**GEBV Phenotype Trait Height prediction**

**using ML & DL Techniques**

**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. These results demonstrate that models can reasonably estimate rice plant height from genotypes. Implementing these techniques in breeding programs enhances agricultural productivity and sustainability.

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

1. **INTRODUCTION**

There are six types of rice varieties, consisting of five cultivated rice subpopulations and one wild rice subpopulation in our research: Indica Rice, Aus Rice, Aromatic Rice, Tropical japonica rice, Temperate japonica rice, Rufipogon (wild rice subpopulation) [[1]](#bookmark=id.n6u91byp87zj). Rice (Oryza sativa L.) is a crucial food crop that sustains more than half of the global population. The Wild rice subpopulation lies between the wild and cultivated varieties, acting as a genetic reservoir that can adapt to environmental changes and enhance crop sustainability [[2]](#bookmark=id.up2yehoce9r8). Garris et al. [[3]](#bookmark=id.bw9qyu4f9dj) proposed a classification of O. sativa into five distinct groups based on model-based structure analysis: Indica, Aus, Aromatic, Temperate japonica, and Tropical japonica. These cultivated rice subpopulations exhibit genetic and taxonomic differences.

Each subpopulation possesses unique desirable genotype and phenotype characteristics, 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]](#bookmark=id.sbbd39d0jamr).

whole-genome prediction or genomic prediction. In this technique, genetic values of untested genotypes (lines or individuals) are predicted on the basis of genome-wide DNA markers such as single nucleotide polymorphisms [[2]](#bookmark=id.up2yehoce9r8).

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]](#bookmark=id.wepbfkhopzrj).

Jun Yan [[1]](#bookmark=id.n6u91byp87zj)  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]](#bookmark=id.nkarulf14u0p). **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 key elements for a better allocation of resources in plant breeding, especially due to the high costs of phenotyping [[6]](#bookmark=id.nkarulf14u0p).

Artificial Neural Networks (ANNs) were implemented in python using TensorFlow. The input layer for the ANNs contained the genetic markers for an individual, the nodes in the hidden layers were all fully connected to all nodes in the previous and following layers (i.e., Multilayer Perceptron). A non-linear

activation function (selected during the grid search, see below) was applied to each node in the input and hidden layers, except the last hidden layer, which was connected with a linear function to the

output layer, the predicted trait value (y) [[7]](#bookmark=id.88lxexk82f7b).

Weights were then optimized using the Adam Optimizer To determine the optimal stopping time for training (i.e., number of epochs), 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 between 0.3% and 1% [[7]](#bookmark=id.88lxexk82f7b).

**Problem statement**

J.L .J and J.P [[6]](#bookmark=id.nkarulf14u0p) 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]](#bookmark=id.up2yehoce9r8)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]](#bookmark=id.sbbd39d0jamr). There is a need to improve predictive ability of breeding values of O. Sativa ensuring it is robust to new data.

Azodi [[7]](#bookmark=id.88lxexk82f7b) 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 Data**

