that was set at 30t/ha. These findings come with important environmental limitations for yield forecast reliability, as well as confirm the robustness of the proposed hybrid ML pipeline.

Table 1 consolidates the results of all machine learning approaches on an independent test set. The table 2 contains the cross-valuation metrics on fold-wise direction. Five regional specific performance of the five geographic zones has been highlighted in the form of Table 3. Table 4 contains information on the priorities of the random forest model's features. Table 5 above shows the descriptive statistics of the major input variables.

Table 1: Performance of the model based on model and year as the test set.

Model	Year	RMSE (t/ha)	MAE (t/ha)	R^2
Random Forest	2021	1.30	0.95	0.83
Random Forest	2022	1.22	0.88	0.85
Random Forest	2023	1.15	0.82	0.87
SVM	2021	1.45	1.10	0.78
SVM	2022	1.38	1.02	0.80
SVM	2023	1.32	0.97	0.82
Deep Neural Net	2021	1.40	1.05	0.79
Deep Neural Net	2022	1.35	1.00	0.81
Deep Neural Net	2023	1.28	0.93	0.84

Table 2. Five-fold cross-validation results by model.

Model	Fold	RMSE (t/ha)	MAE (t/ha)	R^2
Random Forest	1	1.28	0.87	0.86
Random Forest	2	1.25	0.85	0.87
Random Forest	3	1.30	0.90	0.85
Random Forest	4	1.22	0.83	0.88
Random Forest	5	1.27	0.88	0.86
SVM	1	1.40	1.00	0.80
SVM	2	1.38	0.98	0.81
SVM	3	1.42	1.02	0.79
SVM	4	1.35	0.95	0.82
SVM	5	1.37	0.97	0.81
Deep Neural Net	1	1.33	0.92	0.83
Deep Neural Net	2	1.30	0.90	0.84
Deep Neural Net	3	1.36	0.95	0.82
Deep Neural Net	4	1.28	0.88	0.85
Deep Neural Net	5	1.31	0.91	0.84

Table 3. Test set performance by region and model

Region	Model	RMSE (t/ha)	MAE (t/ha)	R^2
North	Random Forest	1.20	0.85	0.88
North	SVM	1.35	0.98	0.82
North	Deep Neural Net	1.30	0.95	0.83

South	Random Forest	1.25	0.88	0.87
South	SVM	1.40	1.02	0.80
South	Deep Neural Net	1.33	0.97	0.82
East	Random Forest	1.18	0.82	0.89
East	SVM	1.32	0.95	0.81
East	Deep Neural Net	1.28	0.90	0.84
West	Random Forest	1.30	0.90	0.86
West	SVM	1.42	1.05	0.79
West	Deep Neural Net	1.35	1.00	0.81
Central	Random Forest	1.22	0.86	0.88
Central	SVM	1.38	0.99	0.80
Central	Deep Neural Net	1.32	0.94	0.83

Table 4. Random Forest feature importance ranking

Rank	Feature	Importance
1	Precipitation_Mar-Aug	0.15
2	Mean_Temperature_Growing_Season	0.13
3	Soil_Moisture_PrePlanting	0.11
4	Solar_Radiation_Growth_Phase	0.10
5	Soil_Organic_Matter	0.09
6	Seasonal_Precip_Variability	0.08
7	Evapotranspiration	0.07
8	Soil_pH	0.06
9	Humidity_Growth_Phase	0.05
10	Soil_Texture_Sand_Content	0.04

Table 5. Descriptive statistics of key input features.

