Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 1, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

Harnessing Large Language Models and Stochastic Programming for Optimized Plant Breeding Strategies

ZHOU, Yuqun [1*] CEN, Zuen [2]

[1] University of Wisconsin-Madison, USA
[2] Northern Arizona University, USA

**** ZHOU, Yuqun is the corresponding author, E-mail: yzhou364@wisc.edu***

Abstract: The convergence of Generative AI (GenAI) and stochastic programming introduces unprecedented opportunities for optimizing plant breeding strategies under uncertainty. This paper presents a hybrid framework that integrates Large Language Models (LLMs) with stochastic programming to enhance decision-making in crop improvement. LLMs are employed to analyze vast datasets, generate insights on genotype-environment interactions, and simulate breeding scenarios, while stochastic programming optimizes the selection of genotypes for maximum yield and resilience. Case studies demonstrate the effectiveness of this approach in addressing challenges such as climate variability and evolving market demands, offering a transformative solution for sustainable agriculture.

Keywords: Large Language Models, Stochastic Programming, Plant Breeding, Optimization Strategies, Genetic Improvement, Crop Yield Prediction, Predictive Analytics, Decision Support Systems, Agricultural Technology, Data-driven Modeling.

Disciplines: Biological Sciences. **Subjects:** Genetics.

DOI: https://doi.org/10.70393/616a6e73.323632 **ARK:** https://n2t.net/ark:/40704/AJNS.v2n1a03

1 INTRODUCTION

Plant breeding faces growing challenges in meeting global food security demands due to uncertainties in environmental conditions, genetic traits, and market dynamics. Traditional optimization methods often fail to capture the complexity and variability inherent in breeding systems. Recent advances in Generative AI (GenAI) and Large Language Models (LLMs) offer novel capabilities for synthesizing knowledge from diverse datasets, predicting outcomes, and generating innovative strategies. This paper proposes a hybrid framework that combines LLM-driven data analysis with stochastic programming to address these challenges. The LLMs are used to analyze scientific literature, experimental data, and environmental trends, providing insights into genotypeenvironment interactions and guiding scenario generation. These scenarios are then fed into a stochastic programming model, which optimizes breeding decisions to maximize yield, resilience, and profitability.

2 SHORTLIST OF MACHINE-LEARNING APPLICATIONS FOR CROP IMPROVEMENT AND PRODUCTION

With emerging new technologies and approaches, large datasets are generated from different agricultural domains, particularly from the crop production domain[1]. These vast datasets can easily feed into machine-learning approaches to help all beneficiaries optimize crop improvement systems. Even though machine-learning applications are extensive, their subcategories, mainly in crop quality , crop phenotyping, crop weed identification , disease detection, crop recognition, crop-related microbiome improvements, and yield prediction, were separated into crop development, production, and improvement[2-4].

3 ESSENTIAL CONCEPTS

We discuss several fundamental ideas in machine learning and, whenever possible, present examples from agricultural literature to clarify these concepts[5].

3.1 BASIC TERMS IN MACHINE LEARNING

A dataset consists of several instances, or data points, that are conceptualized as individual experimental observations. Several fixed features describe each data point[6-9]. Phenotype, genotype (SNPs), product price, and climatic parameters are a few examples of these features[10]. Whatever we aim to do with a machine-learning model is specified objectively by a machinelearning task. For instance, we could predict the rate of price fluctuation at a particular

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 1, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

point in time for a specific agricultural product with an experiment examining the cost of the crop product over time[11-16]. In this instance, the features "cost of crop product" and "time" could be referred to as input features. The conversion rate, which would represent the anticipated output of the target model at a specific moment, is the quantity we are interested in forecasting. Input and output features of a model can be as many as desired. Features could be either categorical (accepting just discrete values) or continuous (continuous numerical values are used). Technically, categorical features are usually binary in nature, meaning they can be 1 (true) or 0 (false)[17].

