GEBV Phenotype Trait Height Prediction of Oryza Sativa using ML & DL Techniques

Dr.B.Kiranmai*[1], Harshith. K[2], Lasya Priya.Y[2], Vivek.V[2], Soumya.K[2], Kavya .B[2]

[1] Sreyas Institute of Engineering and Technology, Bandlaguda

[2] KMIT, Narayanguda-2, India[1,2]

kiranmamtech@gmail.com,[1] 21bd1a054rcsec@gmail.com, lasyayelishetti@gmail.com, vivekvemula21@gmail.com, ksowmya1702@gmail.com, borgamkavya17@gmail.com

Abstract. Plant breeding is crucial for improving crops with desirable traits, and phenotypic prediction plays a key role in this process. Accurately predicting plan height is a challenge but is vital for optimizing crop management, maximizing yield, improving disease resistance, enhancing environmental adaptation, ensuring efficient harvesting and streamlining breeding programs.

Our analysis utilized GWAS and implemented Machine Learning (ML) and Deep Learning (DL) Models. We achieved more than 75% accuracy in classifying rice subpopulation and an accuracy range of 0.64 to 0.76 for predicting rice plant height based on genotype information. Using these models will enhance productivity and sustainability of agriculture. With this methodology and results one can reasonably estimate that deep learning models are best for predicting large data and machine learning models are reasonably good for small set of data.

Keywords: Genotype, Phenotype ,GWAS ,Machine learning ,Breeding value

1 Introduction

Rice (Oryza sativa L.) is a crucial food crop that sustains more than half of the global population.[1].O. sativa is botanical name for rice and they are categorized in to wild rice and cultivated subpopulations. The Wild rice subpopulation will adjust to climatic conditions and that will act as a genetic factor for cultivated varieties of rice. [2] . Broadly classified rice types based on its subpopulation and they are: Indica, Aus Aromatic , Tropical japonica rice, Temperate japonica , Rufipogon (wild rice subpopulation) [1]. Based on model-based structural research, Garris et al. [3]

proposed a categorization of O. sativa into five separate groups: Indica, Aus, Aromatic, Temperate japonica, and Tropical japonica. Among these subpopulations genetic and taxonomic differences arise in this subpopulations.

Each subpopulation has unique genotype and phenotype traits, which can be beneficial in rice production in various ways. The most well-known rice subpopulations, Indica and Japonica, can be differentiated based on genetic information [4]. To identify the rice subpopulations genome -wide DNA markers such as single nucleotide polymorphisms to predict the genetic trait of undiagnosed lines [2].

Genomic prediction (GP) utilising single nucleotide polymorphism (SNP) markers has emerged as a valuable tool in plant breeding. Genomic prediction provides an alternative approach to incorporating genomic information into breeding decisions.
[Meuwissen et al., 2001] [5].

Jun Yan [1] implemented phylogenetic analysis using the 156,502 tagSNPs was performed on the 2556 rice accessions, which were explicitly documented with subpopulation classification and origins.

Factors that affect prediction accuracy include the number of markers used for estimating the GEBVs trait heritability, calibration population size, statistical models, number and type of molecular markers, linkage disequilibrium and effective population size [6]. Generally, larger TRSs (training sets) are required for traits controlled by more genes with smaller effects [10].

Knowing the predictability of a model is one of the important constraint for a better utilization of resources in plant breeding, especially of high cost for predicting phenotyping traits [6].

To predict the trait authors have experimented Artificial Neural Networks Using TensorFlow in Python. Each layer consists of genetic info and are connected to previous and following nodes exempting the final layer. The final layer is connected to the output layer which predicts the trait. [7].

Weights were optimized by using the Adam Optimizer to determine number of epochs training set was further divided for validation set used An early stopping approach was used, where the training set was further divided into training and validation, and early stopping occurred when the change in mean squared error (MSE) for the validation set was in the range of 0.3% to 1%, in order to determine the optimal stopping time for training (i.e., number of epochs). [7].

