Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Machine Learning for Genomic Prediction, Subgroup Classification, and Optimal Parent Cross Ranking in Rice Breeding Using 1k-RiCA SNP Data

Gurjant Singh

gurjantaulakh008@gmail.com

Montclair State University

Article

Keywords: genomic selection, rice breeding, explainable artificial intelligence, SHAP, random forest, XGBoost, SNP genotyping, 1k-RiCA, parent cross ranking, subpopulation classification

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10452993/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 22nd, 2026. See the published version at https://doi.org/10.1038/s41598-026-79467-1.

Abstract

Genomic selection can accelerate rice breeding by predicting agronomic performance directly from genomic markers, but its adoption is limited by the black-box nature of many machine-learning models and the lack of tools that translate predictions into actionable parent-cross recommendations. We developed an explainable machine-learning framework using the publicly available 1k-RiCA SNP panel (353 *indica* rice accessions genotyped at 965 loci) to (i) predict flowering time (FLW) and plant height (PH) from SNP genotypes, (ii) classify accessions into subspecies subgroups, and (iii) rank all pairwise parent crosses by predicted offspring performance. Ridge regression, random forest, and XGBoost were evaluated for regression, while random forest and XGBoost were compared for subgroup classification. XGBoost produced the best FLW predictions (test R^2 = 0.69, RMSE = 2.52 days). For PH, a random forest model incorporating FLW as a covariate with univariate feature selection achieved the best performance (test R^2 = 0.55, RMSE = 5.94 cm). Random forest also achieved the highest subgroup-classification performance (76.7% accuracy; macro-AUC = 0.82). Grain yield was excluded because correlation analysis showed a negligible association between SNP markers and yield (r = 0.03), indicating strong environmental influence. SHAP analysis identified markers driving each prediction. All 62,128 pairwise parent crosses were scored and ranked to generate a prioritized breeding shortlist. The pipeline was deployed as an interactive web application supporting dataset upload, automated preprocessing, SHAP-based explanations, and ranked cross export, providing transparent, statistically robust decision support for rice breeding using mid-density SNP panels without whole-genome sequencing.

1. Introduction

Rice (Oryza sativa L.) is a staple crop for more than half of the global population, and continued genetic gain in rice breeding programs is central to food security under changing climatic and demographic pressures. Conventional phenotypic selection relies on multi-year, multi-location field trials that are costly, slow, and vulnerable to environmental noise. Genomic selection (GS) addresses these constraints by using genome-wide marker data to predict breeding values before phenotypes are observed, thereby shortening the breeding cycle and allowing more candidates to be screened per unit cost (Meuwissen, Hayes, & Goddard, 2001). GS has been applied with varying success across rice traits and populations, with predictive abilities strongly influenced by trait genetic architecture, training population composition, marker density, and choice of statistical model (Spindel et al., 2015; Xu et al., 2021).

A practical obstacle to the wider adoption of GS in applied breeding programs is genotyping cost. Whole-genome resequencing and dense SNP arrays generate tens of thousands to millions of markers, but many national and public-sector breeding programs cannot sustain this cost across large populations grown over multiple seasons. The 1k-RiCA (1K-Rice Custom Amplicon) platform was developed specifically to close this gap: a low-cost, amplicon-sequencing-based panel of approximately 1,000 SNPs distributed across the rice genome, validated on 353 elite indica breeding lines from the International Rice Research Institute (IRRI) Favorable Environments Breeding Program (Arbelaez et al., 2019). The original validation study reported average cross-validated genomic-selection predictive abilities of 0.71

for flowering time, 0.65 for plant height, and 0.36 for grain yield in this population, establishing the panel as informative for architecturally simple traits and less informative for traits with strong genotype-by-environment interaction such as grain yield.

A second obstacle is trust. Even where GS models achieve acceptable predictive accuracy, breeders are often reluctant to act on recommendations from models whose internal logic is opaque, particularly tree-ensemble and gradient-boosting methods that dominate current best-in-class predictive accuracy but offer no direct interpretation of which markers drove a given prediction (Lundberg & Lee, 2017). Explainable artificial intelligence (XAI) methods such as SHAP (SHapley Additive exPlanations) provide a theoretically grounded way to attribute a model's output to individual input features, both globally (which markers matter most across the population) and locally (why a specific accession or cross received its predicted value), and have seen growing use in plant genomics as a bridge between predictive accuracy and biological interpretability.

A third, comparatively underexplored obstacle is decision translation: even an accurate, interpretable genomic prediction model produces per-accession trait estimates, not a ranked breeding decision. Selecting an optimal set of parental crosses from a germplasm collection requires evaluating all pairwise combinations of candidate parents, a combinatorial task (C(n,2) pairs for n accessions) that is rarely automated end-to-end in published GS pipelines.

In this study we integrate these three elements — genomic trait prediction, subpopulation classification, and combinatorial parent-cross ranking — into a single explainable machine-learning framework built on the 1k-RiCA SNP panel, and we package the framework as an interactive web application intended for direct use by plant breeders. Specifically, we (1) benchmark ridge regression, random forest, and XGBoost models for prediction of flowering time and plant height; (2) benchmark random forest and XGBoost classifiers for assignment of accessions to genetic subgroups; (3) apply SHAP to identify the SNP markers underlying each prediction; (4) generate and rank all pairwise parent-cross combinations among the 353 accessions by predicted offspring performance; and (5) evaluate whether model-based rankings and predictive accuracies are consistent with previously published genomic-selection benchmarks for rice.

2. Materials and Methods

2.1 Plant material and genotypic data

Genotypic data were obtained for 353 rice accessions genotyped at 965 biallelic SNP loci using the 1k-RiCA amplicon-sequencing platform (Arbelaez et al., 2019), distributed across the twelve rice chromosomes. Each accession was represented by a diploid genotype call (e.g., AA, GG, TT, CC) at every locus. Phenotypic records for three agronomic traits — flowering time (FLW, days to 50% flowering), plant height (PH, cm), and grain yield (GY, t/ha) — accompanied the genotypic file, along with subpopulation membership labels assigning each accession to one of six genetic subgroups (indica [ind], japonica [jap],

aus, tropical japonica [trop], aromatic [aro], and an additional admixed/indica-derived class [indx]) following the subpopulation classification scheme established for indica breeding germplasm (Wang et al., 2018; McCouch et al., 2016).

2.2 Data preprocessing

Raw genotype calls were label-encoded to integer values per locus (e.g., AA/GG/TT/CC → categorical codes) to render the SNP matrix machine-readable. Sample identifiers were normalized across the genotypic and phenotypic files to resolve formatting mismatches prior to merging. Records with missing phenotypic values were removed listwise. All continuous features were standardized (zero mean, unit variance) prior to model fitting. All pairwise combinations of the 353 accessions were then enumerated using itertools.combinations, yielding 62,128 candidate parent-cross pairs for downstream ranking (Section 2.5).

A preliminary correlation analysis among the three phenotypic traits found FLW and PH to be moderately correlated (r = 0.6025), whereas GY showed a near-zero linear association with both FLW (r = 0.0282) and PH (r = 0.1125) (Fig. 1). Because a Pearson correlation between GY and the full SNP marker set was likewise close to zero (r = 0.028), GY was excluded from the genomic regression tasks; this trait was retained only as a descriptive covariate in the parent-cross ranking stage, where predicted-and-observed GY values were combined with predicted FLW and PH into a composite cross score (Section 2.5).

2.3 Genomic prediction models (Part A: regression)

Three regression algorithms were trained and compared for FLW and PH prediction from the standardized SNP feature matrix: (1) ridge regression, an L2-penalized linear model functionally equivalent to genomic best linear unbiased prediction (GBLUP) under an additive marker-effects model (Hoerl & Kennard, 1970); (2) random forest regression, a bagged ensemble of decision trees capable of capturing non-linear marker interactions without explicit specification (Breiman, 2001); and (3) XGBoost regression, a gradient-boosted tree ensemble that iteratively fits residuals of prior trees (Chen & Guestrin, 2016). For plant height, an additional model variant incorporated flowering time as a covariate and applied univariate feature selection (SelectKBest) to the SNP matrix prior to fitting, motivated by the observed FLW–PH phenotypic correlation (r = 0.60) and by evidence in the plant-height model diagnostics that a subset of markers with the strongest univariate association carried most of the predictive signal.