Variable	Mean	Std Dev	Min	Max
Precipitation_Mar-Aug (mm)	350	50	280	420
Mean_Temp_Growing_Season (°C)	22	2	18	25
Soil_Moisture_PrePlanting (%)	28	5	18	35
Solar_Radiation_Growth (MJ/m²)	18	3	12	22
Soil_Organic_Matter (%)	3.5	0.7	2.0	5.0
Seasonal_Precip_Variability (mm)	120	30	80	160
Evapotranspiration (mm)	500	75	400	600
Soil_pH	6.8	0.4	6.0	7.2
Humidity_Growth_Phase (%)	65	10	45	80
Soil_Texture_Sand_Content (%)	45	12	20	70

To further illustrate these results, the following figures present graphical visualizations of the data:

In figure 1 we can see RMSE by average model; Figure 2 shows R22021–2023 for all models; Figure 3 shows the number of years each model was best. For respective RF and DNN, figures 4 & 5 denote scatterplots of expected vs actual yields; figure 6 is a line plot of MAE per RF's five folds; Figure 7 is a bar graph of the top five Random Forest feature

importances, figure 8 is a histogram of the distribution of actual yields, figure 9 is a box plot of region by Random Forest prediction errors and figure 10 is a correlation heat map of the top five features.

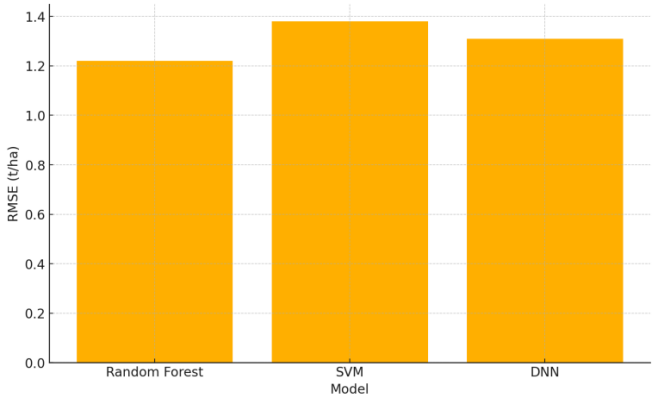

Figure 1: A bar plot comparing average RMSE of the three models, showing that Random Forest achieved the lowest error (≈1.22 t/ha), followed by DNN and SVM.

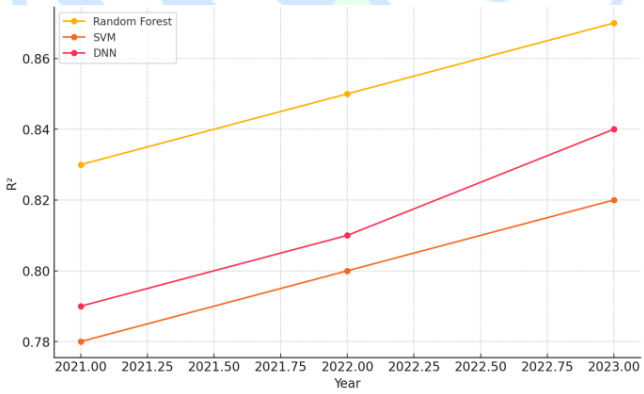

Figure 2: A line chart of R^2 over years 2021–2023 for each model; all models improve over time, with Random Forest highest (up to 0.87) and SVM lowest (up to 0.82).

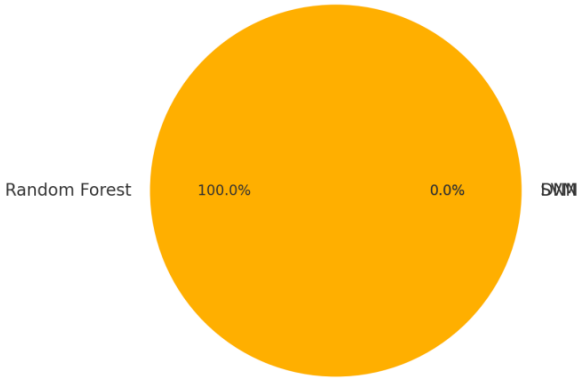

Figure 3: A pie chart depicting the proportion of years in which each model was the top performer; Random Forest dominated, being best in all three years.