1.1 Problem statement

J.L .J and J.P [6] have done comparative analysis of genotyping by sequence markers which gave predictive ability around 60 to 70%. Further improvement is required by applying various model analysis and techniques. Kiranmai [2] applied ml-bvps method for genomic data and gained the accuracy of 78% on test data which isn't mapping to the classifier accuracy [4]. There is a need to improve predictive ability of breeding values of O. Sativa ensuring it is robust to new data.

Azodi [7] applied ANN by imputing seeded weights and claimed the corrcoef of 0.5 and mean accuracy falling between 50 and 60 for rice data. Further improvement is required. So, we developed our model GEBV-DL for that, We balanced our data, a pattern formed that could be beneficial for the neural network model whether to classify or predict the target variable.

2 Proposed Methodology

Our proposed methodology is depicted in Fig. 3. Here we are implementing biological classification and followed by Machine Learning / Deep Learning Techniques. Initially data is prepared and GWAS file is generated. in the second step Phylogenetic tree is constructed using UPGMA , on this Machine learning techniques or Deep Learning Techniques are applied . A comparison of Machine Learning and Deep Learning techniques are illustrated in Table 1.

(i) Data collection

The data for rice cultivars, comprising phenotype, GWAS, and genotype information is collected from [11] http://ricevarmap.ncpgr.cn/. The dataset specifically includes phenotype values and GWAS information related to plant height. Imputed genotype information is also obtained from the database. We then associated phenotype information with the corresponding genotype data.

(ii) Data cleaning and preprocessing

Sequences contains "N" value which represent different types of missing or uncertain genotypes was replaced by corresponding primary allele. All the "DEL" values in the dataset which represent null values or missing values are replaced with their corresponding secondary alleles respectively, during the data cleaning. Phylogenetic analysis is performed, and clustered data is extracted to obtain a structured representation of the data. Now the encoding is performed by replacing the genomic data which contains A, T, C, G with 'A':1.5,'T':2.5,'C':0.5,'G':0.75 respectively in the sequences.

Handling class imbalance, enhance the model training process and improve model perfor mance. And to address the class imbalances within the dataset, they are carefully examined, and the Synthetic Minority Over-sampling Technique (SMOTE) is applied. This approach enhances the representation of the minority class by reducing the bias

towards the majority class, improving the model's ability to learn and make predictions on both the majority and minority classes more effectively.

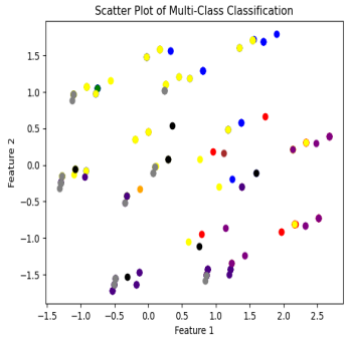


Fig 1 Before Class Imbalance.

