Artificial Intelligence and Machine Learning in Breeding Programs

V. Sandeep Varma

Plant breeding is crucial for addressing global challenges like food security, climate change resilience, and sustainable agriculture. The integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques has revolutionized traditional breeding methods, enabling the development of improved crop varieties. AI and ML algorithms are used for tasks such as genotype-phenotype prediction, genomic selection, trait discovery, and optimization of breeding schemes. These technologies help identify genetic markers associated with desirable traits, enabling breeders to select plants with desired characteristics more efficiently. AI-driven models can predict the performance of novel genotypes under different environmental conditions, aiding in the development of resilient and high-yielding crop varieties. AI-powered tools can optimize breeding strategies by simulating breeding outcomes, reducing time and resource constraints. However, challenges such as data quality, model interpretability, and ethical considerations need to be addressed. Additionally, the accessibility of advanced computational resources and expertise remains a barrier for many breeders, especially in developing countries. The future of AI and ML in plant breeding holds great promise, with continued advancements in computational biology, genomics, and data analytics. Collaboration between breeders, data scientists, and biotechnologists is essential for leveraging AI and ML technologies to their full potential in addressing global agricultural challenges.

Keywords: *Artificial intelligence, Machine learning, Plant breeding.*

V. Sandeep Varma
Ph.D. Scholar, Department of Genetics and Plant Breeding, Agricultural College, Bapatla, Andhra Pradesh, India.
*Email: sandeepvunnam81@gmail.com

Publisher: Cornous Publications LLP, Puducherry, India.
Recent Advances in Plant Breeding (Volume 1)
Editors: Dr. Manoj Kumar Tripathi, Riya Mishra
ISBN: 978-81-976294-6-4
DOI: https://doi.org/10.37446/volbook102024/62-69

Introduction

Evolution of Artificial intelligence and machine learning in plant breeding

AI and ML initially served as data analysis tools in traditional breeding methodologies. With high-throughput sequencing technologies and genomic data growth, AI and ML have become indispensable assets in modern breeding programs. Early stages of AI and ML adoption in plant breeding involved statistical models and computational algorithms for genotype-phenotype prediction tasks. As AI and ML technologies advanced, breeders began exploring efficient techniques like deep learning algorithms for genotype-phenotype prediction and trait discovery. Deep learning models like CNNs and RNNs showed superior performance in analyzing large-scale genomic and phenotypic datasets.

Applications of artificial intelligence and machine learning in breeding programs

It includes

- Genotype-phenotype prediction
- Genomic selection
- Trait discovery and characterization
- Breeding scheme optimization

1. Genotype-phenotype prediction

Genomic data processing and feature extraction

Chen et al., (2018), Ringnér (2008), and other researchers have contributed significantly to the field of genomic data analysis. They have highlighted the importance of preprocessing in ensuring the quality and reliability of genomic data. The preprocessing process involves data cleaning, normalization, and feature extraction. These steps help to remove noise, errors, and inconsistencies, ensuring the data is comparable and suitable for downstream analyses. Feature selection and dimensionality reduction techniques are also crucial in genomic analysis, as they help identify the most relevant subset of features for a particular analysis. By selecting informative features, researchers can focus computational resources on variables that contribute most to the prediction or classification task, enhancing the efficiency and effectiveness of downstream analyses.

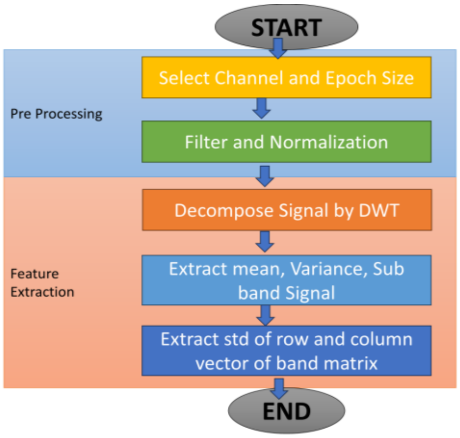
START
Pre Processing
Select Channel and Epoch Size
Filter and Normalization
Feature Extraction
Decompose Signal by DWT
Extract mean, Variance, Sub band Signal
Extract std of row and column vector of band matrix
END

Figure 1. Flow chart of preprocessing and feature extraction (Chum et al.,2011)

Prediction models and algorithms

Gianola et al., (2009) research on genotype-phenotype prediction focuses on traditional regression models like linear and logistic regression, which are widely used in association studies and quantitative trait locus mapping to understand the complex relationships between genetic variation and plant traits.