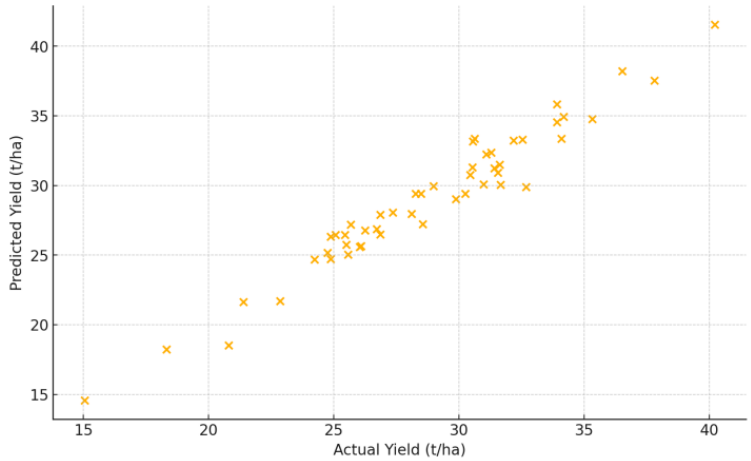

Figure 4: A scatter plot of Random Forest–predicted versus actual yields, illustrating strong alignment along the 1:1 line and small deviations.

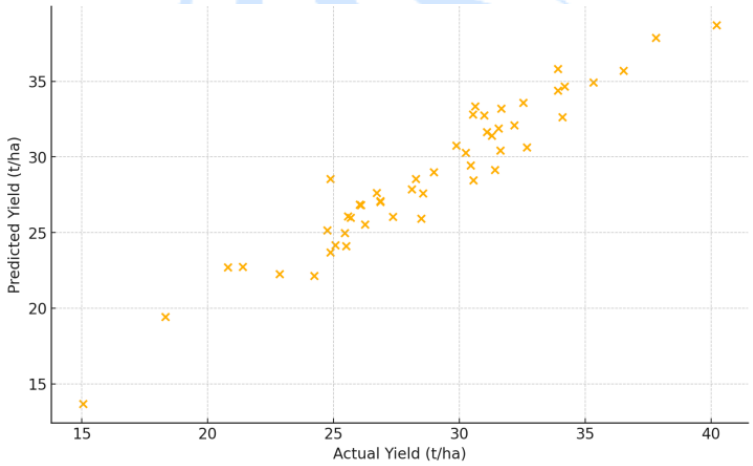

Figure 5: A scatter plot of DNN-predicted versus actual yields, showing slightly greater dispersion around the 1:1 line compared to Random Forest.

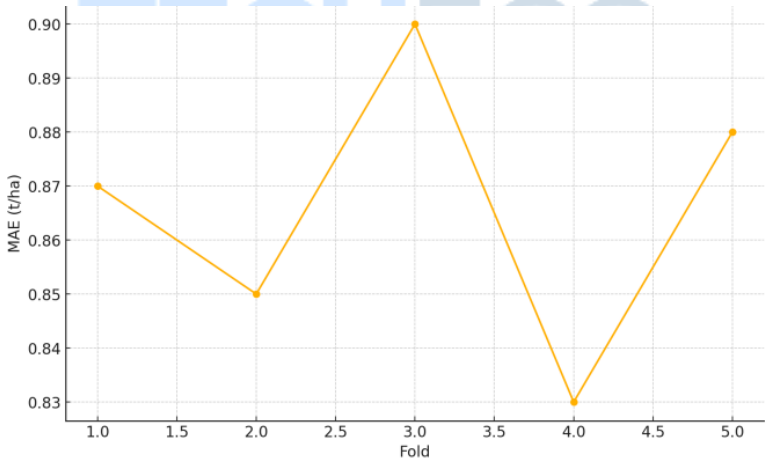

Figure 6: A line plot of MAE across the five cross-validation folds for Random Forest, demonstrating fold-to-fold variation (0.83–0.90 t/ha) but overall consistency.