1. **Data collection**

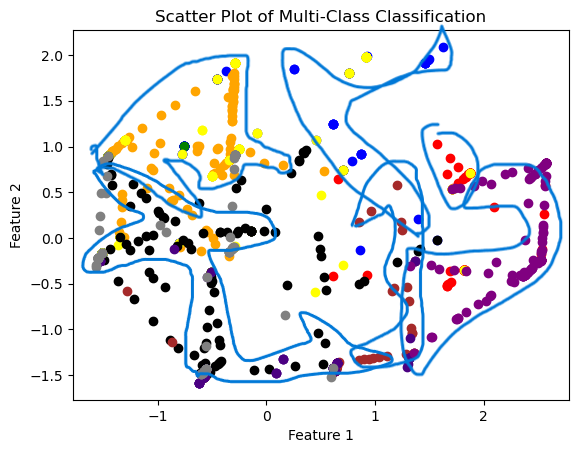
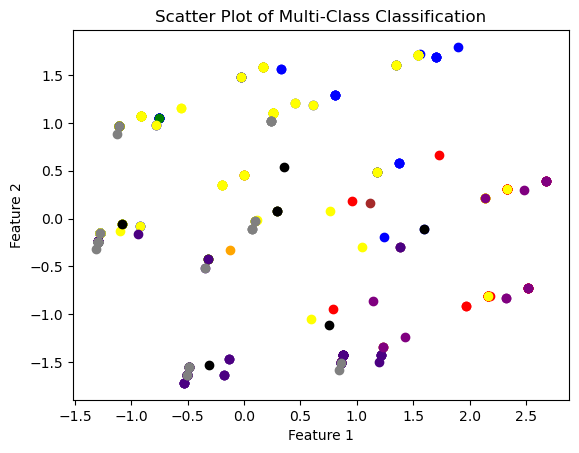
The data for rice cultivars, comprising phenotype, GWAS, and genotype information is collected from [11] <http://ricevarmap.ncpgr.cn/>. The dataset specifically includes phenotype values and GWAS information related to plant height. Imputed genotype information is also obtained from the database. We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

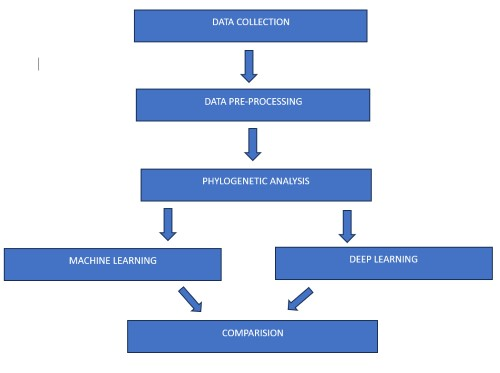
1. **Data cleaning and preprocessing**

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 performance. 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.

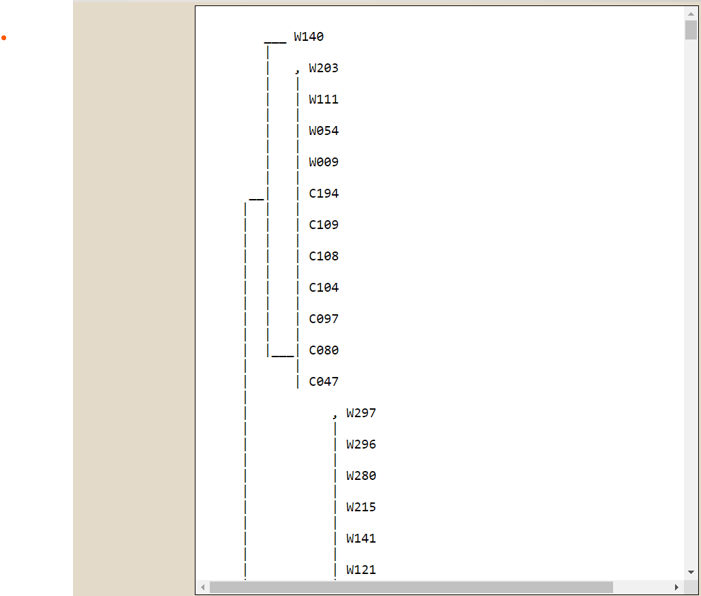
**Figure. 1** Before balancing classes **Figure. 2** After balancing classes



** Figure. 3** Flow of the project

1. **Phylogeny**

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), 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.



**Figure. 4**

**4 Models and Classification**

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, Logistic Regression (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](#fig1a). 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.