Types of machine learning algorithms

Random forests: It is a popular ensemble learning method, use the power of decision trees to predict phenotypic traits by aggregating predictions from multiple trees trained on different subsets of the data (Breiman, 2001).

Neural networks: It is particularly deep learning architectures, have emerged as powerful tools for capturing intricate genotype-phenotype relationships and extracting latent features from genomic data.

Convolutional neural networks (CNNs) and **recurrent neural networks (RNNs):** They have been adapted to genomic sequence data, enabling researchers to predict phenotypic outcomes based on DNA sequences, gene expression profiles, and epigenetic modifications (Alipanahi et al., 2015; Angermueller et al., 2016).

2. Genomic selection

AI and ML-based genomic prediction models

Genomic best linear unbiased prediction (GBLUP): a statistical approach that estimates breeding values by fitting a linear mixed model to genomic data. GBLUP uses genomic relationships among individuals, captured through marker-based kinship matrices, to estimate genetic effects across the genome while accounting for population structure and genetic relatedness (VanRaden, 2008).

Bayesian method: It represents another class of algorithms employed in genomic prediction, offering flexibility and the ability to incorporate prior knowledge and uncertainty into the modeling framework. Bayesian regression models, such as Bayesian LASSO and Bayesian Ridge Regression, enable the estimation of marker effects and shrinkage of coefficients, effectively capturing the polygenic architecture of traits while avoiding overfitting (Habier et al., 2007).

Integration of phenotypic and genomic data

Jiang, and Reif (2015) have explored the integration of phenotypic and genomic data in plant breeding programs. They proposed the Genomic Best Linear Unbiased Prediction (GBLUP) framework, which integrates phenotypic measurements as fixed effects in prediction models, enhancing prediction accuracy by capturing additional sources of variation and reducing residual error. Another strategy is multi-trait prediction models, which jointly predict breeding values for multiple correlated traits using genomic and phenotypic information (Figure 2). These models exploit genetic correlations among traits to improve prediction accuracy and facilitate multi-trait selection strategies. The high-throughput phenotypic and genotypic data collected from large crop germplasm and breeding populations can be integrated with AI technology, such as phenotypic diversity, SNPs polymorphisms, QTL analysis, GWAS analysis, genomics selection, and genome sequence. AI technologies are applied to predict crop phenotype and produce novel breeding strategies through computation and training models.

Case studies demonstrate the successful application of genomic selection (GS) in crop improvement:

Maize breeding: In a study by Rincent et al., (2014), GS successfully improved maize breeding program yield and stress resistance by integrating genomic and phenotypic data, leading to the development of elite hybrids with superior performance.

Wheat breeding: A study by Rutkoski et al., (2012) demonstrated GS effectively predicted grain yield and Fusarium head blight resistance in wheat breeding populations, resulting in significant gains compared to conventional breeding methods.

Rice breeding: Researchers and breeders successfully developed blast-resistant rice varieties using genomic information, reducing disease yield and ensuring food security in rice-dependent regions through collaborative efforts. (Spindel et al.,2015)

Horticultural breeding: In a study by Kumar et al., (2012), GS was used to predict breeding values for fruit firmness and acidity in apple breeding populations, leading to the development of new apple cultivars with improved post-harvest storage characteristics and consumer appeal.

