Machine Learning-Based Breeding Values Prediction System (ML-BVPS)

S. V. Vasantha and B. Kiranmai

Abstract Understanding the connection between a genotype and its phenotype is a key challenge in predicting the breeding values. However, genotype-to-phenotype prediction presents significant challenges for machine learning algorithms, limiting their use in this context. The data's high dimensionality makes generalization difficult and limits the scalability of most learning algorithms. The accurate prediction of phenotypes is needful in improving crop breeding. We analyzed GWAS and implemented a strategy (ML-BVPS) for the prediction of rice plant height based on its genotype. We implemented Machine learning algorithms for the classification of rice subpopulation and phenotype prediction. We achieved 94% accuracy in classifying the rice population. We achieved an accuracy range of 0.82–1.0 in the prediction of phenotype value based on its Lead SNP markers. We also recommend genotype and its corresponding GWAS information for each subpopulation category to obtain a better breeding value.

Keywords Genotype · Phenotype · GWAS · Machine learning · Breeding value

1 Introduction

Rice (*Oryza sativa* L.) is a major food crop that provides food for more than half of the world's population [1]. Isozyme analysis [2] was the first widely used molecular classification of rice groups, which found six varietal groups. Following DNA analysis, five subpopulations of rice were discovered within the two Oryza subspecies. The aus subpopulation is a group of people who live in Australia [3]. The most popular rice subpopulation, indica and japonica, are differentiated by genetic information.

Intermediate between wild family and cultivars, it provides important information for crop improvement by creating a genetic reservoir that can adapt to environmental changes and increase crop sustainability. Garris et al. [4] proposed that *O. sativa* can be divided into five different groups based on model-based structure analysis:

S. V. Vasantha · B. Kiranmai (✉)
Department of CSE, KMIT, Narayanguda, India

D. Gupta et al. (eds.), *Proceedings of Data Analytics and Management*,
Lecture Notes on Data Engineering and Communications Technologies 90,
https://doi.org/10.1007/978-981-16-6289-8_22

indica, aus, aromatic, temperate japonica, and tropical japonica rice. As a result, there are still differences in the genetic structures or taxa of cultivated rice species. Each subpopulation has desirable cultivar characteristics, breaking down genetic barriers between them will allow rice breeders to freely introduce desirable genes from different subpopulations into an elite line. This will help to reduce breeding costs and ensure the long-term viability of rice production, which is threatened by ongoing climate change [5, 6].

The reduction of genotyping cost and availability of high-throughput genotyping services has made feasible to deploy various genomic selection approaches in regular plant breeding programs [7]. Oryza Sativa genetic markers are used as a genotyping tool for genetic analysis and marker assisted breeding. The restriction fragment is the first type of molecular marker used in genetic analysis [8].

Genomic Selection (GS) is a plant breeding technique that uses genome-wide marker data to predict the genetic value of untested lines. The technique has been extensively tested in both simulated and real-world plant breeding programs.

The data used in the GS model, such as the size of the training population, relationships between individuals, marker density, and use of pedigree information, has been shown to affect the accuracy of GS [9]. In hybrid breeding, genomic selection (GS) is more effective than traditional phenotype-based methods [10].

Genomic prediction (GP) using single nucleotide polymorphism (SNP) markers has emerged as a useful tool in a variety of human healthcare settings [11]. Singh et al. [12] looked at allelic sequence variation in three Sub1 genes in a panel of 179 rice genotypes and came up with some interesting results.

Kim [6] used phenotyping variables and the logistic regression model (LRM) to classify indica/japonica. They tailored LRM to classify indica and japonica based on seven phenotypic factors. Jeon et al. [13] used GP modeling to create a prediction model with the best marker set. The interaction effects of the SNP markers are considered indirectly through modeling, and statistical, machine, and/or deep learning techniques are used and put to the test on actual datasets with four different phenotypes. The prediction results were better than those obtained by using the entire set of markers or the GWAS-top markers, which are commonly used in prediction studies. Despite its flaws, the GMStool is expected to help researchers predict quantitative phenotypes in a variety of studies. Li et al. [14] noticed that the RF and particularly GBM algorithms are efficient in identification of SNPs subset having direct correspondence to the candidate genes that are responsible for growth trait.