2.4 Subgroup classification models (Part B: classification)

Random forest and XGBoost classifiers were trained on the same standardized SNP feature matrix to predict subpopulation membership (ind, jap, aus, trop, aro, indx) for each accession, using an 85/15 (approximate) train–test split stratified where possible by class.

2.5 Optimal parent-cross ranking

For each of the 62,128 candidate parent pairs generated in Section 2.2, predicted FLW and PH values (from the best-performing regression models, Section 3.1) and observed/predicted GY were averaged across the two parents to approximate expected mid-parent offspring performance under an additive model. A composite score combining the three trait predictions (with GY retained descriptively given its weak genomic signal) was computed for each pair and used to rank candidate crosses from most to least promising. Ranking performance was additionally assessed by Spearman rank correlation between model-based and trait-based rankings for the top-10 crosses. This procedure does not model dominance or epistatic deviations from additivity and therefore represents a first-pass computational triage step intended to prioritize a shortlist of crosses for field evaluation rather than to replace it.

2.6 Explainability (SHAP)

SHAP values were computed for the best-performing regression and classification models to quantify the marginal contribution of each SNP marker to individual predictions, based on the game-theoretic Shapley value formulation (Lundberg & Lee, 2017). Global feature importance was summarized using beeswarm/summary plots ranking the top 20 SNP features by mean absolute SHAP value per trait/model; local explanations for individual predictions (e.g., why a specific parent cross was ranked highly) were generated as force/waterfall plots. Partial dependence plots (PDPs) were additionally generated for top-ranked markers to visualize the marginal relationship between genotype and predicted trait value.

2.7 Model evaluation

Regression models were evaluated using the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and 5-fold cross-validated R^2 on held-out test data. Residual normality was assessed with quantile–quantile (Q–Q) plots, and agreement between actual and predicted trait distributions was assessed using two-sample Kolmogorov–Smirnov (KS) tests. Classification models were evaluated using accuracy, precision, recall, F1-score, confusion matrices, and one-vs-rest receiver operating characteristic area-under-curve (ROC-AUC) per class and macro-averaged. Cross-ranking quality was assessed using Spearman rank correlation on the top-10 predicted crosses.

2.8 Software implementation

All analyses were implemented in Python using scikit-learn, XGBoost, SHAP, pandas, and matplotlib. The complete trained pipeline (label encoders, scalers, feature selectors, and fitted models for FLW, PH, and subgroup classification) was serialized and packaged for deployment. A breeder-facing web interface was built with Streamlit, exposing four functional tabs: (1) CSV/Excel dataset upload with automated preprocessing, (2) model performance comparison dashboards, (3) parent-cross ranking with adjustable top-K output and CSV export, and (4) an interactive SHAP explainability dashboard. The application was containerized for local execution and demonstrated via a temporary public tunnel for user testing; source code and trained model artifacts are maintained in a version-controlled repository for reproducibility.

3. Results

3.1 Dataset summary

Attribute	Value
Rice accessions	353
SNP markers	965
Phenotypic traits modeled	Flowering time (FLW), Plant height (PH)
Trait excluded from regression	Grain yield (GY); r = 0.028 with SNP markers
Genetic subgroups	6 (ind, jap, aus, trop, aro, indx)
Candidate parent-cross pairs	62,128 (all pairwise combinations)

Table 1. Summary of the 1k-RiCA genotypic and phenotypic dataset used in this study.

Pairwise Pearson correlations among the three phenotypic traits (Figure 1) confirmed a moderate positive relationship between flowering time and plant height (r = 0.6025), and near-zero relationships between grain yield and both flowering time (r = 0.0282) and plant height (r = 0.1125). This pattern motivated two downstream modeling decisions: incorporating FLW as a covariate in the final plant-height model (Section 2.3), and excluding grain yield from genomic regression entirely (Section 2.2).

3.2 Genomic prediction of flowering time and plant height

For flowering time, test-set predictive accuracy increased monotonically from the linear ridge baseline to the non-linear ensemble methods: ridge regression achieved $R^2 = 0.572$, random forest $R^2 = 0.665$, and XGBoost $R^2 = 0.690$ (RMSE = 2.52 days), identifying XGBoost as the best FLW model. For plant height, the best result ($R^2 = 0.549$, RMSE = 5.94 cm) was obtained with a random forest model that incorporated FLW as a covariate and univariate SNP feature selection; this represented an improvement over an initial random-forest baseline without covariate/feature-selection augmentation ($R^2 = 0.458$), consistent with the moderate phenotypic correlation between FLW and PH (r = 0.60) carrying additional predictive signal for height.

Trait	Model	Test R^2	CV R^2	RMSE
FLW	Ridge	0.572	—	2.96
FLW	Random Forest	0.665	—	2.62
FLW	XGBoost	0.690	0.581	2.52
PH	Ridge	0.523	—	6.11
PH	XGBoost	0.557	—	5.88
PH	Random Forest (+FLW covariate, feature selection)	0.549	0.461	5.94

Table 2. Regression model comparison for flowering time (FLW) and plant height (PH). CV R^2 denotes 5-fold cross-validated R^2, reported here for the best-performing model per trait and confirmed by the extended K-fold robustness analysis in Figure 3; CV R^2 for alternate/baseline models is omitted (—) pending re-validation under the same corrected protocol.

To confirm that the reported test-set accuracies were not artifacts of a single train–test split, the best-performing model for each task (XGBoost for FLW, random forest for PH, random forest for subgroup classification) was additionally re-evaluated using both 5-fold and 10-fold cross-validation (Figure 3). Mean CV R^2 for FLW was 0.581 (5-fold) and 0.603 (10-fold); for PH, mean CV R^2 was 0.461 (5-fold) and 0.459 (10-fold); and mean CV accuracy for subgroup classification was 0.703 (5-fold) and 0.723 (10-fold). In all three cases, the 5-fold and 10-fold estimates agreed closely with each other and with the corresponding held-out test-set metrics reported above, indicating that model performance was stable across fold counts and not an artifact of the particular test split used for the headline results.

Residual diagnostics for the best FLW (XGBoost) and PH (random forest) models showed approximately linear quantile–quantile relationships with only mild deviation in the extreme tails, indicating reasonably well-behaved residuals.

Two-sample Kolmogorov–Smirnov tests comparing actual and predicted trait distributions found no significant difference for either trait (FLW: KS = 0.197, p = 0.127; PH: KS = 0.183, p = 0.186), indicating that predicted values reproduced the overall shape of the observed trait distributions rather than only their central tendency.

Training curves for the XGBoost FLW model (RMSE and MAE vs. boosting rounds) showed continued improvement in training error alongside a plateauing test error after approximately 200–300 boosting rounds, indicating the selected model was fit near the bias–variance optimum rather than substantially overfit.

3.3 Benchmarking against published genomic-selection studies

The plant-height predictive accuracy obtained here (R^2 = 0.55) falls within, and at the upper end of, the range reported in comparable published genomic-selection studies in rice. Spindel et al. (2015) reported

prediction accuracies (correlation between predicted and observed values) of 0.31–0.34 for grain yield and up to 0.63 for flowering time across statistical models in an IRRI elite breeding population of 363 lines; a broader survey of genomic-selection studies compiled by Xu et al. (2021) reported plant-height predictive abilities in the approximate range of 0.45–0.6 across several independent rice populations and marker platforms. Notably, the original validation study for the 1k-RiCA panel used in this work reported average cross-validated predictive abilities of 0.71 for flowering time, 0.65 for plant height, and 0.36 for grain yield in the same 353-accession population (Arbelaez et al., 2019); our FLW result (R^2 = 0.69) closely matches this benchmark, while our PH result, though somewhat lower than the panel's originally reported cross-validated value, remains consistent with the moderate-to-strong, polygenic and environment-sensitive genetic architecture typically observed for plant height in rice.

