Creation of an Integrated Evaluation Algorithm to Genetic Divergence Analysis in Rice (*Oryza sativa* L.)

Shubham Kumar[1], Varun Bansal[1] and Shivani[2]
[1]*Department of Computer Science & Engineering, Shobhit University, Gangoh, India*
[2]*School of Agriculture & Environmental Sciences, Shobhit University, Gangoh, India*

***Abstract*—**Appropriate crop development programs depend on accurate evaluation of genetic differentiation. Although more conventional statistical techniques, such as the principal component test and Mahalanobis D^2 test, are widely used, there has not been much work done to incorporate them into systematic computing programs in the subject of agriculture. In this work, a sophisticated algorithm for evaluating multivariate genetic divergence of rice (*Oryza sativa* L.) is presented and justified. Twenty-six rice genotypes were examined for morphological, yield, and quality characteristics in the field. The approach computes the objective genotype ranking and estimates the divergence by combining principal component extraction, cluster analysis, and path coefficient modelling into a single procedural chain. The results showed a considerable degree of variance among genotypes, and the harvest index, days to flowering, and spikelets per panicle all contributed significantly to overall divergence. The proposed evaluation approach enhances reproducibility, lessens the impact of computation, and helps with data-driven parent selection in breeding projects.

***Keywords*—**Genetic divergence; Evaluation algorithm; Rice; Principal component analysis; Cluster analysis; Yield traits.

I. INTRODUCTION

Since rice (Oryza sativa L.) provides the primary food supply for more than half of the world's population, it is one of the most strategically important cereal crops in the world. Cytogenetically, rice is a diploid species (2n = 24) that self-pollinates and belongs to the Poaceae family. The Oryza genus includes two domesticated species — O. sativa (Asian rice) and O. glaberrima (African rice) — and several wild members that offer a promising source of genetic diversity in stress tolerance, yield advancement, and adaptation evolution [1], [2].

Genetic variety is the cornerstone of crop development initiatives. Heterotic hybrids and transgressive segregants are more likely to result from genetically diverse parental lines. Breeders can quantify genetic diversity to assess the degree of variability in germplasm collections, as well as their distribution and structure. One of the effective multivariate methods used is multivariate Mahalanobis D^2 statistics, which takes into account covariance structure across characteristics in addition to evaluating genetic distance [5], [6].

The analysis of diversity is further supported by secondary statistical techniques such as Principal Component Analysis (PCA) and hierarchical clustering. By transforming correlated variables into orthogonal variables that best explain maximum variance, PCA lowers dimensional complexity [9]. A methodical selection of heterogeneous parents is made possible by cluster analysis, which uses multivariate similarity indexes to classify genotypes into relatively homogeneous groupings [11]. Additionally, path coefficient analysis separates correlation coefficients into direct and indirect effects, making it possible to identify the characteristics that actually have causal effects on yield [12], [13].

Despite their strength in analysis, these statistical procedures are frequently applied independently, which results in a fragmented interpretation and decreased reproducibility. Coordinated computation systems that can integrate conventional biometric technologies into a single analytic pipeline are necessary due to the growing complexity of high-dimensional phenotypic data. Algorithm-based systems promote analytical transparency by offering standardised pretreatment, automated covariance estimation, objective grouping logic, and repeatable genotype ranking [15].

Modern breeding techniques have been redefined by the integration of computer science and agricultural sciences. Crop yields, phenotypic stress, genotype-environment interaction, and genomic selection are all predicted using techniques based on data mining, pattern recognition, machine learning, and artificial intelligence [16]–[18]. The significance of open algorithmic forms in agricultural practice is further highlighted by recent advances in explainable artificial intelligence (XAI) and interpretable machine learning [20].

Hence, the current study aims at designing and developing an integrated assessment algorithm to study the genetic divergence in rice agricultural genetic research. The suggested framework integrates Mahalanobis D^2 statistics, Principal Component

Analysis, hierarchical clustering and path coefficient modelling into a common set of computations. The algorithm is validated with multivariate field data on a variety of rice genotypes to prove its applicability to objective genotype classification and parent selection [25].

II. MATERIALS AND METHODS

A. Experimental Materials and Field Evaluation

The experimental material comprised twenty-six genetically different rice (Oryza sativa L.) genotypes (Table 1), selected to represent wide genetic variation from various agro-ecological areas. The field experiment was carried out in the Kharif season in irrigated lowlands, designed as a Randomized Block Design (RBD) with three repetitions to reduce environmental variation and increase statistical precision.