Wang et al. [10] assess the predictive ability of GS for hybrid performance in rice using the NC II scheme, in which 115 inbred rice lines were crossed with 5 male sterile lines, and to show how predicting characteristics can be used to predict potential crosses between the 115 inbred lines and other genotyped varieties. They have worked on GBLUP for predicting genetic values and phenotypes using whole genome markers. Yan et al. [11] devised subpopulation classification using Phylogenetic Tree and predicted Genomic Estimation Breeding values for four phenotypes.

2 Problem Statement

Significant work carried out for predicting Breeding Values of *O. Sativa* populations such as, Yan et al. [11] solution for predicting *O. Sativa's* Genomic Breeding values and its accuracy highly varied from 0.23 to 0.90. Prediction accuracy of solutions presented by Joen [13] and [10] are 0.52 and 0.88, respectively. Thus, there is a need for improving the prediction accuracy of Breeding Values (BVs) of Oryza species. Hence, the ML based Breeding Values Prediction System (ML-BVPS) for O. Sativa is designed.

3 Proposed Model

It mainly focuses on improving prediction accuracy of Breeding Values (BVs) in *O. Sativa.* It involves a strategy that initially finds out the type of subpopulation of each given sample. And then based on the identified type, it predicts the BVs. It also recommends genotypes for better breeding values for each subpopulation type.

The proposed model comprises of three main modules.

(i) Classification of Oryza samples based on subpopulation types
(ii) BVs Prediction
(iii) Genotype Recommendation system for Gene Editing.

Data Gathering and Preprocessing

Rice Cultivars data related to phenotype, GWAS, and its genotype information is collected from the Data-Source: [15] http://ricevarmap.ncpgr.cn/. Phenotype values and GWAS information are extracted for the plant height. Next its corresponding imputed genotype information is extracted. Datasets are prepared by normalizing and scaling the extracted data for further classification and BVs prediction tasks.

Classification of Oryza samples based on subpopulation types

This module deals with classification of given *O. Sativa* samples based on its subpopulation type. It uses imputed SNP markers which are spread across twelve chromosomes. Usually, SNP markers, which improve classification accuracy, are selected for identifying the type of subpopulation, but the proposed model is designed to improve BVs prediction accuracy, thus LEAD SNP markers specific to phenotype are selected.

Various popular ML algorithms are applied to classify the given cultivars such as logistic regression, Naive Bayes, decision trees, random forests, and many more. Among these algorithms, random forest classifier exhibited most promising results. Data is partitioned into training sets and testing sets out of which, 80% is used for training and 20% is used for testing.

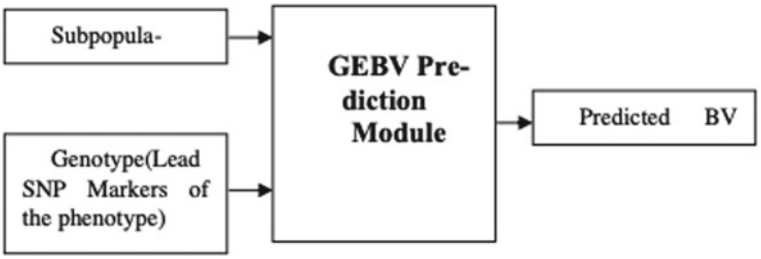

Fig. 1 BV prediction module

BVs Prediction

This module is designed to predict BVs of the given samples with improved prediction accuracy, which is depicted in the Fig. 1. It takes genotype, that is, the Lead SNP markers set specific to the phenotype to be predicted along with subpopulation type predicted in first module as inputs. These inputs are processed by ML algorithms, to predict a phenotypic value range. Here the ML model is trained based on the genotype along with its formulated phenotypic value range of the samples for each subpopulation category so as to improve BV prediction accuracy.

K-means Algorithm is used to formulate phenotypic value range of the samples. Various well known ML techniques are applied to identify appropriate phenotype range based on the given sample's genotype. 80% of dataset is considered for training and 20% of it is considered for testing. Random forest classifier expressed most promising prediction accuracy among other ML techniques.

Genotype Recommendation system for Gene Editing

This module provides suggested genotype for given phenotype with respect to each subpopulation category, which is shown in the Fig. 2. Here one of the significant phenotypic traits of *O. Sativa* known as plant height is considered and for which the genotype is recommended for better BV.