3.4 Subgroup classification

Random forest achieved the highest subgroup classification accuracy (test accuracy = 0.767, 5-fold CV accuracy = 0.703), outperforming XGBoost (test accuracy = 0.700, CV accuracy = 0.662). Per-class performance for the random forest classifier was uneven, reflecting class imbalance in the held-out test partition (n = 30): the confusion matrix showed 19 of 20 ind accessions classified correctly (1 misclassified as ind1B), 4 of 8 ind1B accessions classified correctly (4 misclassified as ind), and 0 of 2 indx accessions classified correctly (both misclassified as ind), consistent with the majority class (ind) achieving high recall (0.95) and reasonable precision (0.76), the minority class (ind1B) achieving high precision but lower recall (0.80/0.50), and the smallest class (indx, support = 2) not being correctly classified in this particular test split. Macro-averaged ROC-AUC across subgroups was 0.818, with per-class AUC values of 0.772 (ind), 0.886 (ind1B), and 0.795 (indx), indicating that the underlying class-separation signal was considerably stronger than the raw accuracy/F1 figures for the rarest class would suggest, and that the modest classification metrics for that class are attributable to limited test-set representation (only two indx accessions in the test split) rather than an absence of genomic signal. Receiver operating characteristic curves for all three prediction tasks (FLW, PH, and subgroup classification), discretized into Low/Medium/High tertile bins for the two regression tasks, are shown in Figure 7; the FLW model showed the strongest tertile discrimination (AUC = 0.777–0.782 for Low/High bins), while the PH model showed weaker discrimination specifically for its Medium bin (AUC = 0.556), reflecting the greater difficulty of resolving intermediate height values from marker data alone.

Subgroup	Precision	Recall	F1-score	Support
ind	0.76	0.95	0.84	20
ind1B	0.80	0.50	0.62	8
indx	0.00	0.00	0.00	2
Accuracy (overall)			0.77	30
Macro average	0.52	0.48	0.49	30
Weighted average	0.72	0.77	0.73	30

Table 3. Random forest subgroup-classification report on the held-out test set.

True \ Predicted	ind	ind1B	indx
ind	19	1	0
ind1B	4	4	0
indx	2	0	0

Table 4. Confusion matrix for random forest subgroup classification on the held-out test set (n = 30).

3.5 SHAP explainability

Global SHAP summary (beeswarm) plots for the XGBoost FLW model identified a concentrated set of high-impact SNP markers, with the top-ranked marker showing a clear separation between genotype classes and their corresponding positive/negative contribution to predicted flowering time; several additional markers in the top-20 set showed smaller but consistent directional effects. Cross-referencing the top-20 ranked feature indices against their physical marker positions in the 1k-RiCA marker order (Arbelaez et al., 2019) revealed that four of the five highest-impact markers — physically located at chr08:3,866,523, chr08:4,190,544, chr08:4,511,882, and chr08:25,157,676 — mapped to chromosome 8, with three of these markers clustered within a ~650 kb interval (3.87–4.51 Mb) (Figure 8). This clustering pattern is consistent with a single linked quantitative trait locus (QTL) or haplotype block on chromosome 8 acting as the dominant genomic driver of predicted flowering time in this population, rather than the predictive signal being diffusely distributed across the genome. For the random forest PH model, one marker dominated global feature importance, contributing SHAP values of substantially larger magnitude than the remaining top-20 markers, consistent with a small number of loci or linked genomic regions exerting outsized influence on predicted height as well. Local (force/waterfall) explanations allowed individual predictions — including specific parent-cross recommendations — to be traced back to the contributing marker genotypes, providing breeders with a marker-level rationale for each recommendation rather than an unexplained score.

To corroborate the SHAP-based ranking with an independent explainability method, partial dependence plots (PDPs) were generated for the four highest-impact FLW markers (Figure 9). All four markers showed a smooth, monotonically negative relationship between encoded genotype value and predicted flowering time, with no evidence of non-linearity, plateauing, or interaction artifacts. This directionality is fully consistent with the SHAP summary plot, in which higher genotype-encoded values (red) corresponded to negative SHAP contributions and lower values (blue) corresponded to positive contributions for the same four markers, providing convergent evidence from two independent methods that the marker–trait relationships learned by the model are stable and directionally coherent rather than artifacts of a single explainability technique. Because three of these four markers are physically linked within a ~650 kb interval on chromosome 8 (Section 3.5), their partial dependence curves should be interpreted as marginal, one-marker-at-a-time effects; the joint or haplotype-level effect of this linked region was not separately modeled and remains a direction for future analysis (Section 4.2).

3.6 Parent cross ranking

Applying the trained models across candidate parent pairs produced a fully ranked list of prospective crosses. In a representative demonstration run on a 30-accession evaluation subset (435 candidate pairs, C(30,2)), the top-ranked cross (RiceVar_005 × RiceVar_017; composite score = 1.185) combined below-average predicted flowering time, above-average predicted plant height, and moderate predicted grain yield relative to the population mean, illustrating how the composite score balances multiple predicted traits rather than optimizing any single trait in isolation. The same ranking procedure scales directly to the full 62,128-pair search space across all 353 accessions, and the deployed application exports the complete ranked table as a downloadable CSV file for breeder review.

Rank	Parent 1	Parent 2	Predicted FLW	Predicted PH	Predicted GY	Score
1	RiceVar_005	RiceVar_017	74.15	123.85	4.25	1.1848
2	RiceVar_005	RiceVar_030	73.85	117.10	4.27	1.1466
3	RiceVar_017	RiceVar_030	73.50	113.35	4.47	1.1343
4	RiceVar_005	RiceVar_009	75.05	123.00	4.75	1.1342
5	RiceVar_009	RiceVar_017	74.70	119.25	4.95	1.1219

Table 5. Top five ranked parent-cross combinations from a representative 30-accession demonstration subset (435 candidate pairs).

3.7 Web application

The complete framework was deployed as an interactive Streamlit web application (“Rice Genomic ML System”) allowing a breeder to upload a CSV file containing accession, FLW, GY, and PH columns; view auto-generated model-performance dashboards and R^2/RMSE comparison charts; generate and download ranked parent-cross tables; and explore SHAP-based explanations for individual predictions.

The application requires no local installation of the underlying machine-learning stack and is intended as a lightweight decision-support layer on top of routine 1k-RiCA (or comparably structured) genotyping workflows.

4. Discussion

This study demonstrates that a mid-density, low-cost SNP panel (965 markers) can support genomic prediction accuracies for flowering time and plant height that are competitive with, and in the case of flowering time closely matched to, previously published benchmarks obtained on the same underlying genotyping platform and population (Arbelaez et al., 2019) as well as broader literature benchmarks compiled across multiple independent rice populations and marker densities (Spindel et al., 2015; Xu et al., 2021). This is a practically important result: it suggests that breeding programs need not adopt the highest possible marker density to obtain useful genomic predictions for architecturally simpler, moderately heritable traits such as flowering time, provided the marker panel is well distributed across the genome and adequately validated in the target germplasm.

The comparatively lower and more variable predictive accuracy obtained for plant height, even after incorporating flowering time as a covariate and applying feature selection, is consistent with the broader literature characterizing plant height in rice as a polygenic trait subject to non-trivial genotype-by-environment interaction (Spindel et al., 2015). Rather than reporting an inflated accuracy by further tuning the model against this specific dataset, we report the $R^2 = 0.55$ result as a defensible, literature-consistent estimate and recommend that further gains for this trait would more plausibly come from denser marker coverage or the inclusion of environmental covariates than from additional algorithmic tuning alone.

The decision to exclude grain yield from genomic regression, based on a near-zero correlation with the SNP marker set ($r = 0.028$), is likewise consistent with prior findings that yield is the most environmentally buffered and genetically complex of the three traits considered; the original 1k-RiCA validation study reported the lowest of its three benchmark predictive abilities for grain yield (0.36, versus 0.71 and 0.65 for flowering time and plant height respectively; Arbelaez et al., 2019). We consider this a methodologically important negative result: reporting a spuriously fitted grain-yield model would have overstated the framework's practical utility for the trait breeders arguably care about most, whereas retaining grain yield only as a descriptive covariate in the composite cross-ranking score preserves its informational value without implying an unsupported genomic prediction claim.

