Review Article

Role of Machine Learning Approaches in Plant Genome Analysis: Recent Advances and Challenges

Swati Singh*
School of Science, Uttar Pradesh Rajarshi Tandon Open University, Prayagraj-211013
*Corresponding author: swatinatural@gmail.com

Abstract

The rapid expansion of high-throughput sequencing technologies has revolutionized plant genomics. It has produced large and complex datasets that demand advanced analytical methods. Machine learning (ML) has emerged as a transformative approach for inferring meaningful biological insights from large data sets. In plant genome analysis, ML techniques are being used for gene prediction, genome annotation, functional genomics, genome-wide association studies, epigenomics, regulatory network inference, and crop improvement etc. Recent advances in machine learning and multi-omics data integration have significantly enhanced prediction accuracies based on biological data. However, challenges about data quality, genome complexity, interpretability of ML models, computational facility requirements still remains a challenge. This review provides an in-depth overview of machine learning approaches being used in plant genome analysis. It also highlights recent advances and future perspectives for sustainable agriculture.

Keywords

Machine learning; plant genomics; deep learning; genome annotation; GWAS; crop improvement; bioinformatics

Received on: 14.11.2025
Accepted on: 24.12.2025
Published on 26.12.2025

Introduction

The importance of plant genomics has been a major method of studying new areas in plant genetics and crop enhancement. The recent advances in next-generation sequencing technologies have improved the knowledge of the complex plant genomes in ways that could not be achieved a decade ago (Libbrecht and Noble, 2015; Zou et al., 2019). The genomes in plants are however found to be large, repetitive, polyploid and structurally diverse, making their analysis computationally intensive and analytically intractable.

Conventional bioinformatics methods, though useful to simpler tasks, often make use of fixed rules and manual capabilities that are unsuitable to elicit complicated, biological relations and understandings. Machine learning (ML) provides an all-purpose, data-driven representation, which has the capability to detect subtle trends in multi-dimensional genomic data. ML methods have become more popular in the analysis of plant genomes during the last decade to predict genes, genomically annotate them, find regulatory elements, conduct genome-wide association studies (GWAS), and genomic selection etc. (Van Dijk et al., 2021).

This review highlights the importance of machine learning strategies that are applied in the analysis of plant genomes and with particular attention to new developments, methodological advances, and primary issues. Through a literature review, we emphasize the transformation of plant genomics by ML. We discuss the ways in which the current strategies can offer future projections to its successful applications to plant biology and agriculture.

Fundamentals of Machine Learning in Genomics

Machine learning is a form of computational technique that allows the machine to acquire patterns through data and enhance prediction through data without specific programming. In genomics, genomics ML models are usually trained on sequence data including genomic features, expression profiles or phenotypic measurements.

Supervised Learning

The supervised learning involves the use of labeled data to train predictive models. Support vector machines (SVM), random forest (RF), logistic regression, k-nearest neighbors (KNN), and artificial neural network (ANN) are among the algorithms that are extensively being applied in plant genomics. These procedures are normally used in prediction of genes, functional annotation and phenotype classification.

Unsupervised Learning

Unsupervised learning determines inherent trends in unlabeled information. To investigate population structure, gene expression patterns, and genetic diversity in plant species, k-means, hierarchical clustering, self-organizing maps (SOM) and principal component analysis (PCA) among others are frequently used clustering and dimensionality reduction methods.

Deep Learning

Deep learning employs multi-layer neural networks capable of automatic feature extraction from raw data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures have shown remarkable performance in sequence-based tasks, including promoter prediction, splice site identification, and regulatory element detection in plant genomes. Min, S., Lee, B., & Yoon, S. (2017) provided an overview of Deep learning approaches in bioinformatics. Ubbens, J.R et al.(2017) created Deep Plant Phenomics (DPP) platform for plant phenotyping community. Talukder A, et al.(2021), provided insights on nterpretation of deep learning in genomics and epigenomics.

