International Journal of Agriculture and Biosciences 2023 12(4): 277-283.
https://doi.org/10.47278/journal.ijab/2023.078

SHORT COMMUNICATION
eISSN: 2306-3599; pISSN: 2305-6622

Integrating Phylogenetic Analysis and Machine Learning for Enhanced Phenotype Prediction in Oryza Sativa

Dr. B. Kiranmai[1*] and Rishabh Gurbani[2]

[1]Associate Professor, Department of CSE, KMIT, Hyderabad, India
[2]Department of IT, KMIT, Hyderabad, India
*Corresponding author: kiranmaimtech@gmail.com

ABSTRACT

Genomic selection has revolutionized plant breeding by enabling the efficient and accurate selection of elite genotypes. Traditional approaches require resource-intensive phenotyping at all stages of artificial selection. However, genomic selection reduces this burden by leveraging genotyping data and machine learning techniques to predict agronomically relevant phenotypic traits. In this paper, we present a two-level prediction system that incorporates both phylogenetic analysis and machine learning models to predict the height of *Oryza Sativa* L. (Rice) plants based on their gene sequences. The only input for our model is a genomic sequence, whose length does not have to be equal to 24 (number of SNPs considered for our model). At the first level, we employ phylogenetic analysis to classify the plants into subpopulations, capturing the inherent genetic diversity within the dataset. This approach addresses the limitations of existing research, as it incorporates population structure information from gene sequences that is often overlooked in machine learning-based approaches. Subsequently, at the second level, we leverage the population structure information and genomic data to train a machine learning model for accurately predicting plant height. We compare and evaluate various methods employed at both levels to identify the most effective approach. Hence our approach of predicting Phenotype with reference to genotype is accurate compared with other existing systems. Two level classification has done well in identifying phenotype and performed well in predicting subpopulation.

Keywords: Genomic prediction, Phylogeny, Machine Learning, GWAS, Breeding value

Article History
Article # 23-381
Received: 11-Aug-2023
Revised: 10-Oct-2023
Accepted: 22-Oct-2023

INTRODUCTION

Predicting Phenotype

The ability to predict an organism's phenotype from its genotype and environment is a central problem in genetics with significant societal implications (Zafar et al., 2021; Zafar et al., 2022a; Zafar et al., 2022b). This problem extends to various domains, including medicine, where understanding the relationship between genotype, environment, and disease phenotype is crucial (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). In agriculture, predicting crop phenotypes such as yield and drought resistance is essential for sustaining the world's growing population (Zafar et al., 2022c; Desta & Ortiz 2014).

Advances in DNA sequencing technology have revolutionized our ability to characterize an organism's genotype by employing thousands of genetic markers (genome-wide association studies, GWAS) (Yang et al., 2010; Desta & Ortiz, 2014). However, the traditional focus of GWAS on identifying genetic markers associated with phenotypes has limitations in capturing the complex nature of phenotypic traits and evaluating their predictive utility (Lynch & Walsh, 1998). To overcome these challenges, a more direct and operational approach to phenotype prediction has emerged, aiming to learn predictive functions that leverage an organism's genotype and environment to predict its phenotype (Haroon et al., 2022a).

In the field of plant breeding, genetic markers, particularly single nucleotide polymorphisms (SNPs), have revolutionized the process by enabling marker-assisted selection (MAS) (Vignal et al., 2002; Ali et al., 2023). MAS relies on the association of genetic markers with target genes to rapidly select genotypes with desired phenotypic traits. However, MAS is limited to major-effect loci and does not account for the complexity of agronomically relevant traits controlled by minor-effect loci (Desta & Ortiz, 2014). To address this limitation, genomic selection (GS) has emerged as an advanced breeding approach that utilizes genome-wide genetic markers to capture all quantitative trait loci (QTL) for a trait (Jannink et al., 2010; Razzaq et al., 2022). GS overcomes the need for QTL mapping and enables the consideration of minor-effect QTL, increasing its prediction power compared to MAS (Desta & Ortiz, 2014). The rise of GS has transformed plant breeding by integrating phenotypic and genotypic information through statistical machine learning models (Jannink et al., 2010). This predictive methodology, leveraging reference information containing both phenotypic and genotypic data, has demonstrated its potential in accelerating genetic gain in crops such as maize, wheat, and chickpea (Desta & Ortiz, 2014; Haroon et al., 2022b; Haroon et al., 2023). The cost reduction in genotyping technologies, coupled with the