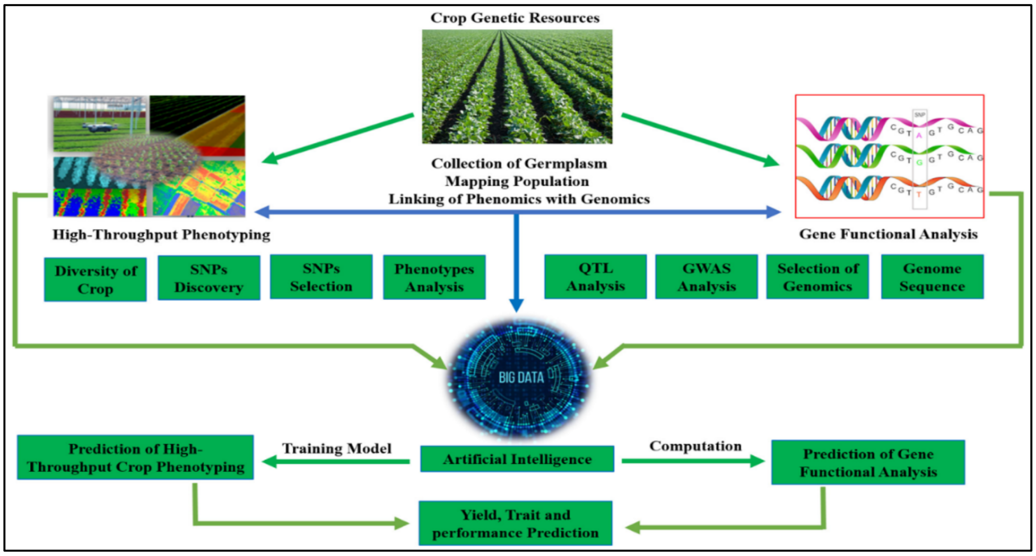

Figure 2. Artificial Intelligence used as a powerful tool for the prediction of high-throughput crop phenotyping and gene functional analysis in modern crop breeding. (Khan et al., 2022)

3. Trait discovery and characterization

Identification of genetic markers: GWAS, a statistical method, uses AI algorithms like logistic regression, random forests, and gradient boosting to identify genetic markers associated with desirable traits. Deep learning architectures like Convolutional Neural Networks and Recurrent Neural Networks help analyze genomic sequences and identify regulatory elements. By integrating AI and ML techniques into genetic marker discovery pipelines, breeders can accelerate the identification of key genetic loci, enabling targeted selection and genomic prediction in breeding programs.

Multi-Omics integration: The integration of multi-omics data, including genomics, transcriptomics, metabolomics, and epigenomics, offers a comprehensive approach to trait characterization in plant breeding, enabling a deeper understanding of the molecular mechanisms underlying complex traits and facilitating more precise trait prediction and selection like abiotic stress tolerance (Figure 3.). These data layers enable a comprehensive understanding of complex traits, enabling precise trait prediction and selection. Hirsch et al., (2014) work on genome-wide association studies (GWAS) and linkage mapping helps identify candidate genes and genetic markers associated with target traits. Shi et al., (2020) work on transcriptional changes and regulatory networks provides insights into gene regulatory mechanisms. Saito and Matsuda (2010) work on metabolic profiles and biochemical pathways reveals metabolic signatures associated with trait phenotypes. Epigenomic data provides insights into DNA methylation, histone modifications, and chromatin accessibility.

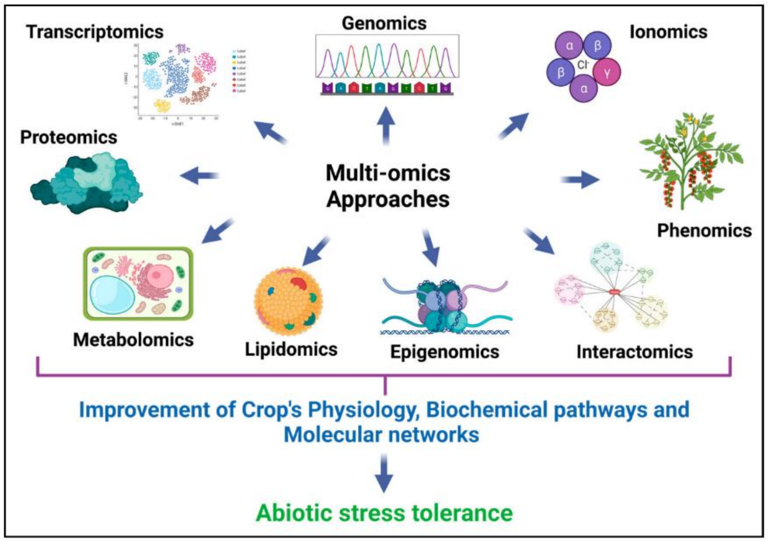

Figure 3. Integrative multi-omics approaches to confer abiotic stress tolerance in plants. (Roychowdhury et al.,2023)

4. Breeding scheme optimization

Simulation and optimization models

Simulation models: Simulation models play a crucial role in designing and optimizing breeding schemes, allowing breeders to explore different breeding strategies, evaluate their performance, and identify optimal decision pathways to achieve breeding objectives efficiently.

Optimization models: These models utilize mathematical programming and algorithmic approaches to identify optimal breeding strategies and decision pathways that maximize genetic gain while satisfying breeding constraints and objectives (Crossa et al., 2017). Optimization models formulate breeding problems as mathematical optimization problems, where decision variables represent breeding actions (e.g., selection intensity, mating designs) and objective functions quantify breeding goals (e.g., maximizing genetic gain, minimizing breeding costs) (Whishart et al., 2019).