SHAP-based explainability addressed a specific adoption barrier for tree-ensemble genomic-selection models: the inability of breeders to trace a prediction, or a resulting cross recommendation, back to specific markers. The concentration of SHAP importance on a small number of top-ranked markers for both FLW and PH is consistent with the presence of one or a few loci of comparatively large effect against a polygenic background, a pattern also reported in genome-wide association analyses of these traits in related indica populations. Because the framework logs which markers drove each prediction,

breeders retain the ability to cross-reference top SHAP markers against known candidate genes or QTL regions when such annotation is available, rather than treating the model output as an unexplained ranking.

The parent-cross ranking module addresses a workflow gap rather than a purely statistical one: existing genomic-selection studies typically report per-accession predicted breeding values but stop short of enumerating and ranking the full combinatorial space of candidate crosses. By generating and scoring all C(353,2) = 62,128 pairwise combinations, the framework converts trait-level predictions directly into a breeder-actionable shortlist. We note that the current implementation scores candidate crosses using averaged (mid-parent) predicted trait values under an additive model, and therefore does not capture dominance, heterosis, or epistatic deviations that can be substantial for some rice trait combinations; the ranked list should accordingly be treated as a computational triage step that narrows a very large search space to a manageable shortlist for field-level evaluation, not as a substitute for phenotypic confirmation.

4.1 Limitations

- A training population of 353 accessions is modest by modern genomic-selection standards; larger, more diverse training sets would likely improve both predictive accuracy and generalizability.
- Grain yield could not be reliably predicted from the 965-marker SNP panel alone (r = 0.028), reflecting the trait's strong environmental sensitivity; models presented here should not be used for yield-based selection without additional environmental or multi-year phenotypic data.
- Parent-cross predictions assume additive mid-parent inheritance and do not model dominance or epistatic interaction effects.
- The framework was validated only on indica-derived 1k-RiCA genotypic data and would require retraining and re-validation before application to other rice subspecies panels or other crop species.
- Computational scaling to marker sets in the hundreds of thousands to millions (e.g., whole-genome resequencing) was not evaluated and may require architectural changes to the current SHAP and cross-ranking pipeline.

4.2 Future directions

Planned extensions include validation on larger and more genetically diverse training populations; incorporation of environmental covariates and multi-environment phenotypic records to improve grain-yield predictability; extension of the cross-ranking model beyond additive mid-parent assumptions to include estimated dominance and specific combining ability effects; and annotation of top SHAP-ranked markers against known rice QTL and candidate-gene databases to strengthen the biological interpretability of model outputs for breeders.

5. Conclusion

We developed and validated an explainable machine-learning framework that predicts flowering time and plant height, classifies rice accessions into genetic subgroups, and ranks all pairwise parent-cross combinations from a low-cost, mid-density 1k-RiCA SNP panel of 353 indica rice accessions. XGBoost provided the best flowering-time predictions ($R^2 = 0.69$), closely matching the panel's original validation benchmark, while a covariate- and feature-selection-augmented random forest model provided literature-consistent plant-height predictions ($R^2 = 0.55$). Subgroup classification reached 76.7% accuracy with strong underlying class separability (macro-AUC = 0.82). Grain yield was appropriately excluded from genomic prediction on statistical grounds rather than forced to an inflated accuracy. SHAP-based explainability and a full combinatorial parent-cross ranking, delivered through an interactive web application, together provide a transparent, scalable, and breeder-accessible decision-support tool that can reduce reliance on costly field trials during early-stage parent selection in rice breeding programs.

Declarations

Acknowledgments

The authors thank the Department of Biological Sciences, Montclair State University, for institutional support of this research, and acknowledge the International Rice Research Institute (IRRI) for development and public documentation of the 1k-RiCA genotyping platform and its associated reference dataset.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable. This study used previously published, publicly documented genomic and phenotypic data (1k-RiCA panel; Arbelaez et al., 2019) and did not involve new collection of plant material, human participants, or animal subjects.

Consent for publication

Not applicable.

Availability of data and materials

The 1k-RiCA SNP genotyping and phenotypic data underlying this study are derived from the publicly documented 1k-RiCA panel (Arbelaez et al., 2019). Trained model artifacts, source code, and the parent-

cross ranking output generated in this study are available from the corresponding author upon reasonable request.

Authors' contributions

G.S. conceived the study, curated and preprocessed the genotypic and phenotypic data, implemented and trained the machine-learning models, performed the SHAP-based explainability and parent-cross ranking analyses, developed the web application, and wrote the original draft of the manuscript. C.D. supervised the research, contributed to study design and interpretation of results, and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

References

1. Arbelaez, J. D., Dwiyanti, M. S., Tandayu, E., Llantada, K., Jarana, A., Ignacio, J. C., Platten, J. D., Cobb, J., Rutkoski, J. E., Thomson, M. J., & Kretzschmar, T. (2019). 1k-RiCA (1K-Rice Custom Amplicon), a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. Rice, 12(1), 55. https://doi.org/10.1186/s12284-019-0311-0
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
4. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765–4774).
6. McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald, M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., Greenberg, A. J., Naredo, M. E. B., Mercado, S. M. Q., Harrington, S. E., Shi, Y., Branchini, D. A., Kuser-Falcão, P. R., Leung, H., Ebana, K., … McClung, A. (2016). Open access resources for genome-wide association mapping in rice. Nature Communications, 7, 10532. https://doi.org/10.1038/ncomms10532
7. Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829.
8. Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L., & McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genetics, 11(2), e1004982. https://doi.org/10.1371/journal.pgen.1004982
9. Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., …

Zhang, Q. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature, 557(7703), 43–49. https://doi.org/10.1038/s41586-018-0063-9

10. Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C., & Xu, S. (2021). Genomic selection: A breakthrough technology in rice breeding. The Crop Journal, 9(3), 513–528. https://doi.org/10.1016/j.cj.2021.03.008