Table 1: List of cultivated rice varieties used in present investigation

S. No.	Name of Variety	S. No.	Name of Variety
1.	NDR-97	14.	Taramon
2.	Pusa Basmati-1	15.	IR 83668-35-2-2-2
3.	NDR-118	16.	Ayaar
4.	NDR-359	17.	IR 68144-2B-2-2-3-1-127
5.	Shusk Samrat	18.	Ngobanyo Red Cover
6.	NDR-1	19.	IR 91167-133-1-1-2-3
7.	Nagina-22	20.	Maigothi
8.	Kuhusoi-Ri-Sareku	21.	Sarjoo-52
9.	IR 91167-31-3-1-33	22.	Barani Deep
10.	R-RHZ-2	23.	IR-64
11.	IR 92960-75-1-3	24.	Gopalbhog (Local)
12.	IR 68144-2B-2-2-3-1-120	25.	Nedu
13.	Amker	26.	Saponyo

Eleven quantitative traits were measured: seedling vigor (visual score), days to 50% flowering, plant height (cm), panicle bearing tillers per plant, spikelets per panicle, grains per panicle, spikelet fertility percentage, test weight (g), biological yield per plant (g), harvest index (%), and grain yield per plant (g). Data were taken on five competitive randomly selected plants per plot and averaged for statistical analysis.

B. Statistical Model of Genetic Divergence Analysis

The experimental data were analyzed using ANOVA before computational modeling to ensure significant genetic variation across genotypes. Multivariate analysis was done using mean values. Standardization was performed to normalize the dataset and remove scale bias. Mahalanobis D^2 was used to determine the generalized genetic distance between genotypes:

$$D^2 = (X_i - X_j)' S^{-1} (X_i - X_j)$$

where X_i and X_j are standardized trait vectors of the i-th and j-th genotypes respectively, and S^{-1} is the inverse of the pooled covariance matrix. Hierarchical clustering (Ward minimum variance method) was applied using D^2 values to identify genotype clusters. PCA was conducted to find the relative contribution of individual traits to overall variability. Path coefficient analysis was conducted with grain yield per plant as the dependent variable.

C. Development of the Integrated Evaluation Algorithm

An organized computational pipeline called the Agricultural Genetic Divergence Evaluation Algorithm (AGDEA) was created to increase reproducibility and reduce interpretational bias. The algorithm combines classical biometric procedures into consecutive computational steps.

Algorithm: Agricultural Genetic Divergence Evaluation Algorithm (AGDEA)

Input: Standardized Trait Matrix T (26×11)

Output: Genotype divergence index, cluster classification and trait contribution rank

Stage 1: Data Preprocessing

- Impute any missing values with mean imputation.
- Apply Z-score normalization to trait score matrices.

Stage 2: Covariance Structure Estimation

- Compute pooled covariance matrix.
- Invert covariance matrix.

Stage 3: Genetic Distance Computation

- Calculate pairwise Mahalanobis D^2 distances.
- Construct distance matrix.

Stage 4: Cluster Formation

- Apply hierarchical clustering (Ward's method).
- Identify optimum number of clusters using dendrogram.

Stage 5: Dimensionality Reduction

- Conduct Principal Component Analysis.
- Extract eigenvalues and eigenvectors.
- Compute percentage of variance explained.

Stage 6: Causal Modelling

- Calculate genotypic correlation matrix.
- Run path coefficient analysis.
- Partition direct and indirect effects.

Stage 7: Trait Contribution Estimation

- Calculate percent contribution of each trait to total divergence.
- Rank characteristics according to magnitude of contributions.

Stage 8: Divergence Index Generation

- Combine D^2 distances, PCA loadings and path coefficients.
- Compute Genotype Divergence Index (GDI).
- Rank genotypes for breeding suitability.

End Algorithm

D. Computational Implementation

The algorithm was applied with the help of matrix-based statistical computation procedures compatible with Python (NumPy, SciPy, Scikit-learn libraries) and the R statistical environment. Covariance inversion, eigen decomposition, clustering and path modelling steps were automated for reproducibility. Integrating these multivariate statistical modules into a single calculatory system decreases subjectivity in interpretation and increases analytical clarity.