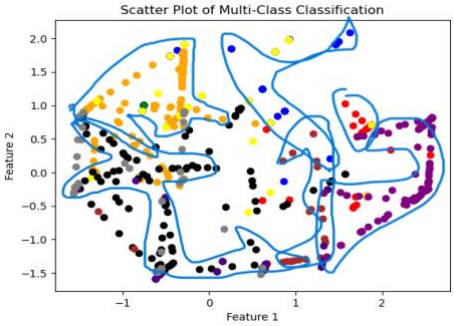


Fig .2 After Class balance

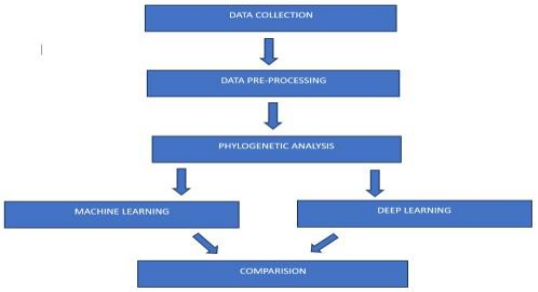


Fig. 3 Flow of the Project

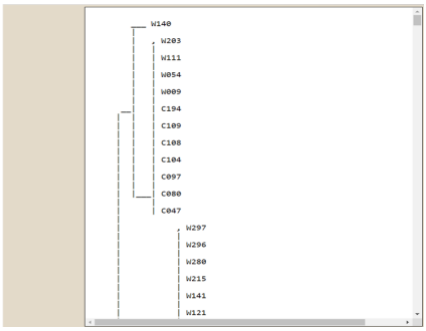


Fig.4 Sampled Biological clusters

2.2 Phylogeny[12]

To study the evolutionary relationships among the species, phylogenetic analysis is performed based on their SNP-containing sequences. First, the sequences are aligned using ClustalW. Then, the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [17] , a distance-based method, is applied to construct the phylogenetic tree. This method groups the sequences into clusters based on their pairwise distances or similarities. After this process, clusters are extracted and labeled. Based on the extracted clusters, the original dataset is reorganized to aid model training." Sample part of the tree is shown in Figure. 4.

2.3 Models and Classification

2.3.1 Using Machine Learning

After data cleaning, we encoded the DNA sequences and subpopulations as target variables in our study. Our goal was to classify the subpopulations of the Oryza sativa species using various multiclass classification techniques. Before applying any predictive technique, we split the data into training and test sets. In our case, the test size for all models was set to 0.33, which corresponds to 33% of the entire dataset. We applied several classification models, including XGBoost Classifier, LogisticRegression (Multinomial), Random Forest Classifier, One-vs-One Classifier, and One-vs-Rest Classifier using the SVC kernel. Among these models, the Random Forest Classifier and XGBoost Classifier performed well, yielding satisfactory accuracy compared to the other models. These models demonstrated promising results in accurately classifying the subpopulations of the Oryza sativa species.

Training and Testing

We proceeded to train our random forest model using the training data and carefully tuned the parameters of the classifier to mitigate overfitting. Similar steps were taken for the other classifiers as well, as illustrated in Fig. 1a. In the case of random forest, selecting the appropriate number of trees and depth is crucial to prevent overfitting, although random forest is generally less prone to overfitting compared to other models. While GridSearchCV can be used to automatically find the best parameters for any model, it may be computationally expensive for large datasets. Therefore, it is recommended to manually tune the parameters.

The accuracy achieved by the random forest model on the test data was 75%, which may not be considered high accuracy. Moreover, it is majority to note that the model is robust and neither overfits nor underfits the given data. High accuracy may not be sufficient to be a better model if it overfits the data, and it may not perform well on data which has not trained. Therefore, achieving a balanced and better robust model is a challenge for predicting unknown traits.

Evaluating

To avoid this overfitting and to achieve high performance we have deployed a model for trained models. we employed the repeated k-fold cross-validation technique. This technique is dividing the dataset into k splits and iteratively test on one split and test on remaining splits. By repeatedly executing this process, we obtain a more reliable estimate of the model's accuracy that considers the whole dataset and identifies any hidden relationships within it. Performance of different models which we have used is being shown in Table. 1.

.

Table 1: Evaluation metrics of various models

Model	Test-Size	Test-Accuracy	Precision	Recall
XG-Boost Classifier	0.33	73.9	71.42	73.9
Onevsrest Classifier (SVM)	0.33	73.64	68.3	67.09
Logistic Regression	0.2	70.51	72.14	70.51
Random Forest	0.2	76.49	76.79	76.49
ANN (Artificial Neural Network)	0.33	73	-	-
KNN	0.2	74.78	76.23	74.78

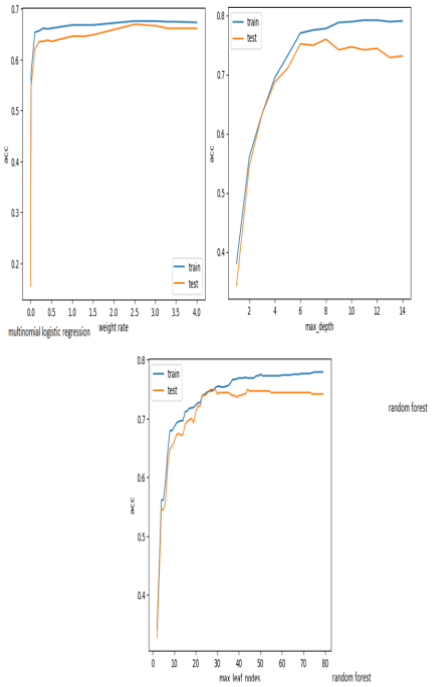


Fig. 5 Performance of multinomial logistic regression and random forest which are being plotted across the hyperparameters