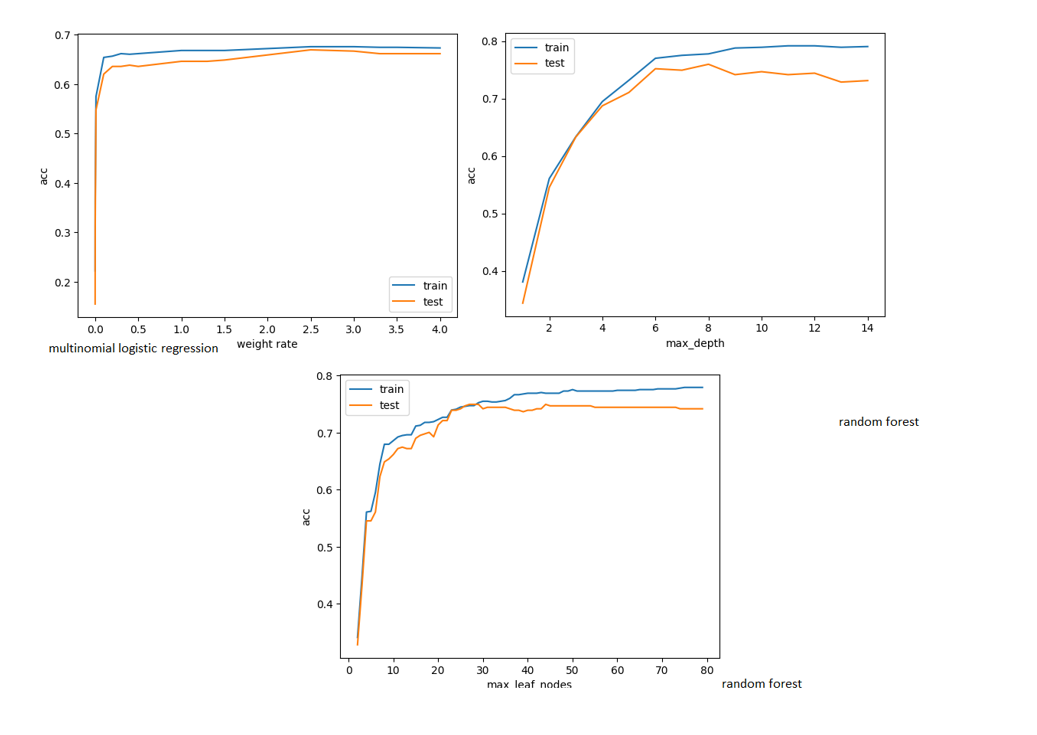
The accuracy achieved by the random forest model on the test data was 75%, which may not be considered high accuracy. However, it is important to note that the model is robust and neither overfits nor underfits the given data. High accuracy alone does not guarantee a better model if it overfits the data, as it may not perform well on unseen data. Therefore, achieving a balanced and robust model is more desirable in this context.

**Evaluating**

The models play a crucial role in selecting the best among the trained models. To accurately assess the performance of the models and prevent overfitting, we employed the repeated k-fold cross-validation technique. This technique involves dividing the dataset into k splits and iteratively using one split as the test data while training the model on the remaining splits. By repeatedly performing this process, we obtain a more reliable estimate of the model's accuracy that considers the entire dataset and captures any hidden relationships within it. This approach helps to ensure that the selected model is robust and performs well on unseen data. Performance of different models which we have used is being shown in [Table. 1](#table1).

**Table. 1** Results of various evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |



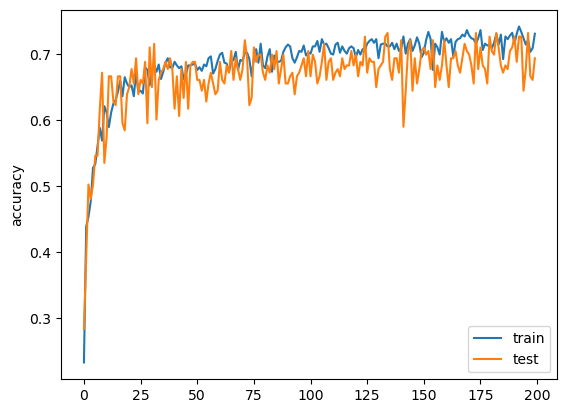
**Fig. 1a** Performance of multinomial logistic regression and random forest which are being plotted across the hyperparameters.