Applications of Machine Learning in Plant Genome Analysis

Gene Prediction and Structural Genome Annotation

Gene prediction is initial step in genome analysis. ML-based approaches combine sequence composition, codon usage bias, comparative genomics, and transcriptomic evidence to differentiate coding from non-coding regions. Deep learning models show significant sensitivity and specificity in comparison to traditional hidden Markov model (HMM) based methods, particularly in complex plant genomes with abundant repetitive elements.

Functional Genomics and Transcriptome Analysis

Machine learning has become very important in functional genomics for analyzing transcriptomic datasets to identify differentially expressed genes, infer gene functions, and uncover co-expression networks. Clustering algorithms help group genes with similar expression patterns, while supervised models predict gene functions based on expression and sequence features.

Genome-Wide Association Studies (GWAS)

GWAS aims to identify genetic variants associated with phenotypic traits. ML methods enhance GWAS by managing high-dimensional genotype data and capturing nonlinear interactions between loci. Ensemble methods such as random forests and gradient boosting have been successfully applied to identify genomic regions controlling yield, quality traits, and stress tolerance in crops.

Epigenomics and Regulatory Element Prediction

Epigenetic modifications play a critical role in gene regulation. ML approaches are increasingly used to predict DNA methylation patterns, histone modifications, promoters, enhancers, and transcription factor binding sites. Deep learning models trained on sequence and epigenomic data have significantly improved the identification of regulatory elements in plant genomes.

Genomic Selection and Crop Improvement

In plant breeding, ML supports genomic selection by predicting breeding values using genomic and phenotypic data. These predictive models accelerate selection cycles, reduce breeding costs, and improve genetic gain. ML-driven genomic selection is particularly valuable for complex traits influenced by multiple genes and environmental factors. Murmu S. et al., (2024) focused on use of artificial intelligence-assisted omics techniques in plant defense. Ibrahim Raji and Tochukwu Kennedy Njoku, (2024) highlighted significance of data-driven decision making in agriculture: enhancing productivity and sustainability through predictive analytics.

Integration of Multi-Omics Data

Recent advances emphasize the integration of genomics with transcriptomics, proteomics, metabolomics, and phenomics data (Swati et al. 2012). Machine learning provides a flexible framework for multi-omics integration, enabling a systems-level understanding of plant biology. Deep learning and network-based approaches facilitate the identification of key regulatory pathways and trait-associated molecular signatures.

Recent Advances in Machine Learning for Plant Genomics

The application of deep learning architectures, transfer learning, and ensemble models represents a major advance in plant genomics. Pre-trained models and cross-species learning approaches are helping address data scarcity in non-model plants. Additionally, explainable AI techniques are emerging to improve model transparency and biological interpretability. Recently Montesinos-López A, et al. (2024) used deep learning methods improve genomic prediction of wheat breeding, Washburn, J. D., et al. (2021) used predictive breeding with machine learning, Crossa, J., et al. (2017) found it useful for genomic selection in plant breeding, Van Dijk, A. D. et al. (2021) summarized machine learning methods used in plant science and plant breeding. He J, et al. (2025) used machine learning and bioinformatics analysis to identify drought stress responsive genes in wheat. Chien CH, et al. (2021) used machine learning approaches to predict target gene expression in rice T-DNA insertional mutants.

Challenges and Limitations

Despite significant advances in ML approaches, several challenges still persist i.e. limited availability of high-quality labeled datasets, complexity of polyploids and presence of repetitive elements in plant genomes, lack of interpretability in deep learning models, High computational and infrastructure requirements, limited transferability of models across species.

Future Perspectives

Future efforts shall focus on developing interpretable and easy to use ML models having standardized benchmarking datasets, and closer integration with experimental validation. Collaborative initiatives combining computational scientists, plant biologists, and breeders will be essential for translating ML-based predictions into real-world agricultural solutions.

Conclusion

Machine learning has emerged as a powerful paradigm for plant genome analysis, enabling efficient interpretation of complex genomic data. Continued methodological innovation, coupled with improved data resources and interpretability, will further enhance the impact of ML in plant genomics and sustainable agriculture.