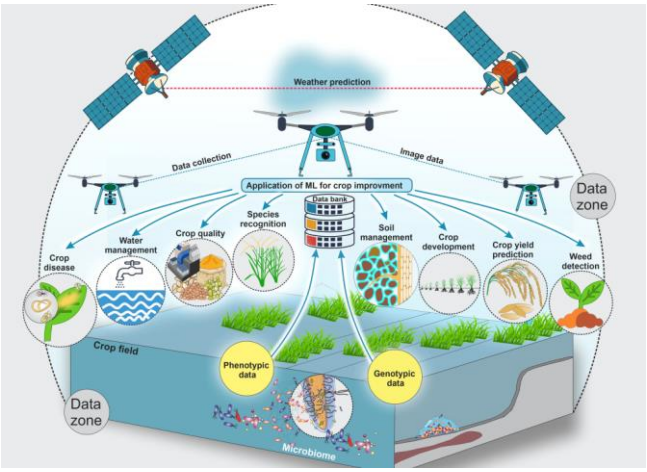

FIGURE 1. THIS SCHEMATIC ILLUSTRATES KEY APPLICATIONS OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR CROP DEVELOPMENT AND IMPROVEMENT, INCLUDING CROP DISEASES, CROP QUALITY, CROP SPECIES RECOGNITION, CROP DEVELOPMENT, CROP YIELD PREDICTION, CROP-RELATED MICROBIOME IMPROVEMENT, WATER MANAGEMENT, SOIL MANAGEMENT, ETC.

Farmers and researchers still encounter numerous obstacles due to employing traditional methods in the crop sector. Artificial intelligence and machine learning are used extensively to address these issues. Also, this figure shows possible data types and collection zones from crop fields to feed different machine-learning models to improve and develop different crops.

3.2 CONCEPT OF SUPERVISED, UNSUPERVISED, SEMI-SUPERVISED, AND REINFORCEMENT LEARNING

Supervised machine learning describes how a model can be fitted to data or part of target data that distinct labels have received for which a ground truth attribute exists; this quality is often determined by experimentation, researchers, or data collectors. In contrast to knowledge derived from inference, ground truth is information verified via direct observation and measurement, thus known to be accurate or real. Among the examples are high-yield prediction and water quality prediction using supervised learning for crop improvement. Laboratory or experimental observations ultimately serve as the source of ground truth in both cases. Contrary to supervised learning, patterns in unlabeled data can be found using unsupervised-learning techniques This approach does not require predetermined labels with ground truth information [18-20]. For example, plant image data can be analyzed using an unsupervised machinelearning technique. Semisupervised learning, in which a significant quantity of unlabeled data is paired with tiny quantities of labeled data, occasionally combines the two methodologies ; for example, weed distribution and density estimation . When obtaining tagged or labeled data is expensive, this can dramatically enhance performance. Another component of machine learning known as reinforcement learning (RL) teaches an agent how to behave and react in a given environment by having it carry out specific tasks and then watching the rewards or outcomes. This technique is already employed in different agricultural domains, such as crop yield prediction and a completely autonomous precision agricultural aerial scouting technique [21-24].

4 CONVENTIONAL MACHINE LEARNING

This section investigates several essential and traditional machine-learning techniques, focusing on their advantages and disadvantages, presents a comparison of several machinelearning techniques along with some applications for crop improvement and production. Figure 4 illustrates a few of the conventional machine-learning techniques[25]. To train these models, several software programs have been available, such as Caret in R, MLJ in Julia , and scikit-learn in Python. When developing machine-learning algorithms for crop improvement-related data, conventional machine learning is typically the first area to investigate to find the most appropriate solution for a given problem. Deep learning is currently prevalent and has the potential to be a robust and valuable method. It is still restricted to the application domains where it performs well, though, such as when a vast quantity of data are accessible, such as extreme data points, when there are several features on each data point or when the features have a lot of structure . Drone images from crop fields and genotypic data (SNPs) are two examples of agricultural data for which deep learning could be effectively used. Even when the other two conditions are satisfied, deep learning may not be the best option because of the need for vast volumes of data. Technically, conventional approaches build and evaluate solutions for a particular problem far more quickly than deep learning. When compared to more conventional models such as random forests and support vector machines (SVMs) , creating the architecture and training a deep neural network might be a computation-intensive and costly process . For a given agricultural prediction problem, even if deep learning seems theoretically doable, it is usually wise to train a conventional technique and evaluate it against a model based on neural networks such as ANN (artificial neural network),

Academic Journal of Natural Science
Journal Home: http://ajns.suaspress.org | CODEN: AJNSAE
Vol. 2, No. 1, 2025 | ISSN 3078-5170 (Print) | ISSN 3078-5189 (Online) | ISSN 3078-5197 (Digital)

if at all possible. Conventional approaches usually assume that every sample in the collection has the same number of characteristics, which is not always feasible. Using SNP data with varying lengths for each case is a clear illustration of this problem. The data can be adjusted using basic techniques such as windowing and padding to make them all the same size and employing standard ways with them. Padding refers to the process that can add zero value to each example up to making the size of each of them equal to the most prominent example in the target dataset. Conversely, the windowing approach condenses each sample to a specific size[26].