Cite this Article as: Kiranmai.B and Gurbani R, 2023. Integrating phylogenetic analysis and machine learning for enhanced phenotype prediction in oryza sativa. International Journal of Agriculture and Biosciences 12(4): 277-283. https://doi.org/10.47278/journal.ijab/2023.078

A Publication of Unique Scientific Publishers

proven utility of GS, has expanded its application beyond annual crops to long-lived species (Desta & Ortiz, 2014). Overall, the integration of genomic data, machine learning techniques, and predictive models in GS holds promise for advancing our understanding of genotype-phenotype relationships, enabling more efficient plant breeding strategies, and addressing the challenges of feeding a growing population.

Phenotype Data

In phenotype prediction, genomic data plays a crucial role, represented by discrete attributes called markers, often single nucleotide polymorphisms (SNPs) (Yang et al., 2010; Desta & Ortiz 2014). Genomic data exhibits a specific structure that influences the application of machine learning methods. The complete genotype of an organism consists of gene arrangements and sequences, with genotype information represented as markers. However, this propositional marker representation overlooks important biological information (Lynch & Walsh, 1998; Bloom et al., 2015).

The availability of fully sequenced genotypes is preferred but not always feasible due to technical or cost constraints. Phenotype prediction problems can have fully sequenced organisms, while others have only partial marker information (Desta and Ortiz 2014). The environment also plays a role and while controlled environments are desirable, they are not always possible in many scenarios (Yang et al., 2010; Bloom et al., 2015). The measurement of phenotype is often the most expensive step, and the number of phenotype attributes may exceed the number of examples (Desta & Ortiz, 2014).

The number of genetic mutations causing a phenotype can vary greatly, with some phenotypes being influenced by a single gene, while others involve multiple genes and environmental effects (Armstead, 2007; Wood et al., 2014). Genetic data exhibits linkage disequilibrium, where markers close together on an organism's DNA are likely to be inherited together (Desta & Ortiz, 2014). In summary, genomic data in phenotype prediction presents challenges and opportunities due to its specific structure, marker representation, availability of sequenced genotypes, controlled environments, measurement costs, causation of phenotype and linkage disequilibrium. Understanding and incorporating these factors are crucial in developing effective predictive models for phenotype prediction.

Present Statistical and Learning Approaches

The analysis of genotype, environment, and phenotype data in agri-genomics has traditionally relied on classical statistical genetics methods, such as univariate and bivariate statistical approaches (Lynch and Walsh 1998; Westfall et al. 2002; Marchini et al., 2007). These methods involve testing each marker or pairs of markers individually for association with a phenotype, ignoring complex causal relationships and potential interactions between markers. Multiple testing issues and limitations associated with p-values have been addressed through approaches like false discovery rate (FDR) and Bayes factors. To overcome the dimensionality problem, techniques like grouping markers into haplotypes have been explored, although identifying meaningful haplotypes remains a challenge (Meng, 2003; Lin & Altman, 2004). More recently, there has been a shift towards multivariate linear models, such as genomic BLUP, penalized regression methods, Bayesian techniques, and linear mixed models, which consider all markers simultaneously and account for population structure and genetic relatedness (Meuwissen et al., 2001; VanRaden, 2008; Gianola, 2006; Li & Sillanpää, 2012; Guan and Stephens 2011).