Accelerating breeding cycles

Strategies for accelerating breeding cycles using AI and ML techniques, such as speed breeding, marker-assisted selection (MAS), and genomic prediction, have become pivotal in modern plant breeding programs.

Speed breeding: Watson et al., (2018) study on speed breeding, utilizing AI and ML algorithms, demonstrates its effectiveness in shortening generation times and accelerating breeding cycles. This technique automates data collection, extracts informative traits, and expedites selection decisions, reducing time for variety development.

Marker-assisted selection (MAS): AI and ML techniques enhance MAS by improving marker-trait association analysis, optimizing marker selection, and predicting phenotypic performance based on genomic data (Crossa et al., 2017). By integrating MAS with genomic prediction models, breeders can prioritize marker-assisted crosses, accelerate trait introgression, and enhance selection accuracy, thereby shortening breeding cycles and accelerating genetic gain.

Genomic prediction: Genomic prediction uses genome-wide molecular markers to predict breeding values and select superior individuals for breeding, without the need for extensive phenotyping (Heffner et al.,2009). AI and ML algorithms enhance genomic prediction by capturing complex genotype-phenotype relationships, improving prediction accuracy, and accelerating breeding progress (Hickey et al.,2017). By incorporating advanced machine learning models, such as deep learning architectures and Bayesian methods, genomic prediction enables breeders to expedite selection decisions, reduce generation intervals, and accelerate variety development in plant breeding programs. In a study by Jarquín et al., (2014), By integrating genomic prediction models with multi-environment trials and field data, breeders were able to accurately predict breeding values for yield and other agronomic traits in soybean breeding populations.

5. Beyond traditional plant breeding

Robotics and automation in phenotyping: Robotics and automation play a crucial role in advancing high-throughput phenotyping (HTP) and data collection in plant breeding and agricultural research. They enable researchers to efficiently collect large volumes of phenotypic data from plant populations, accelerating breeding progress and enhancing the accuracy of trait evaluations.

High- throughput phenotyping: Automated imaging platforms equipped with high-resolution cameras, sensors, and robotic systems can capture detailed phenotypic information, such as plant growth dynamics, canopy architecture, leaf morphology, and physiological traits, with high precision and throughput (Paulus et al., 2014).

Data collection: Robotics and automation technologies are instrumental in streamlining data collection workflows and reducing manual labor in field and laboratory settings. Autonomous vehicles, drones, and robotic systems equipped with sensors and actuators can navigate field environments, collect samples, and perform measurements with minimal human intervention (Haghighattalab et al., 2016).

Controlled environment phenotyping (CEP): Robotics and automation enable the development of controlled environment phenotyping (CEP) facilities like Automated growth chambers, greenhouses, and phenotyping platforms equipped with environmental sensors and robotic systems enable researchers to conduct reproducible experiments under controlled conditions, facilitating the study of genotype-environment interactions and the characterization of plant responses to abiotic and biotic stresses.

Conclusion

AI and ML have the potential to revolutionize plant breeding and address global agricultural challenges. These technologies enable breeders to analyze vast amounts of genetic and phenotypic data more efficiently and accurately than traditional methods, facilitating genotype-phenotype prediction, trait discovery, and breeding value estimation with greater precision and speed (Miotto et al., 2018). AI-driven breeding approaches can develop crop varieties with improved stress tolerance, disease resistance, and nutritional quality, ensuring food security and sustainability in the face of changing climatic conditions and evolving pest pressures (Tallis et al., 2018). Moreover, AI and ML have the potential to democratize access to breeding tools and resources, empowering farmers, particularly in developing countries, to participate in crop improvement efforts and benefit from technological innovations. By harnessing these technologies, researchers, breeders, and policymakers can accelerate innovation, foster resilience, and ensure food security for future generations.

References

Alipanahi, B., Delong, A., Weirauch, M.T., & Frey, B.J. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology, 33*(8):831-838.

Angermueller, C., Parnamaa, T., Parts, L., & Stegle, O. 2016. Deep learning for computational biology. *Molecular Systems biology, 12*(7):878.

Breiman, L. 2001. Random forests. *Machine learning, 45*(1):5-32.

Chen, Y., Wang, Z., Li, Y., Truong, E., & Banerjee, A. 2018. Meta-transfer learning for few-shot learning. *Proceedings of the 35th International Conference on Machine Learning.* Stockholm, Sweden. 10-15 July 2018. 80:4343-4352.

Chum, P., Park, S., Ko, K., & Sim, K. 2011. Optimal EEG Feature Extraction using DWT for Classification of Imagination of Hands Movement. *Journal of Korean Institute of Intelligent Systems, 21*:786-791.