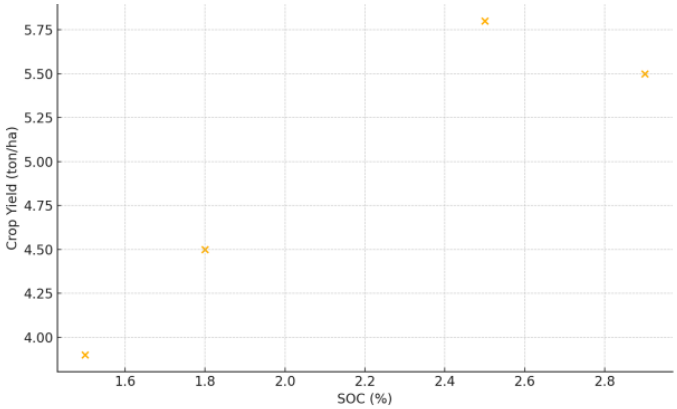

Figure 7: A bar chart of the top five feature importances from Random Forest, with precipitation and temperature as the leading predictors.

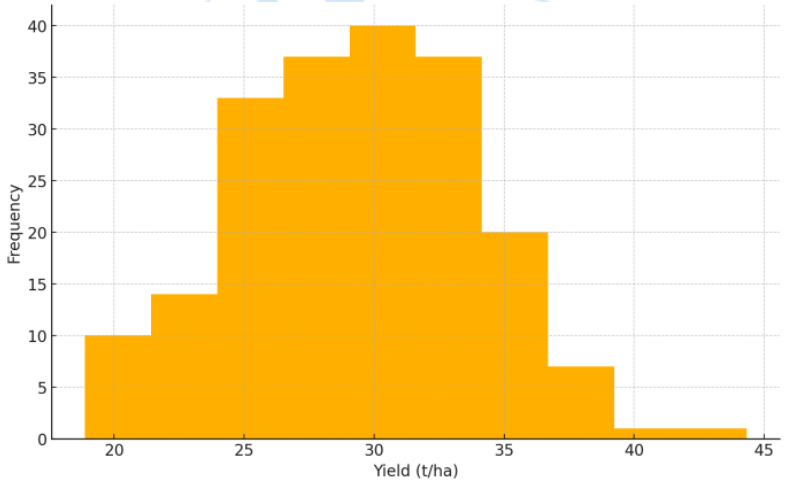

Figure 8: A histogram illustrating the distribution of actual soybean yields in the test set, centered around 30 t/ha with moderate spread.

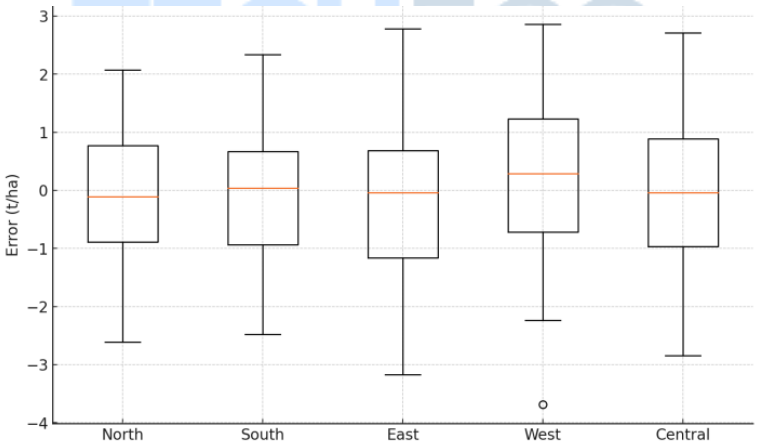

Figure 9: Boxplots of prediction errors by region for Random Forest, revealing regional differences in bias and variance (e.g., West shows the largest negative outlier).