4.1 Application of Regression and Classification Models

Regarding regression problems such as those depicted in Figure 4A, ridge regression (a type of linear regression) is frequently a valuable place to start when building and developing a model since it could offer a quick and clear baseline for a particular responsibility. The value of one variable can be predicted by using linear regression analysis according to the value of another variable. On the other hand, when a model relies on as few features as possible from the given data, then other variations of linear regression, such as elastic net regression and LASSO regression , are also worthy of consideration. Since the correlations between the characteristics in the data are frequently non-linear, using a model such as an SVM is usually a better option in these situations , as shown in Figure 4B. SVMs are a practical kind of classification and regression model that convert non-separable problems into easier-to-solve separable problems by using kernel functions. A kernel function is a technique for transforming input data into the format needed for data processing. Both non-linear (a statistical method called nonlinear regression is used to model non-linear relationships between independent and dependent variables) and linear regression could be carried out with SVMs based on the kernel function that was applied. To quantify, the best idea is to train an SVM through a kernel of a radial basis function and a linear SVM can be used from a nonlinear model, if any. Numerous models that are often employed in regression could be used in classification as well. Another acceptable default starting point for a classification problem is to train an SVM based on the kernel function and a linear SVM. k-nearest neighbors classification (also known as k-NN or KNN) is a further technique that could be used. A non-parametric supervisedlearning classifier, the k-nearest neighbors method employs closeness to classify or anticipate how a single data point will be grouped. XGBoost (Figure 4C) are examples of ensemble-based models, which provide another family of resilient non-linear techniques. These techniques are effective nonlinear models offering feature significance estimations and frequently just need minor adjustments to the hyperparameters. There are often an overwhelming number of variations among the several models available for regression and classification. It can be misleading to try to forecast how well-suited a specific method will be to a given issue in advance; instead, it is usually wiser to use an empirical approach to identify the optimum model via trial-and-error methods. Swapping out these model versions often involves only one line of code change thanks to a novel and robust machine-learning library such as scikitlearn , which can efficiently run in a Python environment. To find the best approach overall, it is an excellent strategy to optimize and train several of the previously described techniques, and then compare the results on a different test set to see which method performed the best on the validation set.

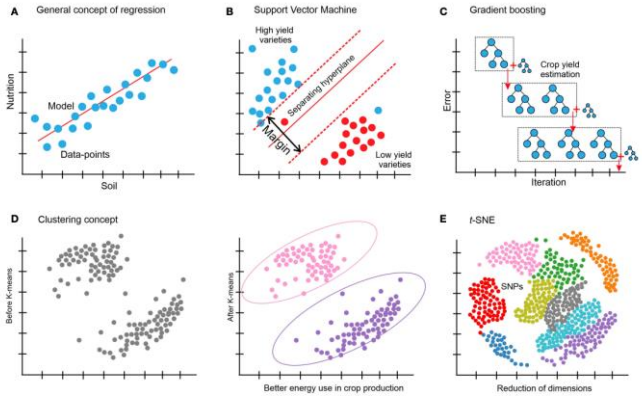

Figure 2. Machine Learning Methods in Agricultural Data Analysis

(A) Regression is the link between a single and/or several independent variables, also known as features, and a dependent variable (the observable attribute) is determined by using regression. A straightforward example is the prediction of crop yield based on one or some of the phenotypic features.

(B) SVM: a support vector machine divides the original input data into several categories by creating a gap as large as feasible between the data in each converted version. One example is a prediction of whether a variety of a specific crop is a low- or high-yield variety.

(C) Gradient boosting makes predictions by combining several weak prediction models, most often decision trees; for example, the prediction of sugarcane yield grade.

(D) Clustering: using one of several algorithms, based on related objects; for example, better energy use in crop production.

(E) t-distributed stochastic neighbor embedding (t-SNE), for example, dimensionality reduction of crop genotypic (SNP) data.

4.2 Application of Clustering Models

Like many other clustering algorithms (Figure 4D), k-means is a powerful multi-purpose clustering technique that requires the number of clusters to be specified as a hyperparameter . An alternate method that is not necessary for a predetermined number of clusters is DBSCAN[27]. For datasets with plenty of features, dimensionality reduction can also be done prior to clustering to enhance performance.