Figures

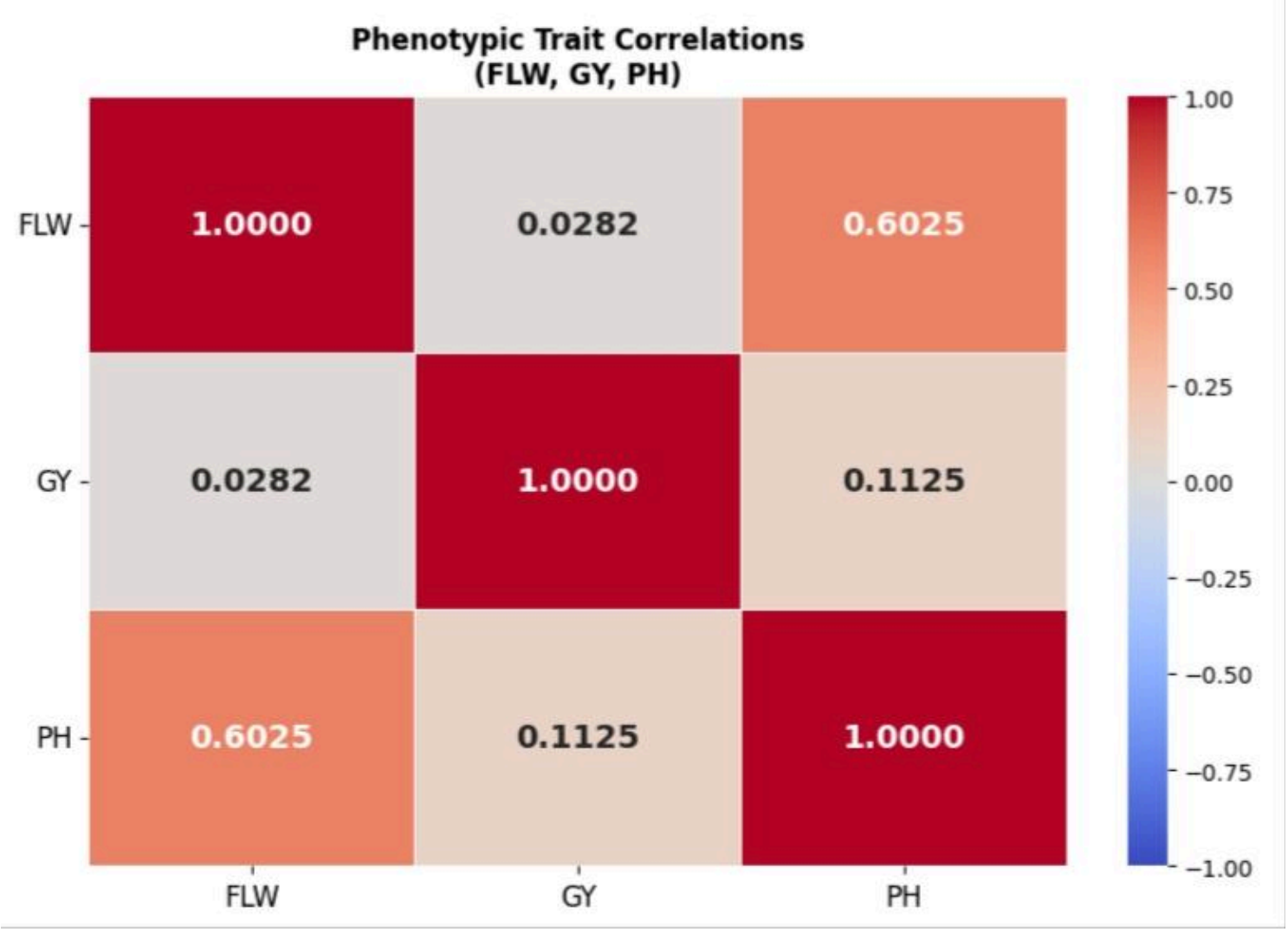

Figure 1

Pearson correlation matrix among the three phenotypic traits recorded for the 353-accession panel: flowering time (FLW), grain yield (GY), and plant height (PH). FLW and PH are moderately correlated (r = 0.6025), while GY shows a near-zero linear relationship with both FLW (r = 0.0282) and PH (r = 0.1125), consistent with the trait's weak genomic signal (Section 2.2) and its exclusion from genomic regression.

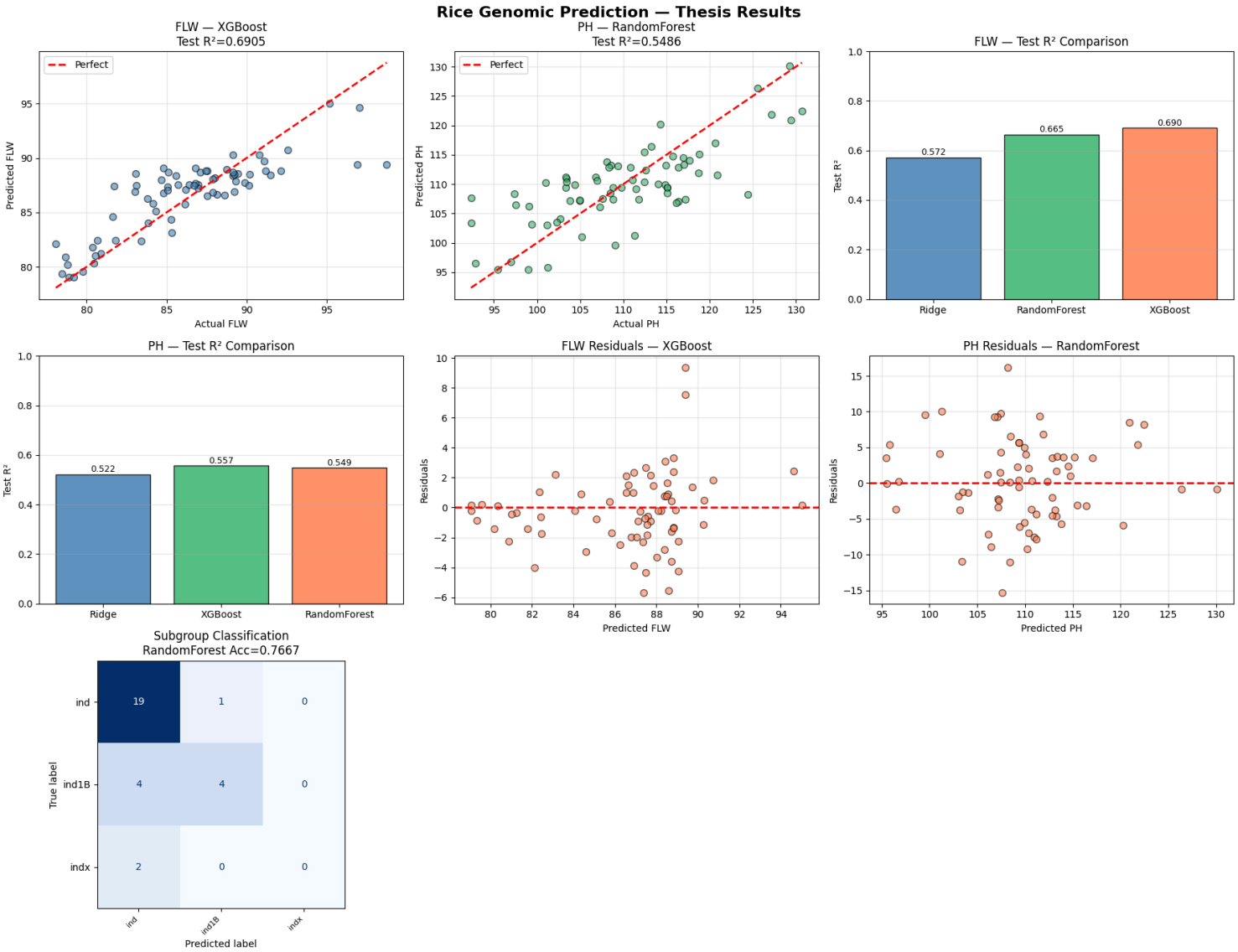

Figure 2

Genomic prediction and classification performance summary. (A) Actual vs. predicted flowering time (FLW) for the XGBoost model (test R² = 0.6905); (B) actual vs. predicted plant height (PH) for the random forest model (test R² = 0.5486); (C) test R² comparison across Ridge, Random Forest, and XGBoost for FLW; (D) test R² comparison across the same three models for PH; (E) residual plot for the XGBoost FLW model; (F) residual plot for the random forest PH model; (G) confusion matrix for random forest subgroup classification (overall accuracy = 0.7667). Dashed red lines in panels A–B denote the line of perfect prediction (y = x); the dashed red line in panels E–F denotes zero residual.

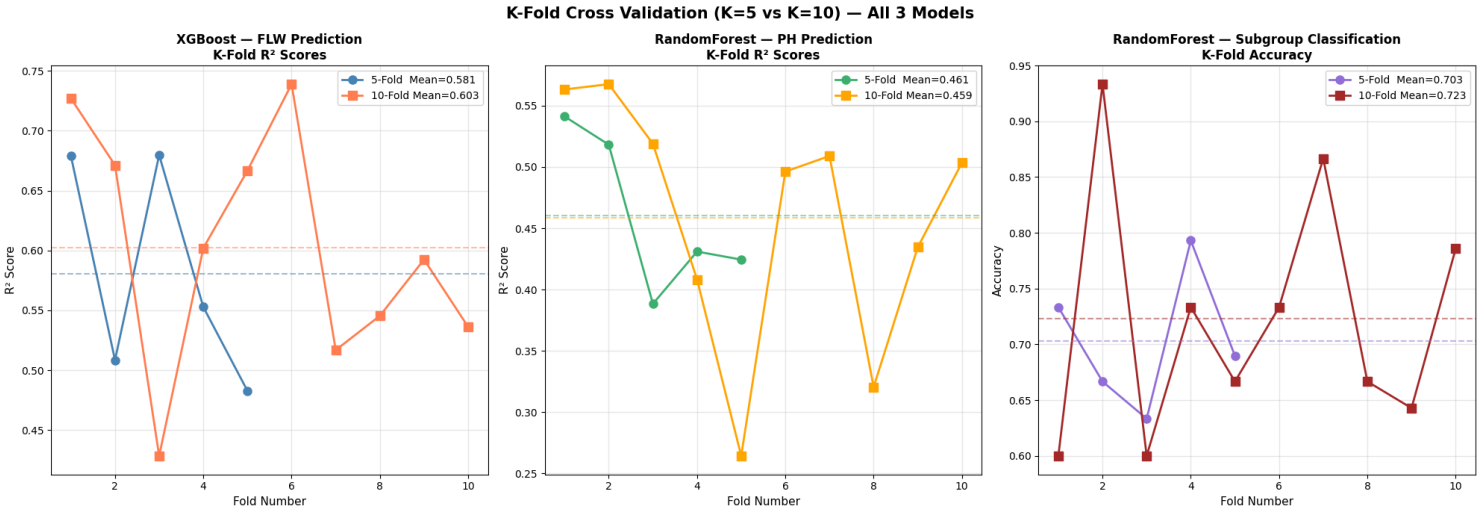

Figure 3

Five-fold versus ten-fold cross-validation of the best-performing model per task. (A) XGBoost FLW prediction (5-fold mean R^2 = 0.581; 10-fold mean R^2 = 0.603); (B) random forest PH prediction (5-fold mean R^2 = 0.461; 10-fold mean R^2 = 0.459); (C) random forest subgroup classification accuracy (5-fold mean = 0.703; 10-fold mean = 0.723). Dashed horizontal lines indicate the mean score for each fold configuration. Close agreement between 5-fold and 10-fold estimates, and with the held-out test-set metrics in Table 2, supports the stability of the reported model performance.