Crossa, J., Perez, P., Cuevas, J., Lopez, M, O., Jarquín, D., De los Campos, G., & Burgueno, J. 2017. Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science, 22*(11): 961-975.

Gianola, D., Okut, H., Weigel, K.A., Rosa, G.J., & Bacheller, L.R. 2009. Predicting complex quantitative traits with Bayesian neural networks: A Case Study with Jersey cows and wheat. *BMC Genetics, 10*(1):37.

Habier, D., Fernando, R.L., Kizilkaya, K., & Garrick, D.J. 2007. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics, 8*(1): 1-15.

Haghighattalab, A., Gonzalez Perez, L., Mondal, S., Singh, D., Schinstock, D., Rutkoski, J., & Poland, J. 2016. Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. *Plant Methods, 12*(1): 35.

Heffner, E.L., Jannink, J.L., & Sorrells, M.E. 2009. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome, 2*(2):191-197.

Hickey, L.T., Hafeez, A.N., Robinson, H., Jackson, S.A., Leal-Bertioli, S.C.M., Tester, M., & Dieters, M.J. 2017. Breeding crops to feed 10 billion. *Nature Biotechnology, 35*(10): 927-937.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., & de Leon, N. 2014. Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell, 26*(1): 121-135.

Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., & Gay, L. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics, 127*(3):595-607.

Jiang, Y., & Reif, J.C. 2015. Modeling epistasis in genomic selection. *Genetics, 201*(2): 759-768.

Khan, M.H.U., Wang, S., Wang, J., Ahmar, S., Saeed, S., Khan, S.U., Xu, X., Chen, H., Bhat, J.A., & Feng, X. 2022. Applications of Artificial Intelligence in Climate-Resilient Smart-Crop Breeding. *International Journal of Molecular Sciences, 23*: 11156

Kuang, Z., Ping, Y., Hao, Y., Fang, Z., Li, J., & Yin, H. 2019. Highly efficient RNA-guided base editing in rabbit. *Nature Communications*, *10*(1): 1-10.

Kumar, S., Garrick, D.J., Bink, M.C., Whitworth, C., & Chagne, D. 2012. Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). *PloS One*, *7*(5):36674.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J.T. 2018. Deep learning for healthcare: Review, Opportunities and Challenges. *Briefings in Bioinformatics*, *19*(6):1236-1246.

Paulus, S., Dupuis, J., Mahlein, A.K., Kuhlmann, H., & Kersting, K. 2014. Automatic early plant disease detection using machine learning-based image analysis. *Plant Pathology*, *63*(6):1302-1312.

Rincent, R., Charpentier, J.P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., & Moreau, L. 2014. Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3: Genes, Genomes, Genetics*, *4*(8):1603-1610.

Ringner, M. 2008. What is principal component analysis?. *Nature Biotechnology*, *26*(3):303-304.

Roychowdhury, R., Das S.P., Gupta, A., Parihar, P., Chandrasekhar, K., Sarker, U., Kumar, A., Ramrao, D.P., & Sudhakar, C. 2023. Multi-Omics Pipeline and Omics-Integration Approach to Decipher Plant's Abiotic Stress Tolerance Responses. *Genes*, *14*(6):1281.

Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J. L., & Sorrells, M. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *Plant Genome*. 5(2):51-61.

Saito, K., & Matsuda, F. 2010. Metabolomics for functional genomics, systems biology, and biotechnology. *Annual Review of Plant Biology*, *61*:463-489.

Shi, T., Shi, L., Fang, H., Weng, Z., & Schadt, E.E. 2020. Investigating and suppressing batch effects in single-cell RNA-Seq data. *Genome Biology, 21*(1):1-19.

Spindel, J., Begum, H., Akdemir, D., Collard, B., Redona, E., Jannink, J. L., & McCouch, S.R. 2015. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4): 395-408.

Tallis, H., Kreis, K., O'Hare, M., O'Connell, D., Hawkins, E., Folarin, A., & Rahman, M. 2018. The role of digital agriculture in food security. London: The Nature Conservancy.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11):4414-4423.

Watson, A., Ghosh, S., Williams, M.J., Cuddy, W.S., Simmonds, J., Rey, M.D., & Hickey, L.T. 2018. Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants, 4*(1):23-29.

Whishart, J., Arief, V.N., & Smith, A.B. 2019. Operations research in agriculture: A review. *Computers and Electronics in Agriculture, 157*:8-27.