4.3 DIMENSIONALITY REDUCTION

High-dimensional data can be transformed into a lowerdimensional format while preserving the different connections and interactions between the data points and pieces using dimensionality reduction techniques. Although more dimensions could be used in machine learning, two or three dimensions are often selected to enable data visualization on several axes. These methods include data transformations that are both linear and nonlinear. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are some of the examples common in the agriculture domain for dimensionality reduction. The circumstance determines which technique to apply. PCA is based on a linear combination of input features; each component preserves the global connections between the data points and could be explainable, implying that it is simple to identify the characteristics that contribute to data diversity. t-SNE is a versatile technique that can uncover structure in complicated datasets and more robustly maintain local links between data points. Concept of artificial neural networks The mathematical principle of artificial neural networks (ANN) has been conceptualized by following and understanding the behaviors and connectivity of human neurons in the human brain. It was created initially to study the workings of the brain. The significant advances in deep neural network training and architecture over the past few decades have increased interest in neural network models. The following section covers the fundamentals of neural networks and common varieties used in research on crop improvement. Figure 5 displays some of these concepts[28]. Concept of neural network fundamentals The capacity of neural networks to approximate functions universally is one of their primary characteristics; this implies that, with minimal presumptions, any mathematical function can be accurately approximated to any degree by a neural network that is set up appropriately. The fundamental units of every neural network model are artificial neurons. A mathematical function that translates (converts) inputs to outputs in a certain way constitutes an artificial neuron. Any number of input values can be fed into a single artificial neuron, which then uses a predetermined mathematical function to produce an output value. Artificial neurons are layered and the output of one layer is the input of the next, which forms a network. In the following subsections, we present several methods for configuring artificial neurons, sometimes called neural network architectures. Combining several architectural styles is also popular. For instance, fully linked layers are typically used to provide the final classification output in a CNN (convolutional neural network) used for classification.

5 CONCLUSIONS

Display significantly greater use of AI and ML approaches in crop science, which could open a new horizon for integrated and valuable solutions in this area. We have undertaken a thorough review of the essential elements, concepts, applications, and machine-learning definitions required for agri-crop improvement. Nowadays, crop science is leveraging tons of available data to obtain deeper insights through AI and ML and offer the best suggestions for following actions and decisions for enhancing crop productivity or for other necessary tasks. Crop improvement and forecasting are made more accessible by combining computer science and agriculture. Offering broad recommendations and guidance for machine learning in agriculture is challenging because of the diversity of agricultural data. Therefore, our article aimed to provide agricultural and crop science researchers with an overview of the many accessible approaches, as well as some suggestions for conducting efficient machine learning through available data. It is vital to recognize that machine learning is inappropriate for all problems and to know when to avoid it: when the available data are insufficient, when it is necessary to comprehend rather than anticipate, or when it is not apparent how to fairly evaluate performance. Also, here we highlighted the application of federated learning in agriculture along with the definition, procedures, and structure, which can be beneficial for researchers in the agricultural sector. Even though there has been huge progress in machine learning in agriculture, many challenges still need to be addressed to mark ML territory in agricultural science. There is no denying that machine learning has influenced and will continue to influence agricultural research significantly.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in

the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

ZHOU, Yuqun

University of Wisconsin-Madison, USA.

CEN, Zuen

Computer Information Technology, Northern Arizona University, Flagstaff, AZ, USA.

REFERENCES

[1] Cen, Z., & Zhao, Y. (2024). Investigating the Impact of AI-Driven Voice Assistants on User Productivity and Satisfaction in Smart Homes. Journal of Economic Theory and Business Management, 1(6), 8-14.

[2] Farooq, M. A., Gao, S., Hassan, M. A., Huang, Z., Rasheed, A., Hearne, S., ... & Li, H. (2024). Artificial intelligence in plant breeding. Trends in Genetics.

[3] Cen, Z., & Zhao, Y. (2024). Enhancing User Engagement through Adaptive Interfaces: A Study on Real-time Personalization in Web Applications. Journal of Economic Theory and Business Management, 1(6), 1-7.

[4] Zhao, Y., & Cen, Z. (2024). Exploring Multimodal Feedback Mechanisms for Improving User Interaction in Virtual Reality Environments. Journal of Industrial Engineering and Applied Science, 2(6), 35-41.

[5] Lin, W., Xiao, J., & Cen, Z. (2024). Exploring Bias in NLP Models: Analyzing the Impact of Training Data on Fairness and Equity. Journal of Industrial Engineering and Applied Science, 2(5), 24-28.

[6] Zhao, Y., & Wu, J. (2024). Enhancing User Engagement and Behavior Change in Healthy Eating Apps: A Human-Computer Interaction Perspective. Journal of Industrial Engineering and Applied Science, 2(6), 27-34.

[7] Wójcik-Gront, E., Zieniuk, B., & Pawełkowicz, M. (2024). Harnessing AI-Powered Genomic Research for Sustainable Crop Improvement. Agriculture, 14(12), 2299.