E. Validation and Reproducibility

To prove robustness, internal consistency of cluster formation was assessed using cophenetic correlation coefficients. PCA results were cross-verified with scree plot analysis. The modular design of the algorithm enables it to scale up to larger genotype datasets and be extended to other crops with relatively simple structural changes.

III. RESULTS

A. Path Coefficient Analysis

Path coefficient analysis was used to partition genotypic correlation coefficients into direct and indirect effects to determine traits with real causal effects on grain yield per plant. Table 2 shows the estimates of direct and indirect effects at the phenotypic level.

Table 2: Direct and indirect effects for different characters on grain yield per plant at phenotypic level in rice genotypes

S. No	Characters	SV	DTF	PH	PBT	SP	GP	SF	TW	BY	HI	GY
1	SV	**0.026**	0.009	0.018	-0.004	0.002	0.003	0.007	0.002	0.011	0.003	0.315
2	DTF	-0.015	**-0.041**	-0.010	-0.001	0.001	0.002	0.005	-0.003	-0.005	0.007	-0.076
3	PH	-0.015	-0.006	**-0.022**	0.002	-0.005	-0.006	-0.005	-0.004	-0.007	-0.004	0.310
4	PBT	0.003	0.000	0.002	**-0.017**	-0.003	-0.003	0.001	-0.001	-0.002	0.008	-0.301
5	SP	0.021	-0.005	0.083	0.060	**0.356**	0.348	-0.007	0.054	-0.015	0.062	0.127
6	GP	-0.037	0.014	-0.085	-0.050	-0.316	**-0.323**	-0.060	-0.044	-0.004	-0.064	0.172
7	SF	0.009	-0.005	0.008	-0.001	-0.001	0.007	**0.035**	-0.002	0.008	0.005	0.216
8	TW	-0.003	-0.003	-0.008	-0.002	-0.006	-0.006	0.002	**-0.042**	-0.005	-0.014	0.263
9	BY	0.251	0.079	0.194	0.054	-0.026	0.008	0.145	0.068	**0.619**	0.075	0.688
10	HI	0.076	-0.118	0.130	-0.342	0.124	0.142	0.093	0.234	0.086	**0.714**	0.791

SV– Seedling vigor; DTF– Days to 50% flowering; PH– Plant height; PBT– Panicle bearing tillers/plant; SP– Spikelets/panicle; GP– Grains/panicle; SF– Spikelet fertility (%); TW– Test weight (g); BY– Biological yield/plant (g); HI– Harvest index (%); GY– Grain yield/plant (g)

The findings revealed that biological yield per plant had the highest positive direct impact on grain yield, followed by harvest index, making these the main determinants of yield. Spikelets per panicle also showed a significant positive direct contribution. The residual effect in the path model is relatively low, implying that the selected traits adequately described most of the variability in grain yield.

B. Principal Component Analysis (PCA)

PCA was applied to decrease dimensional complexity and determine crucial sources of variability among the analyzed genotypes. Table 3 displays eigenvalue estimates and coefficients of the first four principal component vectors.

Table 3: Estimates of canonical roots and coefficient of first four vectors (PCA)

Trait	PC-I	PC-II	PC-III	PC-IV
Eigenvalue (Root)	3.80	2.67	1.67	0.87
% Var. Exp.	34.56	24.29	15.17	7.87
Cum. Var. Exp.	34.56	58.85	74.02	81.89
Seedling Vigor (cm)	0.28	0.41	0.03	0.07
Days to 50% Flowering (days)	0.15	0.19	0.62	-0.26
Plant Height (cm)	0.37	0.32	-0.01	-0.12
Panicle Bearing Tillers/ Plant	-0.09	-0.34	-0.28	-0.54
Spikelets/ Panicle	0.41	-0.30	-0.11	0.11
Grains/ Panicle	0.45	-0.16	-0.26	0.13
Spikelet Fertility (%)	-0.46	0.10	0.24	-0.12
Test Weight (gm)	-0.07	0.27	-0.44	-0.60
Biological Yield/ Plant (gm)	0.18	0.46	-0.05	-0.13
Harvest Index (%)	-0.16	0.36	-0.40	0.28
Grain Yield/ Plant (gm)	-0.34	0.17	-0.20	0.36

The four major components explained approximately 81% of total phenotypic variability, confirming successful dimensional reduction. The first principal component (PC I) was largely related to spikelets per panicle and grains per panicle. Fig. 2 depicts the 3-D ordination of 26 rice varieties on principal component axes I and II, and Fig. 3 shows the 2-D PCA plot.