In contrast to classical statistical genetics methods, machine learning methods have gained attention in agri-genomics due to their ease of use, multivariate nature, and ability to handle attribute selection and capture complex interactions (Dudoit et al., 2002; Ziegler, 2007; Szymczak et al., 2009; Ogutu et al., 2011, Okser et al., 2014; Cherlin 2018). These methods, such as lasso, regression trees, random forest, gradient boosting machines, and neural networks, do not rely on assumptions about the underlying genetic mechanisms. Machine learning methods can handle the p >> n problem, where the number of attributes exceeds the number of samples, although strong underlying signals are required for their effectiveness. However, attribute selection remains a challenging task, particularly when distinguishing between markers associated with a trait and those that are causally linked to it. The goal of machine learning approaches in agri-genomics is to build predictive models and not necessarily mechanistic models, although the underlying biology should be considered (Schaid, 2018; Jaynes, 2003). (Vasantha & Kiranmai, 2022) has utilized machine learning techniques for prediction of height but could not make it for whole population.

Jeong et al. (2020a, b) developed GMStool, a GWAS-based marker selection tool for genomic prediction from genomic data. This tool aimed to improve the efficiency and accuracy of marker selection compared to existing methods. By fitting a statistical model assuming small and similar effect sizes of markers, GMStool successfully identified markers with the largest estimated effects for genomic prediction. In the study by Liu et al. (2019), a deep convolutional neural network (CNN) was utilized for phenotype prediction and genome-wide association study in soybean. The CNN was trained on a dataset of soybean genotypes and phenotypes, enabling accurate phenotype prediction and identification of genetic variants associated with important traits. Bartholomé et al. (2022) provided an overview of the progress and perspectives in genomic prediction for rice improvement. This comprehensive review highlighted the advancements in genomic prediction methods, such as genomic selection, association mapping, and machine learning, and their potential applications in enhancing rice breeding programs.

MATERIALS & METHODS

Proposed Model

Most conventional machine learning models often overlook the population structure information present in the genetic data, leading to suboptimal predictions. In this study, we propose a two-level prediction system shown in (Fig.1) that integrates phylogeny and machine learning to accurately predict the height of Oryza Sativa plants based on their gene sequence data.

1. To address this limitation, our approach leverages phylogenetic analysis as the first level to classify plants into their respective subpopulations with the help of phylogenetic tree.
2. This subpopulation information along with gene sequence, is then utilized in the second level to train a machine learning model specifically tailored for predicting plant height.

The only input for our model is a genomic sequence, whose length does not have to equal 24 (number of SNPs considered for our model).

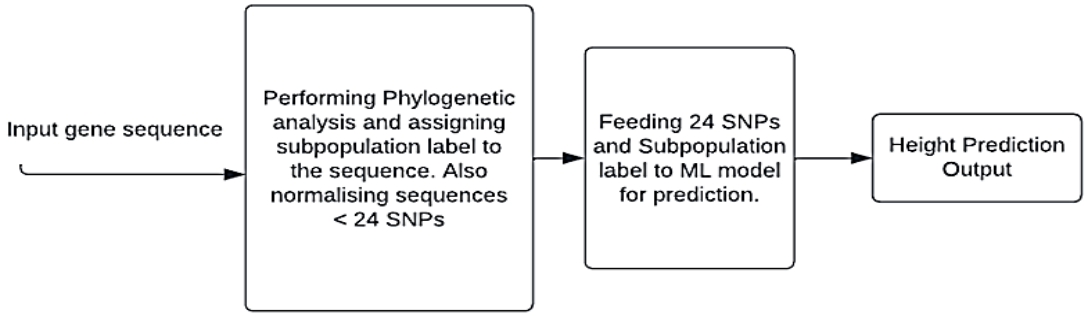

Fig. 1: Proposed model flowchart.

Dataset

Rice (Oryza sativa) is the first crop to have its genome sequenced. In the past decade, thousands of rice accessions in germplasm banks worldwide have been genotyped, and numerous rice variation databases have been constructed (Song et al., 2018). The dataset we worked with, RiceVarMap, was based on the SR4R database. SR4R (Selective Rice Variation Resource) is a couch (Yan et al., 2020) comprehensive resource that emerged from the rice variation database (RVD), a daughter database of the Information Commons for Rice consortium (IC4R) (http://ic4r.org/). To ensure the usefulness of the rawSNP data, it underwent processing to remove low-quality SNPs, including those with missing or low-frequency genotypes and redundant SNPs identified due to linkage disequilibrium (LD).