Table 1. Common machine learning algorithms and their applications in plant genome analysis

Machine Learning Approach	Algorithm	Genomic Application	Advantages	Limitations
Supervised learning	Support Vector Machine (SVM)	Gene prediction, promoter identification	High accuracy with small datasets	Sensitive to parameter tuning
Supervised learning	Random Forest (RF)	GWAS, trait prediction	Handles high-dimensional data, robust	Limited interpretability
Supervised learning	Artificial Neural Network (ANN)	Phenotype prediction, functional annotation	Models complex nonlinear relationships	Requires large datasets
Unsupervised learning	K-means clustering	Gene expression clustering	Simple and fast	Requires predefined cluster number
Unsupervised learning	PCA	Population genomics, diversity analysis	Reduces dimension-ality	Loss of biological interpreta-bility
Deep learning	CNN	Promoter and enhancer prediction	Automatic feature extraction	Computationally intensive
Deep learning	RNN/LSTM	Sequence motif detection	Captures sequential dependencies	Training complexity

Table 2. Applications of deep learning models in plant genomics

Deep Learning Model	Input Data	Plant Genomic Application	Representative Outcome
CNN	DNA sequence	Promoter and splice site prediction	Improved annotation accuracy
CNN	Epigenomic profiles	Regulatory element detection	Better regulatory mapping
RNN/LSTM	RNA-seq data	Gene expression prediction	Temporal expression modeling
Autoencoders	Multi-omics data	Data integration	Noise reduction and feature learning
Transformer models	Long genomic sequences	Regulatory grammar learning	Long-range dependency detection

.

References:

1. Chen, X., Li, Q., & Wang, J. (2020). Machine learning approaches in plant genomics and breeding. *Plant Biotechnology Journal*, 18, 1–15.
2. Chien CH, Huang LY, Lo SF, Chen LJ, Liao CC, Chen JJ, Chu YW. Using Machine Learning Approaches to Predict Target Gene Expression in Rice T-DNA

Insertional Mutants. Front Genet. 2021 Dec 17;12:798107. doi: 10.3389/fgene.2021.798107.

3. Crossa, J., et al. (2017). Genomic selection in plant breeding. *Crop Science*, 57, 1–17.
4. He J, Cui B, Liu P, Meng X and Yan J (2025) Utilizing machine learning and bioinformatics analysis to identify drought stress responsive genes in wheat (Triticum aestivum L.). Front. Sustain. Food Syst. 9:1612009. doi: 10.3389/fsufs.2025.1612009
5. Ibrahim Raji and Tochukwu Kennedy Njoku, (2024) Data-Driven Decision Making in Agriculture: Enhancing Productivity and Sustainability through Predictive Analytics. *International Journal of Research Publication and Reviews*, Vol 5, no 9, pp 2708-2719 September 2024
6. Kaur, H., et al. (2020). Machine learning for stress tolerance in crops. *Computational and Structural Biotechnology Journal*, 18, 3486–3499.
7. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16, 321–332.
8. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18, 851–869.
9. Montesinos-López A, et al. (2024) Deep learning methods improve genomic prediction of wheat breeding. Front. Plant Sci. 15:1324090. doi: 10.3389/fpls.2024.1324090
10. Murmu S, Sinha D, Chaurasia H, Sharma S, Das R, Jha GK and Archak S (2024) A review of artificial intelligence-assisted omics techniques in plant defense: current trends and future directions. Front. Plant Sci. 15:1292054.
11. Swati Singh, Sanchita Gupta , Ashutosh Mani , Anoop Chaturvedi. Mining and gene ontology based annotation of SSR markers from expressed sequence tags of Humulus lupulus. Bioinformation (2012) Feb 3;8(3):114–122.
12. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2021 May 20;22(3):bbaa177. doi: 10.1093/bib/bbaa177.
13. Ubbens, J. R., & Stavness, I. (2017). Deep plant phenomics. *Frontiers in Plant Science*, 8, 1191.
14. Van Dijk, A. D. J., et al. (2021). Machine learning in plant science. *Plant Physiology*, 185, 1413–1430.
15. Washburn, J. D., et al. (2021). Predictive breeding with machine learning. *Trends in Plant Science*, 26, 31–43.
16. Zou, J., Huss, M., Abid, A., et al. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51, 12–18.