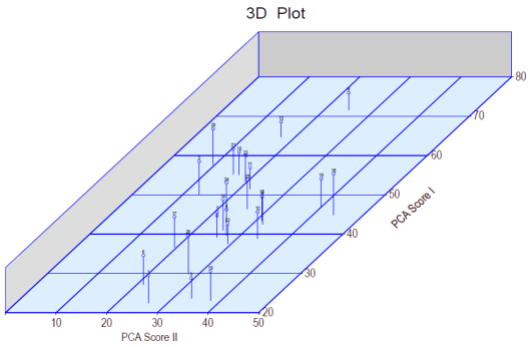

Fig. 2. Three-dimensional ordination of 26 rice varieties on principal component axes I & II.

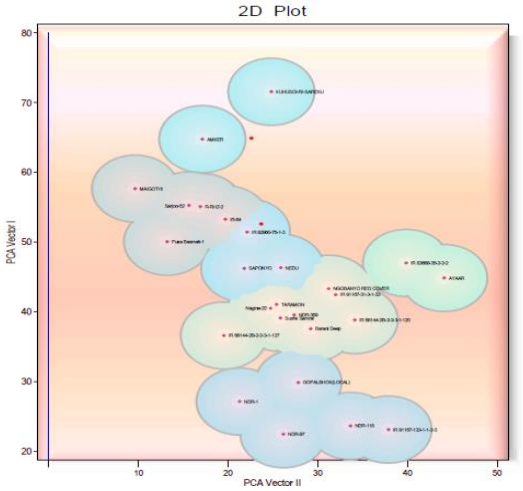

Fig. 3. Two-dimensional plot of 26 rice varieties on Principal Component Vector I and II.

The second major component (PC II) was defined by biological yield and harvest index, indicative of biomass partitioning and resource utilization efficiency. K-means clustering method was also used for genotype grouping, as depicted in Fig. 4.

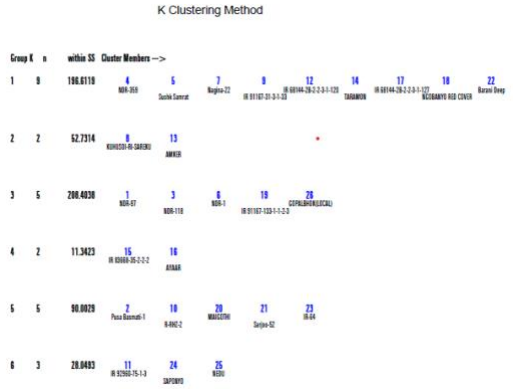

Fig. 4. K-clustering method showing grouping of rice genotypes based on multivariate trait analysis.

C. Cluster Analysis

The Mahalanobis D²-based hierarchical clustering technique revealed six genotype clusters among the twenty-six genotypes, indicating significant genetic heterogeneity. Table 4 shows the clustering pattern of the genotypes. The greatest inter-cluster distance was found between Cluster I and Cluster V, creating strong genetic variance and probable suitability for hybridization.

Table 4: Clustering pattern of 26 rice genotypes on the basis of D² analysis for 10 characters

Cluster No.	No. of Genotypes	Genotype
Cluster I	9	Saponyo, Nedu, IR 91167-31-3-1-33, Ngobanyo Red Cover, Taramon, IR 68144-2B-2-2-3-1-120, Shusk Samrat, Barani Deep, Nagina-22
Cluster II	6	R-RHZ-2, IR-64, Sarjoo-52, IR 92960-75-1-3, Amker, Kuhusoi-Ri-Sareku
Cluster III	8	NDR-97, NDR-1, Gopalbhog (Local), IR 68144-2B-2-2-3-1-127, NDR-118, IR 91167-133-1-1-2-3, NDR-359, IR 83668-35-2-2-2
Cluster IV	1	Pusa Basmati-1
Cluster V	1	Maigothi
Cluster VI	1	Ayaar

The average intra- and inter-cluster D² values are presented in Table 5.