The phylogenetic tree exhibited six major clades, representing five cultivated rice subpopulations and one wild rice subpopulation. The cultivated rice subpopulations include indica rice (Ind), Aus rice (Aus), Aromatic (Aro) rice, tropical japonica rice (TrJ), and temperate japonica rice (TeJ) depicted in Fig. 2.

We conducted our analysis using the rice dataset available at (http://ricevarmap.ncpgr.cn/). This dataset comprises genotype and phenotype data for various rice accessions. In our study, we focused on the imputed dataset, which underwent a process to estimate missing genotypes.

The dataset provided us with information on single nucleotide polymorphisms (SNPs) in rice, and we specifically worked with the top 24 SNP variation IDs. These SNP IDs were selected based on their Pearson correlation coefficients, indicating their relevance to the traits under investigation shown in Table 1. The selected SNP IDs were as follows: vg0112116426, vg0128525986, vg0130976864, vg0131664768, vg0133440209, vg0135617816, vg0135642980, vg0138418739, vg0138428840, vg0138608956, vg0138999212, vg0405463422, vg0405463763, vg0603483061, vg0713178880, vg0719727299, vg0719727339, vg0719834473, vg0819793460, vg0904094998, vg0904282939, vg1019044175, vg1123563633, and vg1207667840.

The dataset was provided in three separate files: Cultivar Information, genomic sequences for the SNP variation IDs, and phenotype information. To conduct our analysis, we combined these files shown in (Table 2) into a unified dataset while addressing missing values and resolving overlaps. The genomic sequences file contained specific codes to represent missing data, such as "DEL" indicating a missing deletion mutation and "N" indicating missing information. We replaced these codes with the primary and secondary alleles for the respective SNP, as per the standard representation. After processing the data, we obtained a Genome-Wide Association Study (GWAS) (https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies) file that contains comprehensive information (24 SNPs, height, subpopulation) about 529 rice cultivars. This file includes details about their gene sequences and corresponding height measurements.

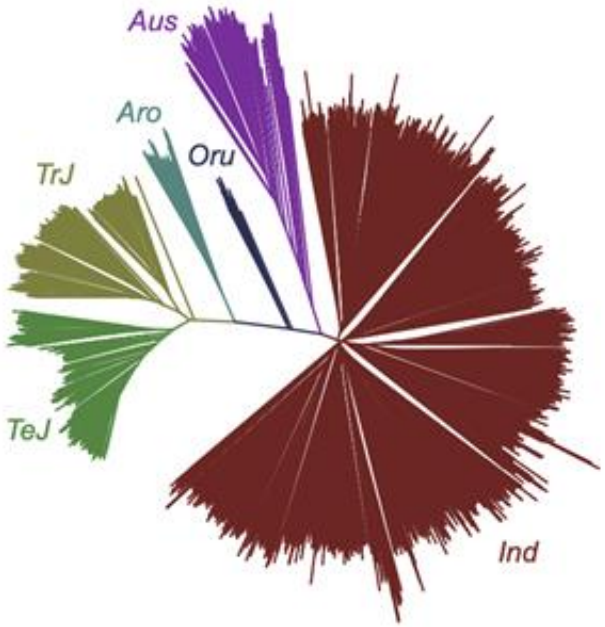

Fig. 2: Distribution of subpopulations of Oryza Sativa derived from phylogeny on genetic sequences of five major subpopulations: Aus, Aro - Aromatic, TrJ - Tropical Japonica, TeJ - Temporate Japonica, Ind - Indica.

Phylogeny

Phylogeny is the field of study that focuses on reconstructing the evolutionary relationships and genetic relatedness among organisms (Jarvis, Holland, & Sumner, 2017). It provides insights into the common ancestry and divergence of species, aiding our understanding of the evolutionary history and genetic factors that shape their characteristics. In our research, phylogeny plays a crucial role in predicting plant height in Oryza Sativa.