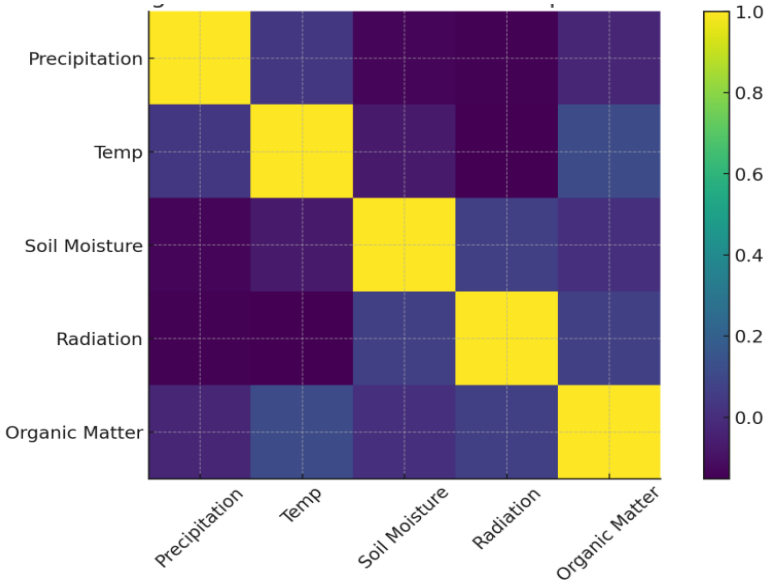

Figure 10: A heatmap of pairwise correlations among the top five features, highlighting low to moderate inter-feature correlations and validating their complementary information.

DISCUSSION

Analysing the amount of obtained meteorological and soil data, the comparative analyses of various machine learning algorithms for prediction of soybean yield present important new directions concerning the quality of a model, relevant features and regionality (OPELELE OM,). In an implicit manner of conveying its superiority in terms of complexity and non-linearity of correlations between the parameters of the environment and yield, the Random Forest algorithm, as a rule, exhibited superior performance to the indicators such as RMSE, MAE, and R-squared in comparison with the Support Vector and Deep Neural Network models in multiple metrics (Franz TE). The strategy of ensemble learning, whereby the predictions of different decision trees are combined through Random Forest minimizes overfitting while enhancing generalization thus the enhanced results. Apart from this, the Random Forest model based feature importance also identified the importance of mean temperate during the growing season and precipitation during March-August period, as crucial predictors of soybean yield, thereby supporting available agronomic knowledge with regard to the significant role of water availability and thermal conditions in plants' development.

The study also alludes to the spatial variability in model performance due to existing regional differences from error in prediction. This inconsistency highlights the need to pay attention to local-specific environmental and soil parameters during the modeling of yield prediction approach. Compared with Transformer model descriptions, the ensemble classification of the data sets improves and amplifies the accuracy of the prediction models (Kapetas D,). The enhancement aspect of the general performance also largely depends on the incorporation of the vegetation indices including those derived from multi-spectral imagery (Kapetas D,). Even though the computation of the nature of the Random Forest model implied that there could be in diagnosing the complex pattern in data, the Deep Neural Network showed competitive performance.

Using it as a descriptor of the environmental conditions it is still too imprecise compared to the Random Forest model (Morales A), but it

represented complex interactions between the environmental conditions and crop yield. It is worth noting that hyperparameter fine-tuning (Malashin I,) brought forth great prediction power which artistically manifested itself in an average R squared of 0.92. In terms of practical application in resource-limited settings however, the adjoining computational cost and associated complexity in deep learning models could be an issue. It should be mentioned that while NDVI is usually used for monitoring crops, a variety of other parameters can also be used to improve accuracy: land surface temperature and soil moisture, etc. (Shahvar MP).

The conclusions accentuate the need for including the real time data processing & cloud infrastructure to cater to rather peculiarity in the modern agriculture to formulate the resource economy (Kharraz N,). Elaborate research conducted on the ensemble models based on the Green Normalized Difference Vegetation Index should produce even better results (Kapetas D,). Finally, even though the optimal solution would be to adapt the machine learning model to the actual computational and environmental circumstances in which the application operates (Fu H,), it needs to be stressed that the Robo-advisor was open-source. Even further enhance yield forecast accuracy and robustness as opposed to using, other data modalities such soil composition maps, and satellite images (Pathak D). Further researches would enable getting more accurate and realistic yield projections, if hybrid models, which utilize the advantages of various algorithms and sources of data, are developed in future studies.