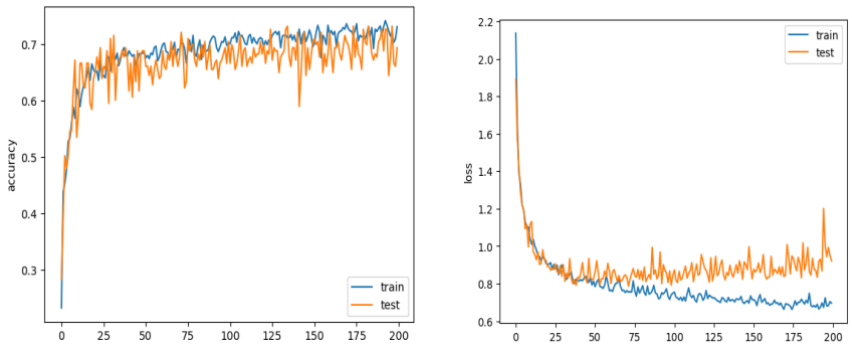
2.4 Classification using ANN :

We employed deep learning techniques to capture the intricate relationships present in DNA sequences. However, it is worth noting tha neural networks typically require a substantial amount of data to outperform traditional machine learning algorithms and produce superior results.

In our study, we developed a relatively simple multi-layer perceptron consisting of only two hidden layers. The architecture of the neural network is defined as (9, 16, 16, 10), where the number 10 represents the number of neurons (which corresponds to the number of classes) in the final layer, utilising the softmax activation function. The remaining layers employ the rectified linear unit (ReLU) activation function. We incorporated kernel initializers such as he_normal and glorot_uniform, depending on the specific layers. Determining the optimal number of neurons and parameters is crucial as it directly impacts the accuracy of the model. We conducted experiments by

trying different combinations of neurons and parameters and selected the best performing configuration. Alternatively, techniques like RandomizedSearchCV [13] can be utilised to automatically search for the optimal hyperparameters, although this approach can be time-consuming.

.The performance of our model across different epochs is illustrated in **Fig. 6a** and **Fig. 6b.**



3. Height Prediction

3.1 Using Deep learning:

We utilised deep learning to predict the height of Oryza sativa (rice). However, traditional machine learning techniques yielded superior results compared to deep learning, primarily due to the requirement of large datasets for neural networks. Additionally, neural networks are prone to overfitting, wherein they become excessively specialised to the training set. To decrease overfitting, we employed dropout regularisation and determined the optimal parameters using RandomizedSearchCV. As a result, we obtained a mean absolute error of approximately 17cm, which falls short of the performance achieved by traditional machine learning algorithms.

The neural network architecture we used consisted of three layers, with 10, 8, and 1 number of neurons respectively. All activation functions were RELU, and the initializers were normal and glorot uniform. We evaluated by checking the correlation coefficient between predicted and real values which rises to be 0.64.

The neural network which we have used is a simple one, but it is sufficient to demonstrate the challenges of using deep learning for rice height prediction. The small dataset size and the high risk of overfitting make it difficult to achieve good results with deep learning. However, as the size of rice height datasets increases and better regularisation techniques are developed, deep learning may become a more viable option for predicting rice height. The performance of the model is being shown in Fig. 3a.

3.2 Using Machine Learning

In our paper our aim is to predict height based on subpopulation and gene sequence. In order to achieve we implemented biological method for classifying asubpopulations and After classifying the subpopulations, the height of each subpopulation was predicted. The obtained values of DNA sequence, subpopulation, and height of subpopulation were used as the target variables. In this paper , a we employed different range of predictive models to estimate the height. Multiple models such as Deep learning techniques and Machine learning techniques were used to comprehensively explore and analyze the prediction of height. As height is a continuous variable, several regression models were employed.