1. **Classification using ANN**

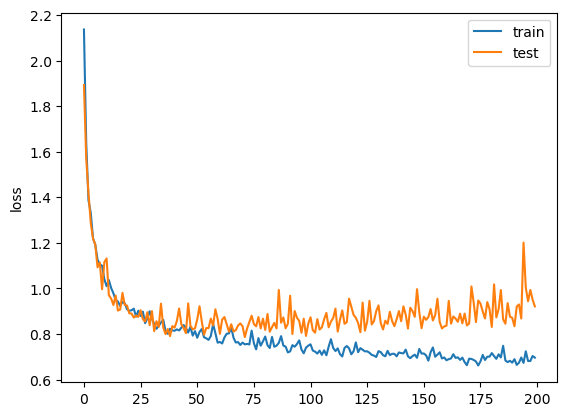
We employed deep learning techniques to capture the intricate relationships present in DNA sequences. However, it is worth noting that neural networks typically require a substantial amount of data to outperform traditional machine learning algorithms and produce superior results.

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 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. 2a** and **Fig. 2b.**



**Fig. 2a**



**Fig. 2b**

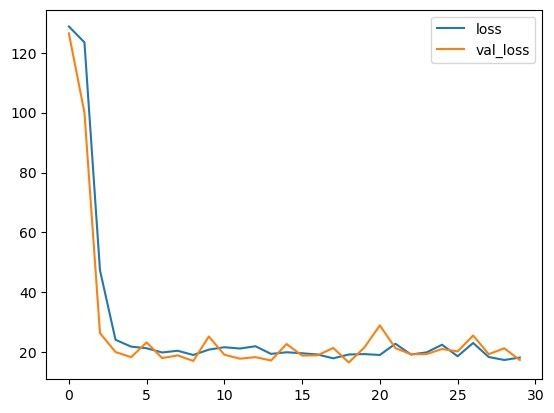
**5 Height Prediction**

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.



**Fig. 3a**

1. **Using Machine Learning**

After classifying the subpopulations, the height of each subpopulation was predicted. The concatenated values of DNA sequence, subpopulation, and height of subpopulation were used as the target variables.

In this study, a diverse range of predictive models were employed to estimate the height. Multiple models were used to comprehensively explore and analyse the prediction of height. In the current study, which involves the prediction of height as a continuous variable, several regression models were employed. These models include:

* Linear Regression: A commonly used regression model that assumes a linear relationship between the predictors and the target variable.
* Random Forest Regressor: A tree-based ensemble model that utilises multiple decision trees to make predictions.
* XGBoost Regressor: An implementation of gradient boosting that combines multiple weak regression models to create a powerful predictive model.
* Support Vector Machines (SVM): SVM is a versatile and widely used algorithm that can handle both linear and non-linear relationships between the predictors and the target variable.

By utilising these regression models, the authors aim to explore and evaluate their effectiveness in predicting height accurately.

**Training and Testing**

Data encoding was performed for the categorical variables of subpopulation and sequences. Count- vectorizer and label encoding were used as encoding techniques. Count vectorizer was applied to the sequences feature, which represents categorical data. This technique converts each sequence into a numerical representation based on the frequency of characters or n-grams (in this case, 3-grams) present in the sequence. The sequences were transformed into numerical features that could be used as inputs for the models.