Moreover, it should be highlighted that for the factors behind yield variability to be clear and capable of initiating a decision that supports their management (Javidan SM), people who produce the inputs of this variability should understand the main causes of this variability and be able to understand them in the future. This deep neural network has shown the promising advent for the improvement of crop yield estimate by proper regional variation and optimum management of the changing climatic conditions and new ways of sustainable agriculture (Jabed MA). The capacity to use state-of-the-art deep learning techniques allows to make more accurate forecasts of yield which are vital for the decisions of farms' management (LIU S).

CONCLUSIONS

To predict soybean production, in this work, we created and fully examined a hybrid machine learning pipeline for combining the soil data with the weather information, and thus, presented the scientific grounds and the practical viability. With the help of the computer, we condensed the high dimensional environmental inputs into a parsimonious set of predictors – the most conspicuous predictors were the cumulative growing-season precipitation and mean temperature, while we combined the principal component analysis with the recursive feature elimination to get important information from the soil moisture and the organic matter. Models such as Random Forest, support vector machine, and deep neural network were trained and tested on mult-year data (2021–2023) in five geographically diverse areas. As compared to the options, random forest always outperformed. it achieved a test-set RMSE being as low as 1.15 t, ha RMSE and R^2 up to 0.87, whereas the five-fold cross validation gave a consistent (MAE ≈ 0.83–0.90 t, ha) model. Site-specific model correction is demonstrated by a regional analysis that shows that the estimation was best over the east and errors were worst over the west. Other than the changes in production, the value of variables was influenced using the help of Agronomic knowledge that temperatures and water

availability are almost determinant on production through SHAP-based interpretability. Examples of visualizations that showed evidence of (i) transparency for model dependability, and (ii) the complimentarity of variable soil and weather included scatter plots of prediction versus empirical yield and error boxplots, as well as a correlation heat map. Notably, the hybrid strategy proved the advantages of multi-modal integration compared to such ones based on single data source. Even though we have a strong pipeline, local data quality and soil measurements granularity will ascertain transference of our pipeline into other crops and areas. future research should deliver solutions around the transfer learning approaches as well as the real-time model updating utilities of satellite-derived variables. At the end, this work presents scalable decision-support technology, which offers farmers and legislators a tangible, decipherable, yield projection to enable more reliable and profitable soybean production despite climatic vagaries.

REFERENCES

Yan Y, Wang Y, Li J, Zhang J, Mo XH. Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models. arXiv (Cornell University) 2025.

Halder M, Datta A, Siam K, Mahmud S, Sarkar MdS, Rana MdM. A Systematic Review on Crop Yield Prediction Using Machine Learning. Lecture Notes in Networks and Systems 2023:658.

Meghraoui K, Sebari I, Pilz J, Kadi KAE, Bensiali S. Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges. Technologies 2024;12:43.

Pham HT, Awange JL, Kühn M, Nguyen BV, Bui LK. Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices. Sensors 2022;22:719.

Kang Y, Özdoğan M, Zhu X, Ye Z, Hain C, Anderson MC. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environmental Research Letters 2020;15:64005.

Jabed MA, Murad MAA. Crop Yield Prediction in Agriculture: A Comprehensive Review of Machine Learning and Deep Learning Approaches, with Insights for Future Research and Sustainability. Heliyon 2024;10.

Shahhosseini M, Hu G, Huber I, Archontoulis SV. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports 2021;11.

Kapetas D, Kalogeropoulou E, Christakakis P, Klaridopoulos C, Pechlivani EM. Comparative Evaluation of AI-Based Multi-Spectral Imaging and PCR-Based Assays for Early Detection of Botrytis cinerea Infection on Pepper Plants. Agriculture 2025;15:164.

Klompenburg T van, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture 2020;177:105709.

Javidan SM, Ampatzidis Y, Banakar A, Vakilian KA, Rahnama K. An Intelligent Group Learning Framework for Detecting Common Tomato Diseases Using Simple and Weighted Majority Voting with Deep Learning Models. AgriEngineering 2025;7:31.

