**RESEARCH PAPER**

**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 generalisation difficult and limits the scalability of most learning algorithms. The accurate prediction of phenotypes is needed to improve crop breeding. We analysed 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 height prediction. We achieved 75% accuracy in classifying the rice population. We achieved an accuracy range of 0.64-0.76 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

**INTRODUCTION**

Rice (Oryza sativa L.) is a major food crop that provides food for more than half of the world’s population [1].The types of Rice varieties in total are 6 in which 5 are cultivated rice subpopulation and 1 is wild rice subpopulation

* Indica Rice
* Aus Rice
* Aromatic Rice
* Tropical japonica rice
* Temperate japonica rice
* Rufipogon

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:Indica,Aus,Aromatic,Temperate japonica,Tropical japonica.As a result, there are still differences in the genetic structures or taxa of cultivated rice species.Each subpopulation has different desirable genotype and phenotype characteristics which can helpful in rice production in different ways.The most popular rice subpopulation, indica and japonica, are differentiated by genetic information.

Genomic data is used for prediction of phenotypic traits of Oryza sativa.It’s genetic markers are used as a genotyping tool for genetic analysis and marker assisted breeding. **Genomic selection (GS)** is a method to predict the genetic value of selection candidates based on the genomic estimated breeding value (GEBV) predicted from high-density markers positioned throughout the genome.The use of genomic selection in plant breeding has several advantages over traditional breeding methods, such as allowing breeders to select for traits that are difficult or expensive to measure, increasing the accuracy and speed of selection, and allowing predictions to be made before the plants are fully grown. This can greatly reduce the time and resources needed to develop new plant varieties with desirable traits, which is especially important in the face of global challenges such as climate change, population growth, and food security.

**Genomic prediction (GP)** using single nucleotide polymorphism (SNP) markers has emerged as a useful tool in plant breeding.**Genomic prediction** provides an alternative method to use genomic information in breeding decisions (Meuwissen et al., 2001). Instead of using only significant marker-trait associations to build up the prediction model, genomic prediction uses all markers simultaneously (Windhausen et al., 2012).

**Phenotype Prediction**

Classification and Regression Models are used for prediction of Data in plant breeding.

Like Linear Regression,Random Forest Regressor,SVM,Random Forest Classifier,XG Boost,

Etc.can be used.

**Project flow**

Building a **GWAS** file mapping Genotype to Phenotype followed by a phylogenetic tree and classifying subpopulation and prediction of height using Machine Learning Models.

At last we built a UI where users can get the results by giving DNA sequences.

**Data**

Rice cultivars data related to phenotype, GWAS, and genotype information is collected from the data source: <https://ricevarmap.ncpgr.cn/>. Phenotype values and GWAS information are extracted specifically for plant height. Imputed genotype information is then extracted, and missing data is replaced with suitable values. Datasets are prepared by mapping phenotype information with the corresponding genotype information. The resulting data is then checked for imbalanced classes. If imbalanced classes are found, SMOTE (Synthetic Minority Over-sampling Technique) is applied to address the imbalance.