Label encoding, on the other hand, was applied to the subpopulation feature. This technique assigns a unique numerical label to each distinct subpopulation category. The subpopulation labels were converted into numerical values, which made it possible for the models to process this categorical information effectively

Following the encoding step, linear regression was performed as the predictive modelling technique. To ensure optimal performance, hyperparameter tuning was conducted using GridSearchCV, a widely used method for systematically searching the hyperparameter space. GridSearchCV was used to explore various combinations of hyperparameters to identify the best parameters for the linear regression model. This approach allowed the model to be fine-tuned and optimised for predicting the target variable. Additionally, using the best parameters helped prevent the model from overfitting.

After training the optimised linear regression model with the best hyperparameters obtained through GridSearchCV, its performance was evaluated by calculating two commonly used regression metrics: mean squared error (MSE) and mean absolute error (MAE).To identify the most effective model for predicting height, a comparative analysis was conducted by applying multiple models. Specifically, linear regression, random forest regressor, XGBoost regressor, and support vector machine (SVM) were used.

**Evaluating**

A comparative analysis of multiple models was conducted to identify the most effective model for predicting height. The results of the analysis showed that the XGBoost regressor consistently produced the lowest error across multiple runs. This suggests that XGBoost may have converged to a certain prediction pattern and is less sensitive to the specific characteristics of the dataset.

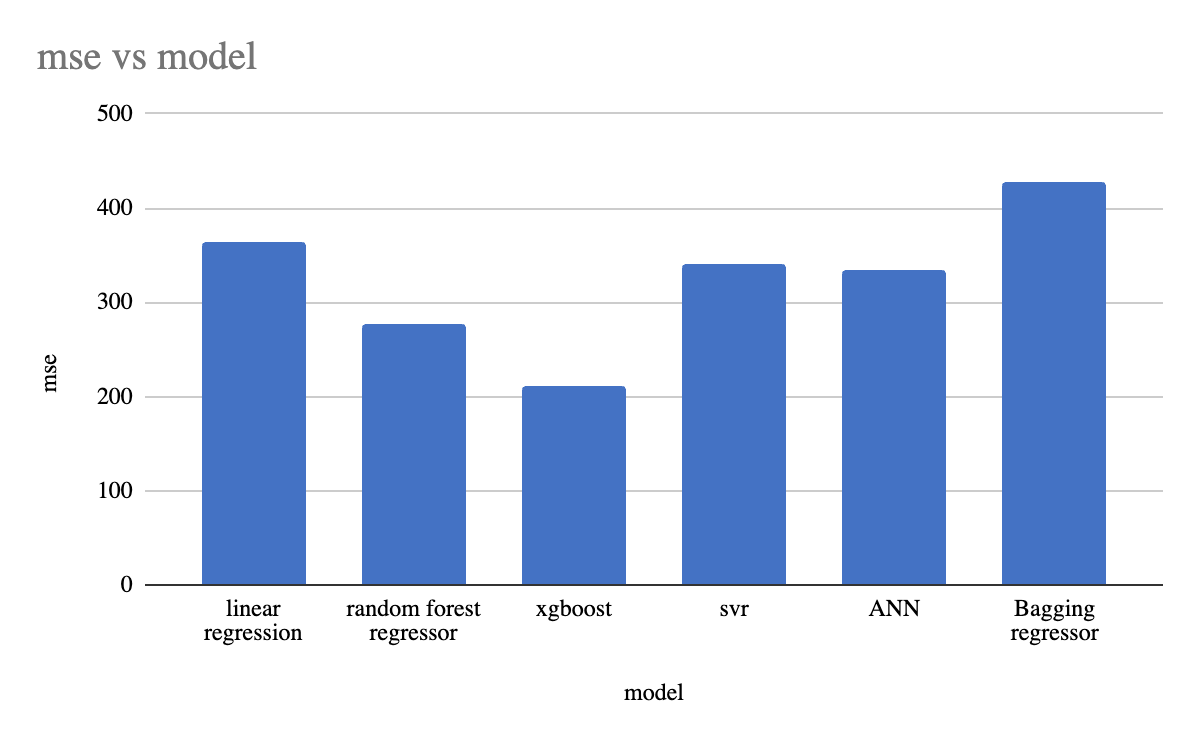
On the other hand, models such as RandomForestRegressor and LinearRegression exhibited varying errors, indicating their sensitivity to the unique characteristics of the data. These findings highlight the diverse behaviours and performance characteristics of the models when applied to the prediction task. Out of all XGBoost gives promising results that uses an ensemble learning approach to combine multiple weak learners (decision trees) into a stronger predictor. This ensemble nature can enhance its predictive power and reduce bias or overfitting and it also had ability to find non-linear relationship between the variables. Fine-tuning the hyperparameters may have contributed to the improved performance of the model.