There are several common methods used in phylogenetic analysis. One widely used approach involves constructing phylogenetic trees, which visually represent the evolutionary relationships between organisms (Baum, 2008). To build these trees, sequence alignment is performed. Sequence alignment is the process of arranging DNA or protein sequences to identify regions of similarity. In our study, we utilized the muscle tool to align the gene sequences and calculate the alignment score,

which reflects the degree of similarity between sequences. The aligned sequences were then used to construct a phylogenetic tree.

Table 1: showing SNP variation IDs taken into consideration along with their correlation to the trait (Height). LR P and LMM P values show their correlation with height.

Variation ID	Chromosome	Position	LR *P*-value	LMM *P*-value
vg0112116426	1	12116426	NA	6.25E-07
vg0128525986	1	28525986	5.01E-42	NA
vg0130976864	1	30976864	NA	7.14E-07
vg0131664768	1	31664768	2.28E-44	NA

Table 2: Final Dataset with Information about Cultivars. Includes Cultivar ID, Subpopulation, Sequence (24 Features) and phenotype height. Each SNP is considered a feature for the model.

Cultivar ID	Subpopulation	Plant Height (cm)	Sequence
C001	Indica	144.13	GGGAGCCATCGTAATGTTTCCCCC
C002	VI/Aromatic	177.62	GGGGGCCACCGTAATGCCTCCCCC
C003	Japonica	141.57	GAGAGCCATAGGAGTGCCAACCTT
C004	Japonica	140.4	GAGGGCCATAGGAGAACTACTCTC
C005	Japonica	163.33	AGGGGCCATAGTAAAGCCAACTTC
C006	Indica	108.23	GGGAGCGGTATTAATGTTACCTTT

Various methods can be employed to construct phylogenetic trees, and we explored different combinations of alignment tools and tree construction algorithms. We tested tools like Cluster and Muscle, and algorithms like UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and Maximum Likelihood (Rocha & Ferreira, 2018). After evaluating the results, we chose to use the Muscle tool coupled with the UPGMA algorithm. This combination resulted in a balanced and evenly distributed phylogenetic tree, which accurately represented the genetic relationships among Oryza Sativa cultivars.

The constructed phylogenetic tree depicted in (Fig.3) served as the basis for forming clusters of cultivars based on their similarity scores. To achieve this, a custom script was developed. In total, we obtained 317 clusters, each representing a distinct subpopulation with cultivars exhibiting similar genetic characteristics. This clustering approach allows us to group cultivars into subpopulations, providing valuable insights into the genetic diversity and population structure of Oryza Sativa.

The main idea behind our approach is to utilize the phylogenetic tree and clusters to predict the phenotype of unidentified sequences. When we encounter an unidentified sequence, we calculate its similarity score against the sequences in our database. The sequence belongs to the cluster with the highest similarity score. The subpopulation assigned to that cluster is determined by the majority of subpopulations present in that particular cluster.

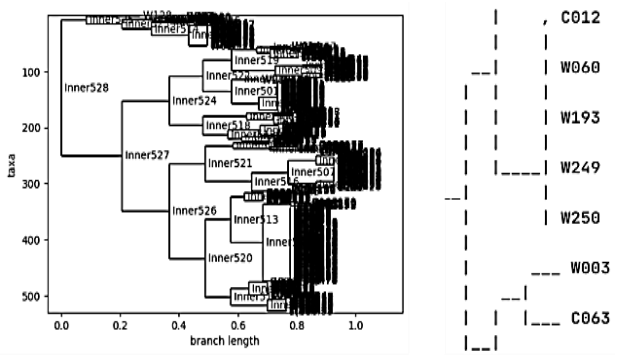

Fig 3: Phylogenetic tree and clusters formed.