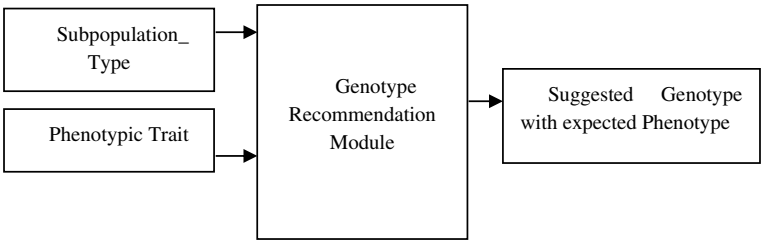

Fig. 2 Genotype recommendation module

Suggested genotype is framed based on the subpopulation's best samples with respect to given phenotype. Thereby, it facilitates Gene editing by providing its corresponding GWAS information.

4 Results Discussion and Comparative Analysis

This section gives details about normalization and scaling of data, results of subpopulation classification and BV prediction and also their comparative analysis with other ML techniques and some contemporary models.

(i) Data Normalization and Scaling

The genotype data is normalized as follows:

A—1, C—2, G—3, T—4, N—5, H—6

Plant height value ranges are formulated with respect to each subpopulation as follows:

Aromatic	129, 130–158, 160–189
Aus	111–159, 161–190
Tropical Japonica	84–141, 143–183
Temporate Japonica	74–98, 99–129, 131–207
Indica	69–132, 133–204

(ii) Subpopulation Classification Results

The subpopulation classification accuracy of given *O. Sativa* samples when RF Classifier is applied along with the parameters tuning information is expressed in below Table 1.

Its Classification Report is as follows:

Table 1 Subpopulation classifier accuracy

Test size	Random_state	Classifier_Accuracy	Classifier prediction accuracy of test data
0.1	800	0.94	0.84
0.15	800	0.93	0.81
0.2	800	0.91	0.78

	precision	recall	f1-score
Aus	0.80	1.00	0.89
Indica I	0.83	0.83	0.83
Indica II	0.82	0.90	0.86
Indica III	0.00	0.00	0.00
Indica Intermediate	0.71	0.56	0.63
Intermediate	1.00	0.50	0.67
Japonica Intermediate	1.00	0.67	0.80
Temperate Japonica	0.92	1.00	0.96
Tropical Japonica	1.00	1.00	1.00
VI/Aromatic	1.00	1.00	1.00
accuracy			0.85
macro avg	0.81	0.75	0.76
weighted avg	0.87	0.85	0.85

(iii) **Comparative Analysis of Subpopulation Classifier Accuracy**

Various popular ML techniques such as random forest, neural network MLP, ridge classifier, ridge classifierCV, Gaussian Naïve Bayes, and decision tree are applied for classifying samples based on the type of subpopulation. Classifier Accuracy of these techniques are expressed in the Figure 3. Random forest classifier is exhibiting the better accuracy compared to other techniques.

(iv) **Phenotype Prediction Results**

Phenotype value of plant height trait is predicted for the whole population, but its accuracy is comparatively less than the accuracy of each subpopulation classified by its type and mean accuracy of all the subpopulations is 0.92. Phenotype prediction accuracy of each subpopulation and whole population along with parameter tuning information is expressed in the Table 2

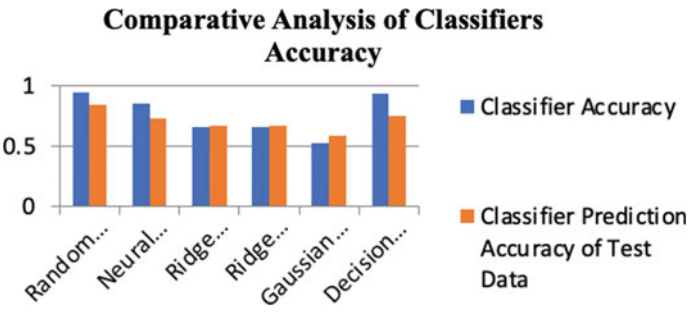

Fig. 3 Comparative analysis of classifiers accuracy

Table 2 Phenotype prediction accuracy of subpopulations

Subpopulation type	Test size	Random_State	Classifier accuracy	Classifier prediction accuracy of test data
Indica	0.25	350	0.99	1.0
Aromatic	0.25	50	1.0	1.0
Temporate Japonica	0.2	350	0.91	0.82
Aus	0.25	500	0.93	0.91
Tropical Japonica	0.2	1800	0.95	0.88
Whole population	0.25	350	0.85	0.57