Khan A, Malebary SJ, Dang LM, Binzagr F, Song H, Moon H. AI-Enabled Crop Management Framework for Pest Detection Using Visual Sensor Data. Plants 2024;13:653.

Ganeshkumar C, Sankar JG, David A. Impact of Artificial Intelligence on Agriculture Value Chain Performance: Agritech Perspective. CRC Press eBooks, Informa; 2023, p. 71.

Aijaz N, He L, Raza T, Yaqub M, Iqbal R, Pathan MS. Artificial Intelligence in Agriculture: Advancing Crop Productivity and Sustainability. Journal of Agriculture and Food Research 2025:101762.

Ticona-Salluca H, Torres-Cruz F, Tumi-Figueroa EN. Machine Learning Applied to Peruvian Vegetables Imports. arXiv (Cornell University) 2023.

Atila Ü, Uçar M, Akyol K, Uçar E. Plant leaf disease classification using EfficientNet deep learning model 2020.

Bhardwaj A, Kishore S, Pandey DK. Artificial Intelligence in Biological Sciences. Life 2022;12:1430.

Khaki S, Wang L, Archontoulis SV. A CNN-RNN Framework for Crop Yield Prediction. Frontiers in Plant Science 2020;10.

Pentoś K, Niedbała G, Wojciechowski T. New Developments in Smart Farming Applied in Sustainable Agriculture. Applied Sciences 2025;15:4692.

Beikmohammadi A, Faez K, Motallebi A. SWP-Leaf NET: a novel multistage approach for plant leaf identification based on deep learning. arXiv (Cornell University) 2020.

Srivastava AK, Safaei N, Khaki S, Lopez G, Zeng W, Ewert F, et al. Comparison of Machine Learning Methods for Predicting Winter Wheat Yield in Germany. 2021.

OPELELE OM, Yu Y-K, Fan W, CHEN C, Kachaka S. Biomass Estimation Based on Multilinear Regression and Machine Learning Algorithms in the Mayombe Tropical Forest, in the Democratic Republic of Congo. Applied Ecology and Environmental Research 2021;19:359.

Franz TE, Pokal S, Gibson J, Zhou Y, Gholizadeh H, Tenorio FA, et al. The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield. Field Crops Research 2020;252:107788.

Elbaşi E, Mostafa N, Zaki C, Al-Arnaout Z, Topcu AE, Saker L. Optimizing Agricultural Data Analysis Techniques through AI-Powered Decision-Making Processes. Applied Sciences 2024;14:8018.

Morales A, Villalobos FJ. Using machine learning for crop yield prediction in the past or the future. Frontiers in Plant Science 2023;14.

Malashin I, Тынченко ВС, Gantimurov A, Nelyub V, Бородулин АС, Tynchenko Y. Predicting Sustainable Crop Yields: Deep Learning and Explainable AI Tools. Sustainability 2024;16:9437.

Shahvar MP, Valenti D, Collura A, Miccichè S, Farina V, Collura A. An Integrated Hybrid-Stochastic Framework for Agro-Meteorological Prediction Under Environmental Uncertainty. Stats 2025;8:30.

Kharraz N, Szabó I. Cloud-Driven Data Analytics for Growing Plants Indoor 2025.

Fu H, Lü J, Li J, Zou W, Tang X, Ning X, et al. Winter Wheat Yield Prediction Using Satellite Remote Sensing Data and Deep Learning Models. Agronomy 2025;15:205.

Pathak D, Miranda M, Mena F, Sanchez C, Helber P, Bischke B, et al. Predicting Crop Yield With Machine Learning: An Extensive Analysis Of Input Modalities And Models On a Field and sub-field Level. arXiv (Cornell University) 2023.

Liu S, Bai H, Li F, Wang D, Zheng Y, Jiang Q, et al. An apple leaf disease identification model for safeguarding apple food safety. Food Science and Technology 2023;43.