Table 5: Estimation of average inter-cluster D² values

Cluster	I	II	III	IV	V	VI
CLUSTER I	1007.40	1625.04	1902.35	3437.11	7994.55	4410.00
CLUSTER II		541.57	818.01	2304.99	3837.43	2157.08
CLUSTER III			569.18	2626.22	3487.87	1390.72
CLUSTER IV				0.00	7209.28	4786.91
CLUSTER V					544.35	1811.89
CLUSTER VI						968.51

Table 6: Cluster means for different characters in rice genotypes

Cluster	SV (cm)	DTF (days)	PH (cm)	PBT/Plant	SP/Panicle	GP/Panicle	SF (%)	TW (gm)	BY/ (g
Cluster 1	29.005	89.467	83.647	9.600	57.533	46.400	81.306	19.253	38
Cluster 2	40.857	91.286	114.732	10.571	92.333	80.048	87.119	23.910	44
Cluster 3	32.880	100.333	94.888	13.400	117.200	94.200	80.771	24.260	42
Cluster 4	50.233	142.000	134.143	8.667	62.000	53.333	85.983	24.367	53
Cluster 5	48.483	91.000	130.700	11.000	194.167	170.667	87.878	24.405	49
Cluster 6	30.111	97.222	103.124	11.222	159.722	133.389	83.635	21.253	37
Mean	35.511	95.974	104.234	11.038	112.641	94.551	83.991	22.524	42
TreatMSS	293.651	659.454	1407.560	11.397	11833.063	8850.502	47.829	25.545	121
ErrMSS	28.321	123.975	202.129	8.108	146.093	96.707	46.491	10.637	104

F Ratio	10.369	5.319	6.964	1.406	80.997	91.519	1.029	2.400	1.158	1.340	1.619
Probability	0.000	0.004	0.001	0.266	0.000	0.000	0.416	0.101	0.338	0.282	0.209

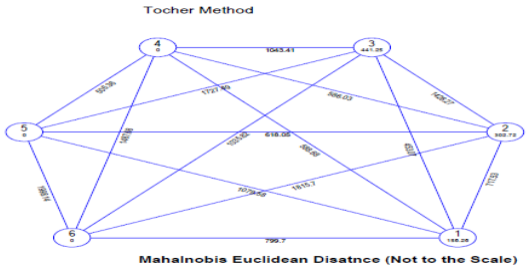

Fig. 5. Ward's minimum variance dendrogram for cluster analysis of 26 rice genotypes.

The proportion of contribution of individual traits to overall genetic divergence was estimated based on their frequency in maximum inter-genotypic distance estimations. Spikelets per panicle contributed the greatest proportion, followed by days to flowering and plant height. Results of trait ranking obtained using the algorithm were consistent with PCA loadings (Table 3) and path coefficient findings (Table 2), confirming the robustness of the combined computational model.

Overall Interpretation

The agreement of path coefficient analysis (Table 2), PCA (Table 3; Fig. 2 and Fig. 3), and cluster analysis (Tables 4–6; Fig. 4 and Fig. 5) proves that a complicated network of morphological and physiological characteristics controls grain yield in rice. The inclusion of multivariate statistical components into an orchestrated computational algorithm contributed to better interpretational understanding, decreased subjective prejudice, and created reproducible classification of genotypes.

IV. DISCUSSION

The current research has shown high genetic diversity among the tested rice varieties based on multivariate divergence indices. The clustering pattern from Mahalanobis D^2 analysis confirmed large genetic differences existing between genotype groups. The high level of deviation between Cluster I and Cluster V gives a high possibility of strategic hybridization.

Path coefficient analysis showed the strongest positive direct effects on grain yield from biological yield and harvest index, consistent with developed yield component models identifying biomass accretion and assimilate partitioning efficiency as productivity determinants. Spikelets per panicle and grains per panicle also showed significant direct and indirect effects.

The multidimensional pattern of variability was supported by PCA. The first principal component was dominated by structural yield components (spikelets and grains per panicle), while the second principal component represented physiological efficiency (biological yield and harvest index). The cumulative variance of approximately 81% for the first four components confirms that most genetic data can be summarized in a reduced trait space.

The overlap of cluster analysis, PCA and path modelling enhances the strength of the findings. When several statistical modules point to similar traits as significant contributors to divergence, selection decisions become more reliable. Combining these methods into one computational pipeline minimizes interpretational subjectivity and guarantees reproducibility — factors increasingly essential in modern agricultural research.