A comparative analysis of multiple models was implemented to identify the most effective model for predicting height. The results has shown that the XGBoost regressor constantly produced the minimum error rate across multiple runs. For our data XGBoost may have emerged to a certain prediction pattern and is less sensitive to the outliers of the dataset. On the other hand, models such as RandomForestRegressor [14] and LinearRegression [15] shown varying errors, indicating their unreliability to the unique characteristics of the data. These findings impact the diverse behaviors and performance characteristics of the models when applied to the prediction task. Out of all XGBoost [16] gives promising results that uses an ensemble learning approach to assert multiple weak learners (decision trees) into a stronger analyzer. This ensemble feature can play a big role in its predictive power and reduce bias or overfitting problem and it also had ability to find non-linear relationship between the variables. Fine-tuning the hyperparameters may have impact to the improved performance of the model.

Model	**MSE**	**MAE**
Linear Regression	364.7	14.73
Random Forest	277.91	12.38
XG Boost	211.12	10.17
SVR	339.58	13.51
ANN	334.9	17.85
Bagging Regressor	428.19	15.4

Table 2: Evaluation Metrics of Various Models.

4 Conclusion

In this study, we classified O. Sativa samples based on their population type, which played a crucial role in predicting BVs. The findings revealed a substantial improvement in prediction accuracy compared to using entire population samples and other alternative methods. Among the various ML models evaluated, Random Forest (RF) and XGBoost demonstrated the most promising results.

In summary, genomic data, including genetic markers and genomic selection methods, play a crucial role in predicting phenotypic traits in Oryza sativa. By harnessing these tools, breeders can accelerate the development of new rice varieties with desired traits, contributing to improved

agricultural practices in the face of global challenge. We observed that if missing values are more the accuracy may be decreased. Among Machine Learning and Deep learning Techniques – Machine learning approaches has done well compared to deep learning techniques.

References

1. Jun Yan 1,#, Dong Zou 2,3,4,#, Chen Li 5, Zhang Zhang 2,3,4, Shuhui Song 2,3,4, *Xiangfeng Wang 1,*. SR4R: An Integrative SNP Resource for Genomic Breeding and Population Research in Rice.
2. Vasantha, S. V., & Kiranmai, B. (2022). Machine Learning-Based Breeding Values Prediction System (ML- BVPS). In *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1* (pp. 259-266). Springer Singapore.
3. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in Oryza sativa L. Genetics 169(3):1631–1638.
4. Onogi, A., Ideta, O., Inoshita, Y., Ebana, K., Yoshioka, T., Yamasaki, M., & Iwata, H. (2015). Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). *Theoretical and applied genetics*, *128*, 41-53.
5. Meuwissen et al.,(2001).Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps:2000-2001
6. Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and applied genetics*, *128*, 145-158.
7. Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., & Shiu, S. H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, *9*(11), 3691-370
8.)Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., ... & Xu, S. (2021). Genomic selection: breakthrough technology in rice breeding. *The Crop Journal*, *9*(3), 669-677.
9. Sehrawat, S., Najafian, K., & Jin, L. (2023). Predicting phenotypes from novel genomic markers using deeplearning. *Bioinformatics Advances*, *3*(1), vbad028.
10. Windhausen et al., (2012). Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments.
11. http://ricevarmap.ncpgr.cn/
12. Bioinformatics algorithms Design and implementation in Python Miguel Rocha Pedro G Ferreira Academic Press ,Elsevier ,2018.
13. Paper, D., & Paper, D. (2020). Scikit-Learn Classifier Tuning from Simple Training Sets. *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*, 137-163.
14. Korstanje, J. (2021). The random forest. In *Advanced Forecasting with Python: With State-of-the-Art-Models Including LSTMs, Facebook's Prophet, and Amazon's DeepAR* (pp. 179-191). Berkeley, CA: Apress.
15. Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275-294.
16. Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, *9*(40), 651-662.
17. Stefan Van Dongen, T., & Winnepenninckx, B. (1996). Multiple UPGMA and neighbor-joining trees and the performance of