Table 3 Comparative analysis of various ML techniques

Classifier/regressor name	Classifier/regressor accuracy	Prediction accuracy of test data
Random forest classifier	0.99	1
Ridge classifier	0.78	0
Ridge classifierCV	0.73	0
Logistic regressor	0.65	0.63
GLM	0.45	0.75

(xxii) Comparative Analysis of Phenotype Prediction Accuracy

Plant height phenotype is predicted by applying various ML techniques such as random forest classifier, ridge classifier, ridge classifierCV, logistic regressor, and GLM. Among all these techniques, random forest classifier expressed most promising prediction accuracy which is shown in Table 3.

Comparative Analysis of Phenotype Prediction Accuracy with contemporary solutions is expressed in Table 4. Our ML-BVP method, mean prediction accuracy is outperforming when compared to other solutions. Prediction accuracy of ML-BVP ranges from 0.82 to 1.0 which shows drastic enhancement when compared to [11] solution.

Table 4 Comparative analysis of contemporary techniques

Method	Prediction accuracy
ML-BVP (Proposed)	0.92
Joen et al.	0.52
Wang et al.	0.88
Yan et al.	0.2–0.9

5 Conclusion

The proposed ML-BVP system is used to classify *O. Sativa* samples based on their population type, the same is used as a key input for predicting BVs. It is observed that the prediction accuracy is improved over whole population samples and other solutions. RF exhibited most promising results when compared other ML models.

References

1. Khush GS (2005) What it will take to feed 5.0 billion rice consumers in 2030. Plant Mol Biol 59(1):1–6
2. Glaszmann JC (1987) Isozymes and classification of Asian rice varieties. Theor Appl Genet 74(1):21–30. https://doi.org/10.1007/BF00290078 PMID: 24241451
3. Ahmed ZU, Panaullah GM, Gauch H, McCouch SR, Tyagi W, Kabir MS, Duxbury JM (2011) Genotype and environment effects on rice (*Oryza sativa* L.) grain arsenic concentration in Bangladesh. Plant Soil 338(1):367–382
4. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169(3):1631–1638
5. Kim B (2018) Classifying Asian rice cultivars (*Oryza sativa* L.) into Indica and Japonica using logistic regression model with publicly available phenotypic data. bioRxiv. 1 Jan 2018 (470351)
6. Kim B (2019) Classifying *Oryza sativa* accessions into Indica and Japonica using logistic regression model with phenotypic data. PeerJ 7:e7259
7. Vinayan MT, Seetharam K, Babu R, Zaidi PH, Blummel M, Nair SK (2021) Genome wide association study and genomic prediction for stover quality traits in tropical maize (*Zea mays* L.). Sci Rep 11(1):686. Published 12 Jan 2021. https://doi.org/10.1038/s41598-020-80118-2
8. Tuhina-Khatun M, Hanafi MM, Rafii Yusop M, Wong MY, Salleh FM, Ferdous J (2015) Genetic variation, heritability, and diversity analysis of upland rice (*Oryza sativa* L.) genotypes based on quantitative traits. BioMed Res Int
9. Robertsen CD, Hjortshøj RL, Janss LL (2019) Genomic selection in cereal breeding. Agronomy 9(2):95
10. Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, Hu Z (2017) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity 118(3):302–310
11. Yan J, Zou D, Li C, Zhang Z, Song S, Wang X (2020) SR4R: an integrative SNP resource for genomic breeding and population research in rice. Genomics Proteomics Bioinform 18(2):173–185
12. Singh A, Singh Y, Mahato AK, Jayaswal PK, Singh S, Singh R, Yadav N et al (2020) Allelic sequence variation in the Sub1A, Sub1B and Sub1C genes among diverse rice cultivars and its association with submergence tolerance. Sci Rep 10(1):1–18
13. Jeong S, Kim JY, Kim N (2020) GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci Rep 10(1):1–12
14. Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 9:237. https://doi.org/10.3389/fgene.2018.00237
15. http://ricevarmap.ncpgr.cn/
16. http://variation.ic4r.org/
17. http://www.ricediversity.org/
18. http://ricepedia.org/rice/rice-as-a-plant/rice-species