[8] Xiao, J., Zhang, B., Zhao, Y., Wu, J., & Qu, P. (2024). Application of Large Language Models in Personalized Advertising Recommendation Systems. Journal of Industrial Engineering and Applied Science, 2(4), 132-142.

[9] Zhao, Y., Qu, P., Xiao, J., Wu, J., & Zhang, B. (2024). Optimizing Telehealth Services with LILM-Driven Conversational Agents: An HCI Evaluation. Journal of Industrial Engineering and Applied Science, 2(4), 122-131.

[10] Zhao, Y., Wu, J., Qu, P., Zhang, B., & Yan, H. (2024). Assessing User Trust in LLM-based Mental Health Applications: Perceptions of Reliability and Effectiveness. Journal of Computer Technology and Applied Mathematics, 1(2), 19-26.

[11] Wu, J., & Xiao, J. (2024). Application of Natural Language Processing in Network Security Log Analysis. Journal of Computer Technology and Applied Mathematics, 1(3), 39-47.

[12] Harfouche, A. L., Jacobson, D. A., Kainer, D., Romero, J. C., Harfouche, A. H., Mugnozza, G. S., ... & Altman, A. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends in biotechnology, 37(11), 1217-1235.

[13] Xiao, J., & Wu, J. (2024). Transfer Learning for Cross-Language Natural Language Processing Models. Journal of Computer Technology and Applied Mathematics, 1(3), 30-38.

[14] Wu, J., Qu, P., Zhang, B., & Zhou, Z. (2024). Sentiment Analysis in Social Media: Leveraging BERT for Enhanced Accuracy. Journal of Industrial Engineering and Applied Science, 2(4), 143-149.

[15] Zhang, B., Yan, H., Wu, J., & Qu, P. (2024). Application of Semantic Analysis Technology in Natural Language Processing. Journal of Computer Technology and Applied Mathematics, 1(2), 27-34.

[16] Qu, P., Zhang, B., Wu, J., & Yan, H. (2024). Comparison of Text Classification Algorithms based on Deep Learning. Journal of Computer Technology and Applied Mathematics, 1(2), 35-42.

[17] Zhong, Y. N. (2024). Optimizing the Structural Design of Computing Units in Autonomous Driving Systems and Electric Vehicles to Enhance Overall Performance Stability. International Journal of Advance in Applied Science Research, 3, 93-98.

[18] Pandey, D. K., & Mishra, R. (2024). Towards sustainable agriculture: Harnessing AI for global food security. Artificial Intelligence in Agriculture.

[19] Zhong, Y. (2024). Enhancing the Heat Dissipation Efficiency of Computing Units Within Autonomous Driving Systems and Electric Vehicles.

[20] Wang, X., Li, X., Wang, L., Ruan, T., & Li, P. (2024). Adaptive Cache Management for Complex Storage Systems Using CNN-LSTM-Based Spatiotemporal Prediction. arXiv preprint arXiv:2411.12161.

[21] Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B. P., ... & Lopez, M. (2024). A foundational large language model for edible plant genomes. Communications Biology, 7(1), 835.

[22] Wang, L., Xu, Z., Stone, P., & Xiao, X. (2024). Grounded curriculum learning. arXiv preprint arXiv:2409.19816.

[23] Sun, Y., & Ortiz, J. (2024). Machine Learning-Driven Pedestrian Recognition and Behavior Prediction for Enhancing Public Safety in Smart Cities. Journal of Artificial Intelligence and Information, 1, 51-57.

[24] Stock, M., Pieters, O., De Swaef, T., & Wyffels, F. (2024). Plant science in the age of simulation intelligence. Frontiers in Plant Science, 14, 1299208.

[25] Sinha, D., Maurya, A. K., Abdi, G., Majeed, M., Agarwal, R., Mukherjee, R., ... & Chen, J. T. (2023). Integrated genomic selection for accelerating breeding programs of climate-smart cereals. Genes, 14(7), 1484.

[26] González-Rodríguez, V. E., Izquierdo-Bueno, I., Cantoral, J. M., Carbú, M., & Garrido, C. (2024). Artificial intelligence: A promising tool for application in phytopathology. Horticulturae, 10(3), 197.

[27] Ferrão, L. F. V., Dhakal, R., Dias, R., Tieman, D., Whitaker, V., Gore, M. A., ... & Resende Jr, M. F. (2023). Machine learning applications to improve flavor and nutritional content of horticultural crops through breeding and genetics. Current Opinion in Biotechnology, 83, 102968.

[28] Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. Scientific reports, 9(1), 17132.