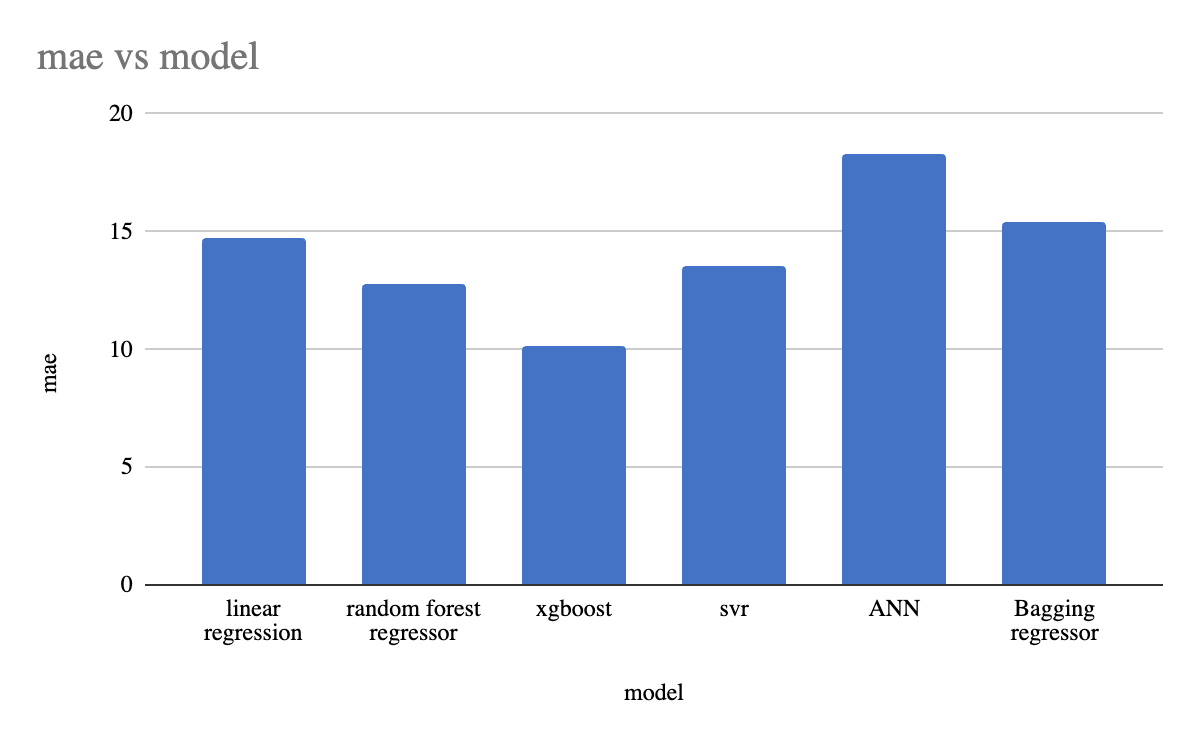
**Results of various regression models**

|  |  |  |
| --- | --- | --- |
| **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 regression models which include MSE (mean squared error) and MAE (mean absolute error) which were imported from the scikit-learn library.

1. **Comparative analysis of various regression models**

****

**Fig. 4a **

**Fig. 4b**

**6 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.

**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/