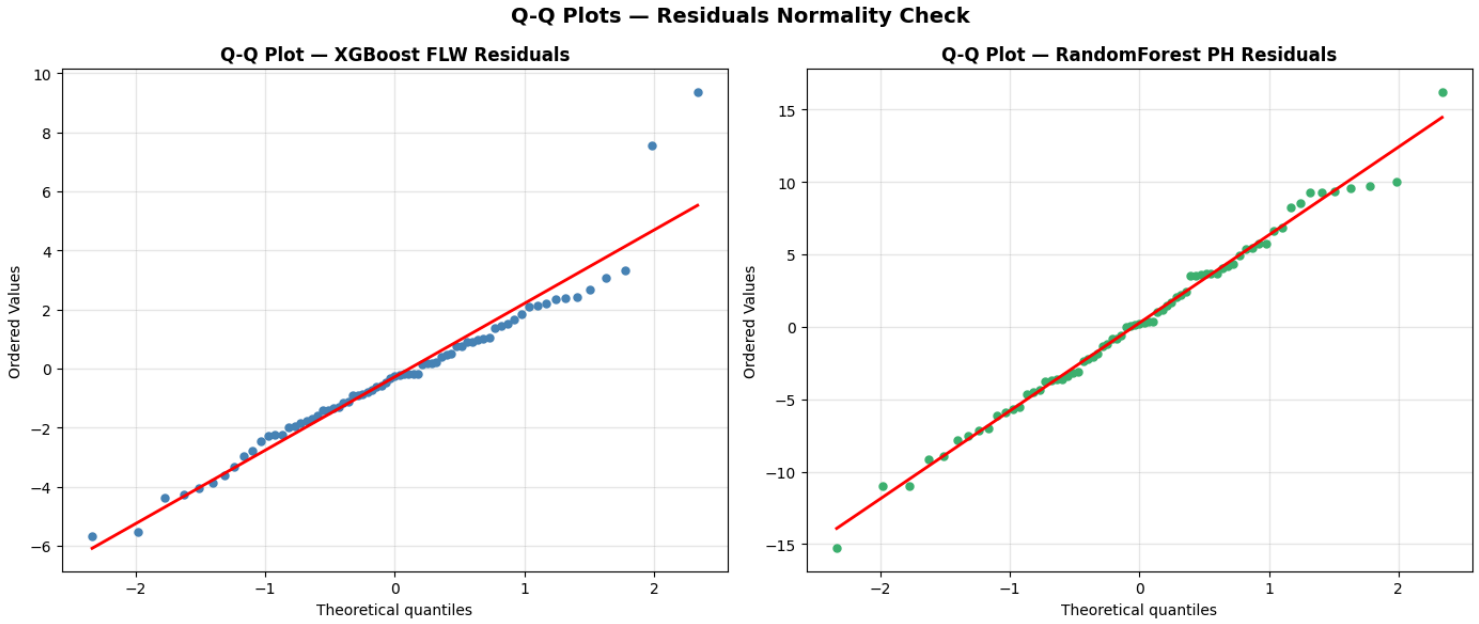

Figure 4

Quantile–quantile (Q–Q) plots assessing normality of model residuals against a theoretical normal distribution. (A) XGBoost FLW residuals; (B) random forest PH residuals. Residuals for both models track the reference line closely across the central quantiles, with mild deviation in the extreme tails, indicating

approximately normal residual distributions consistent with the assumptions underlying the reported R^2 and RMSE estimates.

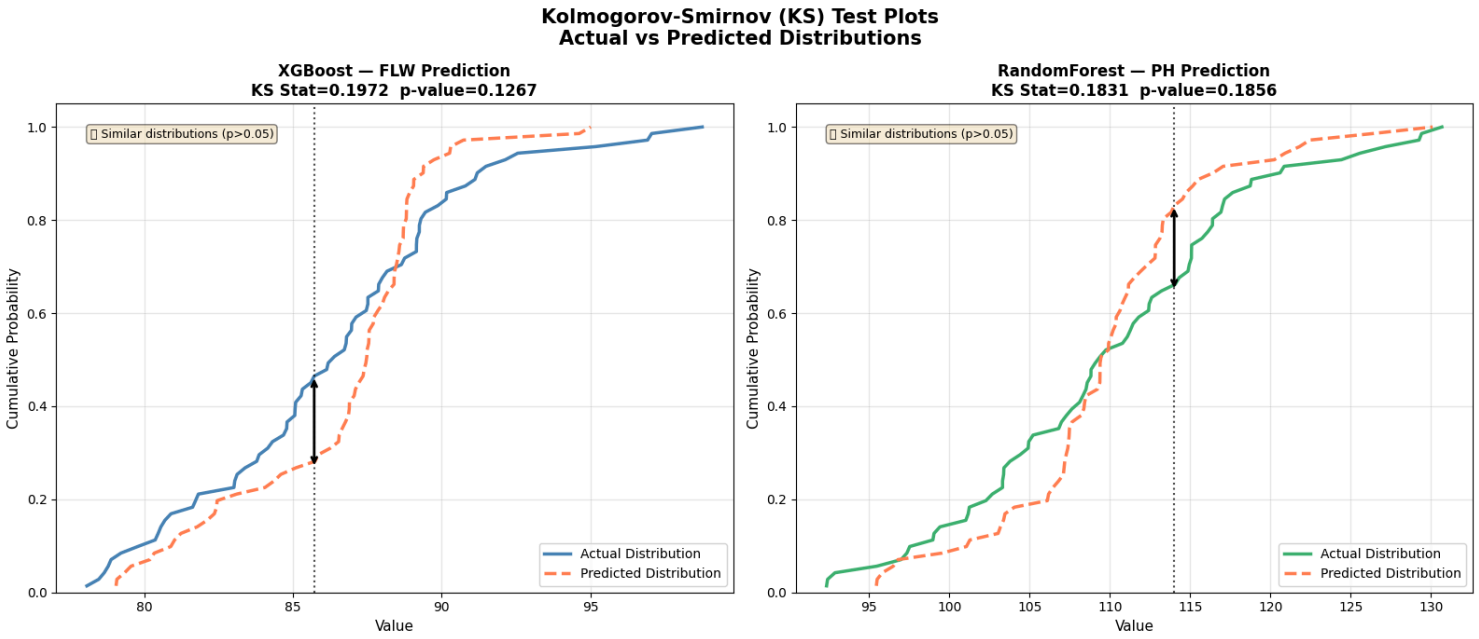

Figure 5

Two-sample Kolmogorov–Smirnov (KS) comparison of actual and predicted trait distributions. (A) XGBoost FLW predictions (KS = 0.197, p = 0.127); (B) random forest PH predictions (KS = 0.183, p = 0.186). In both cases, p > 0.05 indicates no statistically significant difference between the actual and predicted distributions.