The evolution of a Genotype Divergence Index (GDI) in the current research represents a step towards composite ranking systems that combine several statistical outputs into a single selection score, filling the gap between theoretical statistical analysis and practical breeding decisions.

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

The current research illustrated the use of a combined computational model in assessing genetic divergence between rice genotypes. Hierarchical clustering, Mahalanobis D^2 statistics, PCA and path coefficient modelling were combined to give thorough insight into trait interrelationships and genotype differentiation. Biological yield, harvest index and spikelets per panicle emerged as the most influential factors.

The presence of extended distant clusters and distinct genotypes confirms sufficient variability for effective hybridization programs. The statistical results validated the effectiveness of the proposed Agricultural Genetic Divergence Evaluation Algorithm (AGDEA), accelerating reproducibility of classical biometric methods by organizing them into an ordered computational methodology.

B. Future Scope: Artificial Intelligence Integration

The suggested assessment algorithm creates a basic framework that can be enhanced with AI and machine learning. Future improvements include:

- Trait weight optimization using supervised machine learning models trained on yield performance data across environments.
- Genomic data integration to extend phenotypic divergence analysis to genomic divergence using molecular markers.
- Deep learning of phenotypic images via computer vision to automatically extract morphological features in real time.

- Multi-environment stability analysis using AI-derived stability indices to incorporate genotype-environment interactions.
- Cloud-based decision support systems enabling breeders to upload datasets and receive automated divergence reports.
- Explainable Artificial Intelligence (XAI) integration to guarantee transparency in algorithm-based breeding decisions.

REFERENCES

[1] G. Khush, "Origin, dispersal, cultivation and variation of rice," Plant Molecular Biology, vol. 35, pp. 25–34, 1997.

[2] S. Vaughan, H. Morishima and K. Kadowaki, "Diversity in the Oryza genus," Current Opinion in Plant Biology, vol. 11, pp. 144–148, 2008.

[3] D. Fuller et al., "The domestication process and domestication rate in rice," Science, vol. 323, pp. 1607–1610, 2009.

[4] IRRI, Rice Almanac, 4th ed., International Rice Research Institute, 2013.

[5] P. C. Mahalanobis, "On the generalized distance in statistics," Proc. National Institute of Sciences of India, 1936.

[6] C. R. Rao, Advanced Statistical Methods in Biometric Research, Wiley, 1952.

[7] S. K. Singh et al., "Genetic divergence studies in rice," Field Crops Research, vol. 135, pp. 1–8, 2012.

[8] R. Sarawgi and R. Binse, "Genetic diversity in rice genotypes," Oryza, vol. 44, pp. 213–217, 2007.

[9] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.

[10] A. Kumar et al., "Multivariate analysis in rice germplasm evaluation," Euphytica, vol. 186, pp. 37–46, 2012.

[11] J. Ward, "Hierarchical grouping to optimize objective function," Journal of the American Statistical Association, 1963.

[12] D. R. Dewey and K. H. Lu, "A correlation and path coefficient analysis of components of crested wheatgrass seed production," Agronomy Journal, 1959.

[13] S. B. Singh and B. Chaudhary, Biometrical Methods in Quantitative Genetic Analysis, Kalyani Publishers, 2007.

[14] J. Araus and J. Cairns, "Field high-throughput phenotyping," Trends in Plant Science, vol. 19, pp. 52–61, 2014.

[15] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

[16] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[17] A. Kamilaris and F. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[18] S. Liakos et al., "Machine learning in agriculture: A review," Sensors, vol. 18, 2018.

[19] K. Shahhosseini et al., "Machine learning for crop yield prediction," Scientific Reports, vol. 10, 2020.

[20] B. Samek et al., "Explainable artificial intelligence," IEEE Signal Processing Magazine, 2017.

[21] M. Wolfert et al., "Big data in smart farming," Agricultural Systems, vol. 153, pp. 69–80, 2017.

[22] J. Montesinos-López et al., "Multivariate genomic selection models," G3: Genes, Genomes, Genetics, 2016.

[23] P. Crossa et al., "Genomic selection in plant breeding," Trends in Plant Science, 2017.

[24] G. S. Khush, "Green revolution: Preparing for the 21st century," Genome, 1999.

[25] FAO, "Rice market monitor," Food and Agriculture Organization, 2023.