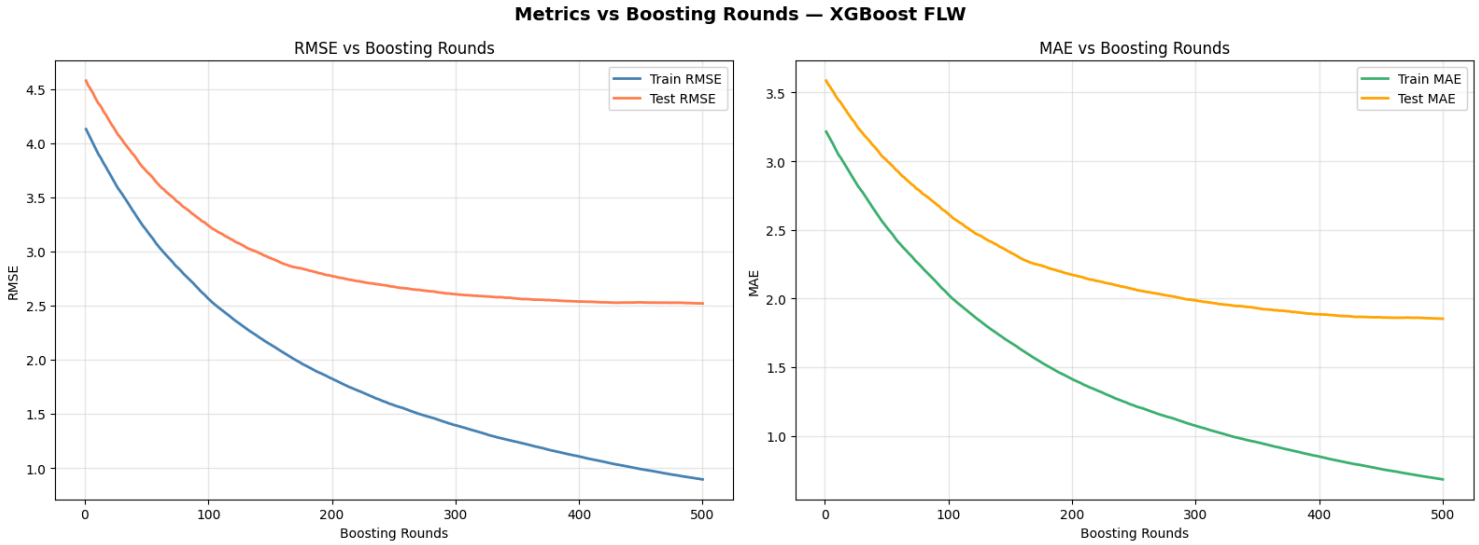

Figure 6

XGBoost training diagnostics for the FLW model across 500 boosting rounds. (A) Root mean squared error (RMSE) for training and test partitions; (B) mean absolute error (MAE) for training and test partitions. Test-

set error plateaus after approximately 200–300 rounds while training error continues to decline, indicating the selected model configuration approximates the bias–variance optimum rather than substantial overfitting.

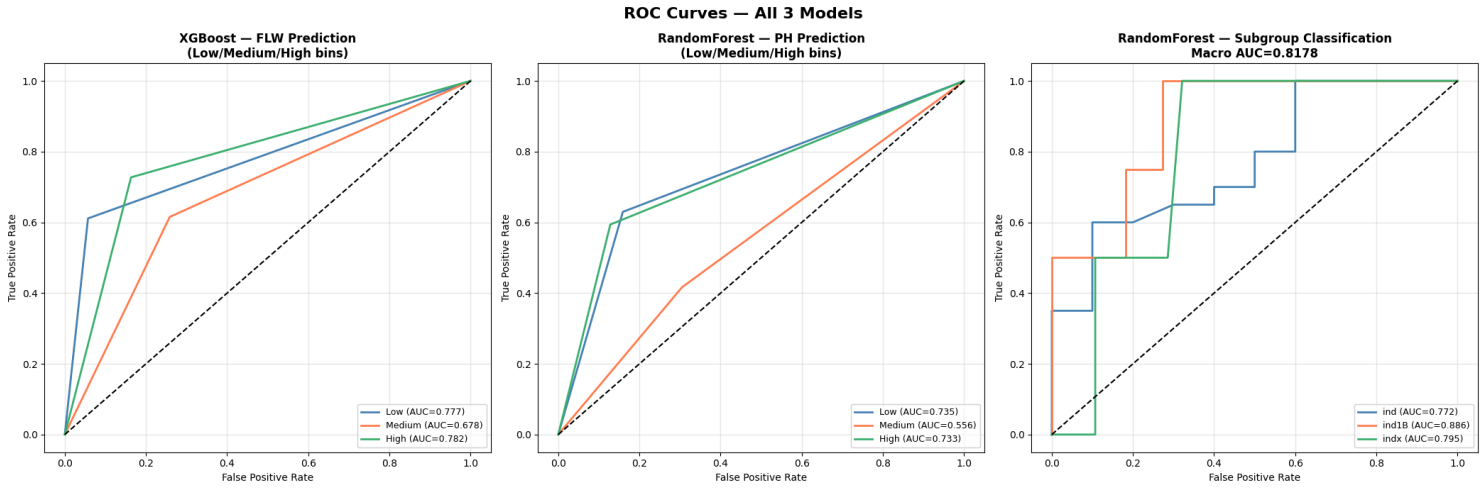

Figure 7

Receiver operating characteristic (ROC) curves for the three prediction tasks, with continuous traits discretized into Low/Medium/High tertile bins for one-vs-rest AUC calculation. (A) XGBoost FLW prediction (AUC = 0.777, 0.678, and 0.782 for Low, Medium, and High bins, respectively); (B) random forest PH prediction (AUC = 0.735, 0.556, and 0.733 for Low, Medium, and High bins, respectively); (C) random forest subgroup classification, one-vs-rest by class (macro-averaged AUC = 0.818; per-class AUC = 0.772 for ind, 0.886 for ind1B, and 0.795 for indx). The dashed diagonal represents random-chance discrimination (AUC = 0.5).

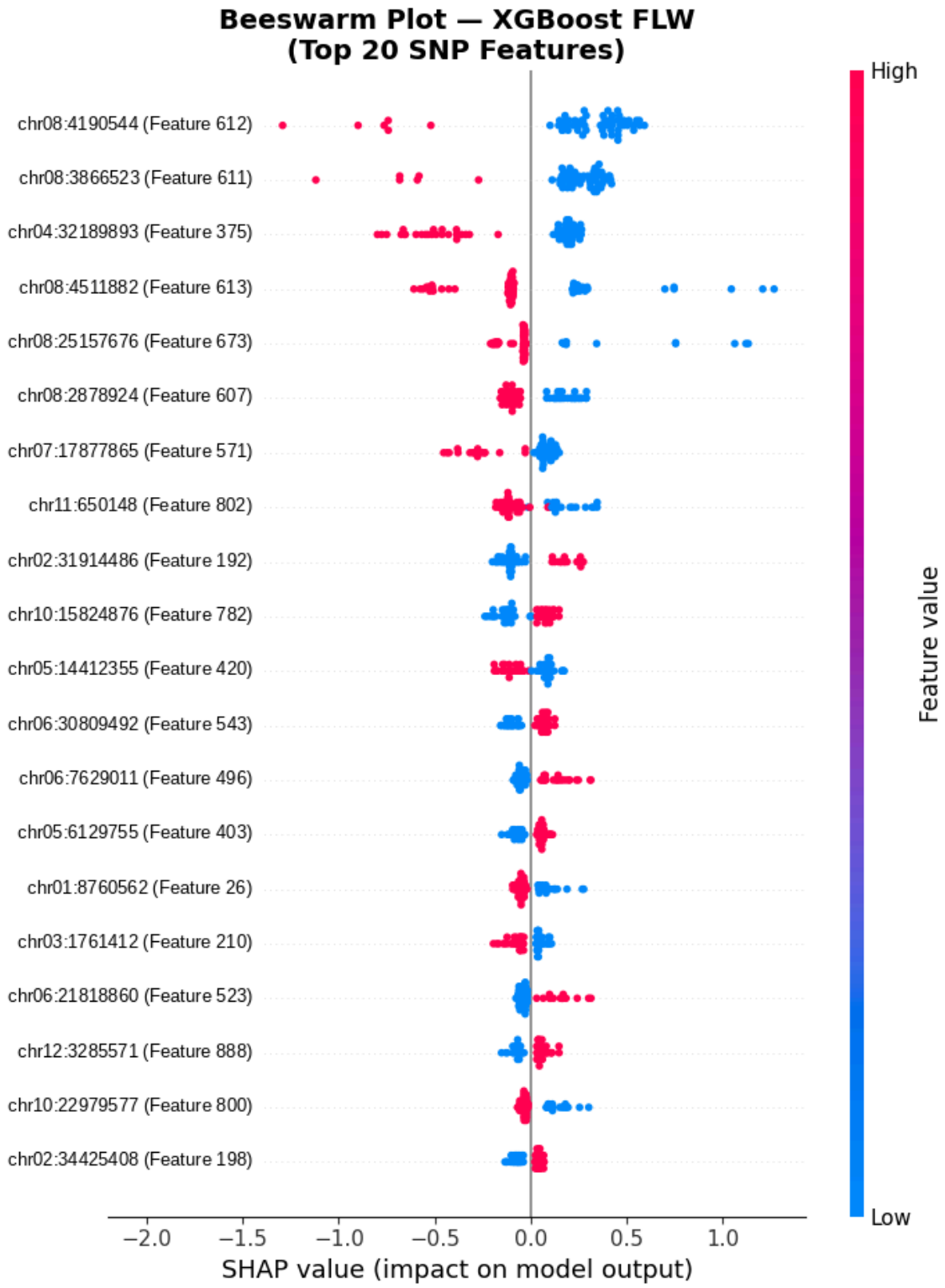

Figure 8

Global SHAP (SHapley Additive exPlanations) summary plot for the top 20 SNP markers contributing to the XGBoost flowering-time (FLW) model, with feature indices relabeled to their corresponding physical marker positions (chromosome:base-pair; mapped from the 1k-RiCA marker order in Arbelaez et al., 2019). Each point represents one accession; horizontal position indicates the marker's SHAP value (impact on predicted FLW) for that accession, and point color indicates the accession's genotype value at that marker (red = high-

value allele, blue = low-value allele). Four of the five highest-impact markers localize to chromosome 8, with three clustered within a ~650 kb interval (3.87–4.51 Mb), suggesting a single linked QTL region as the dominant genomic driver of predicted flowering time.

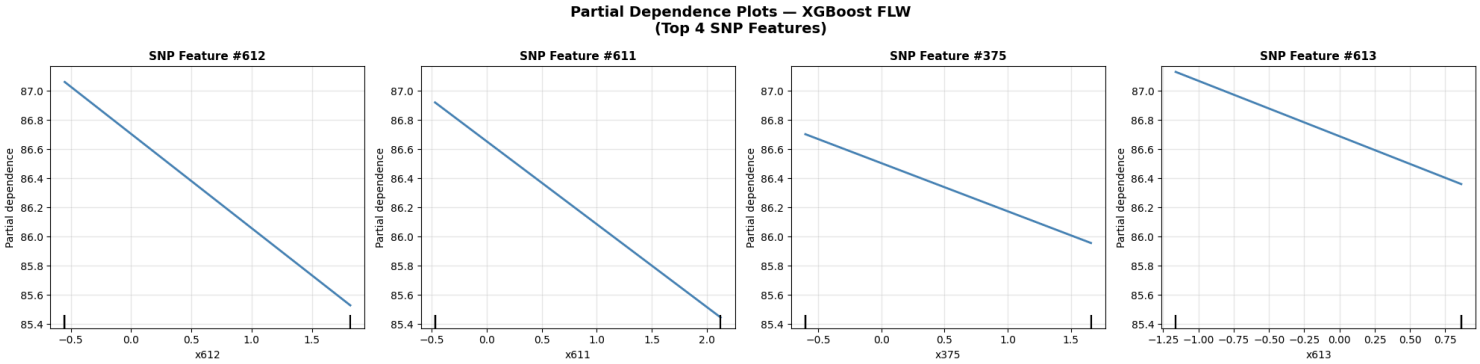

Figure 9

Partial dependence plots (PDPs) for the top four SNP markers contributing to the XGBoost flowering-time (FLW) model (SNP Features #612, #611, #375, and #613, corresponding to chr08:4,190,544, chr08:3,866,523, chr04:32,189,893, and chr08:4,511,882, respectively). Each panel shows the marginal relationship between the encoded genotype value (x-axis) and predicted FLW (y-axis), holding all other markers at their observed distribution; tick marks on the x-axis indicate the deciles of the observed genotype distribution. All four markers show a smooth, monotonically negative dependence, consistent with the directionality of the corresponding SHAP values in Figure 8.

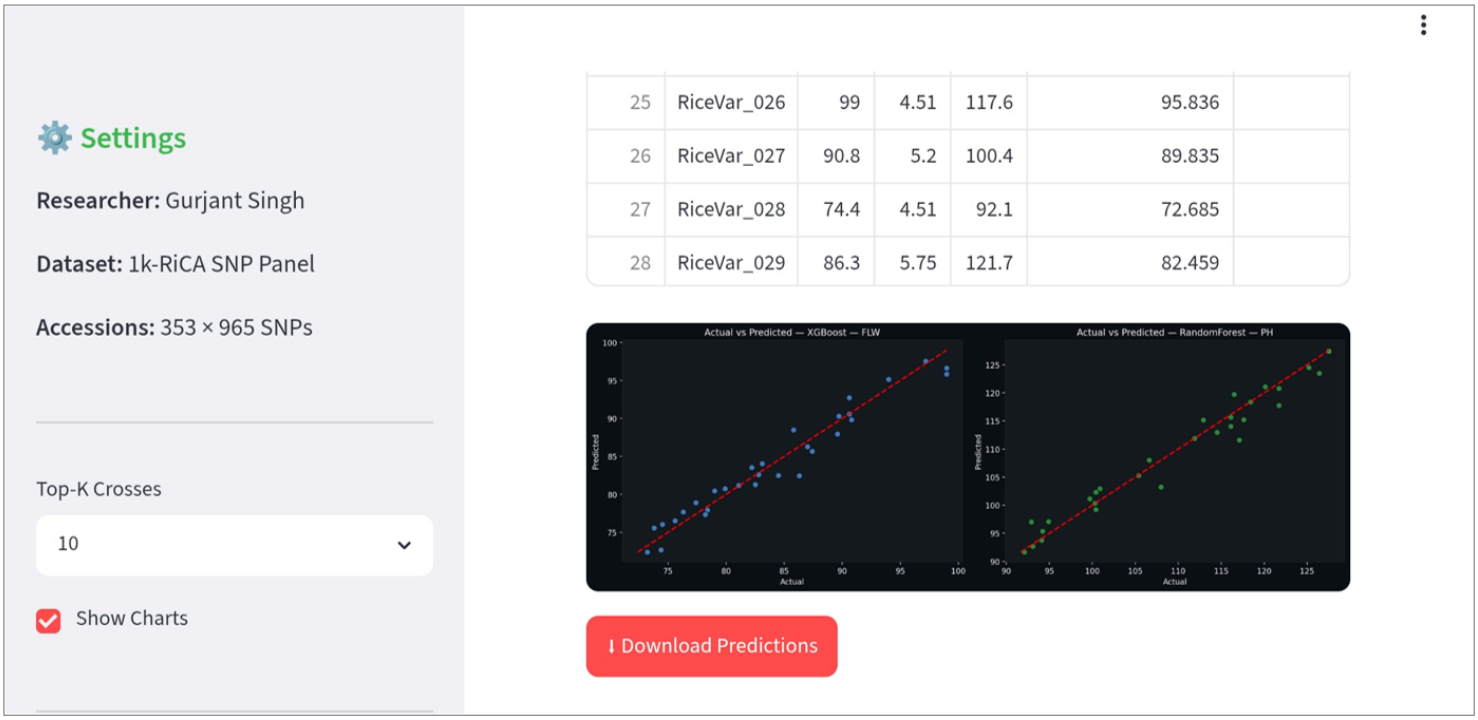

Figure 10

Screenshot of the deployed breeder-facing web application (“Rice Genomic ML System”), showing the Upload & Predict interface with per-accession predicted FLW and PH values alongside observed inputs, and paired actual-vs-predicted scatter plots for the uploaded evaluation set. The application allows CSV upload, automatic preprocessing, and direct export of prediction results for breeder review.