Learning Model

In order to improve the training of our model due to the limited size of our dataset, we employed a synthetic data generation technique. By randomly deleting segments of the sequence and replacing them with "N" (representing missing data), and perturbing the height values by 1-2%, we augmented our original dataset of 529 records to a larger size of close to 5000 records.

Before feeding the data into our machine learning models, we performed preprocessing steps to encode the genetic information. Each position in the DNA sequence can take on one of five possible values: A, T, C, G, or N. To represent this information, we utilized one-hot encoding, resulting in a total of 96 features for the 24 SNPs (single nucleotide polymorphisms). If the value at a particular position was "N," all four corresponding columns were set to zero. Additionally, we applied label encoding to represent the subpopulation information based on the five major subpopulations: indica, japonica, aromatic, aus and intermediate. This resulted in a total of 97 features, including the encoded subpopulation, along with the target output feature, height.

To explore the performance of various machine learning models on our augmented dataset, we trained models such as SVG Regressor (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html), XGBoost Regressor (https://machinelearningmastery.com/xgboost-for-regression/), GBLUP (Clark & van der Werf, 2013), LASSO and Random Forest. Additionally, we employed an Artificial Neural Network (ANN) model to evaluate its effectiveness in predicting plant height.

Furthermore, we trained the XGBoost Regressor on the top 30 features selected by its feature extraction algorithm, aiming to assess the impact of feature selection on model performance.

Prediction

When predicting the height of a plant using our model, we follow a two-step process depicted in Fig.4.

First, given an input DNA sequence, we utilize phylogenetic analysis to determine the most closely matching cultivar in our database and identify the corresponding cluster. This approach allows us to leverage the similarity scores obtained through sequence alignment. Additionally, if the length of the input sequence is less than the number of SNPs (24) our machine learning model is trained on, we utilize the aligned sequences and replace any missing data ("N") or deletions ("_") with the primary and secondary alleles at that particular position. By utilizing the phylogenetic tree, we also obtain the subpopulation information associated with the given sequence.

In the second step, we feed both the subpopulation information and the normalized DNA sequence into our trained machine learning model. The model then processes this combined input to predict the height of the plant.

RESULTS and DISCUSSION

High prediction accuracy is a prerequisite for the successful application of genomic selection. Prediction accuracy is often measured by the correlation between observed phenotypes and the predicted GEBVs (http://nsip.org/wp-content/uploads/2021/02/Overview-GEBV-Article.pdf) or predicted phenotypes of cross-validation (Xu, 2017). To evaluate the performance of our model, we employed a 5-fold cross-validation approach. We then assessed the model's performance using several evaluation metrics. For regression-based evaluation, we calculated metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and the R2 score (Karunasingha, 2022). These metrics provide insights into the accuracy and precision of our height predictions. Lower values for MAE, MSE and RMSE indicate better performance, while an R2 score closer to 1 indicates a higher degree of correlation between the predicted and actual heights represented in Table. 3.

Furthermore, we examined the Pearson correlation coefficient to measure the linear relationship between the predicted and actual heights. A higher correlation coefficient value suggests a stronger linear association between the predicted and actual values.

The results of the experiments are given in the figure 6.

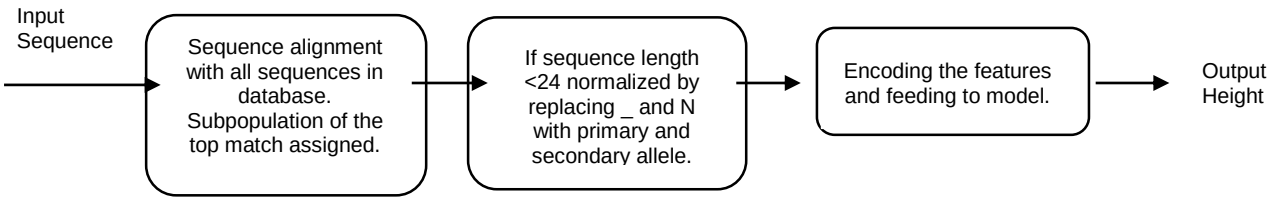

Fig. 4: Prediction process flowchart.

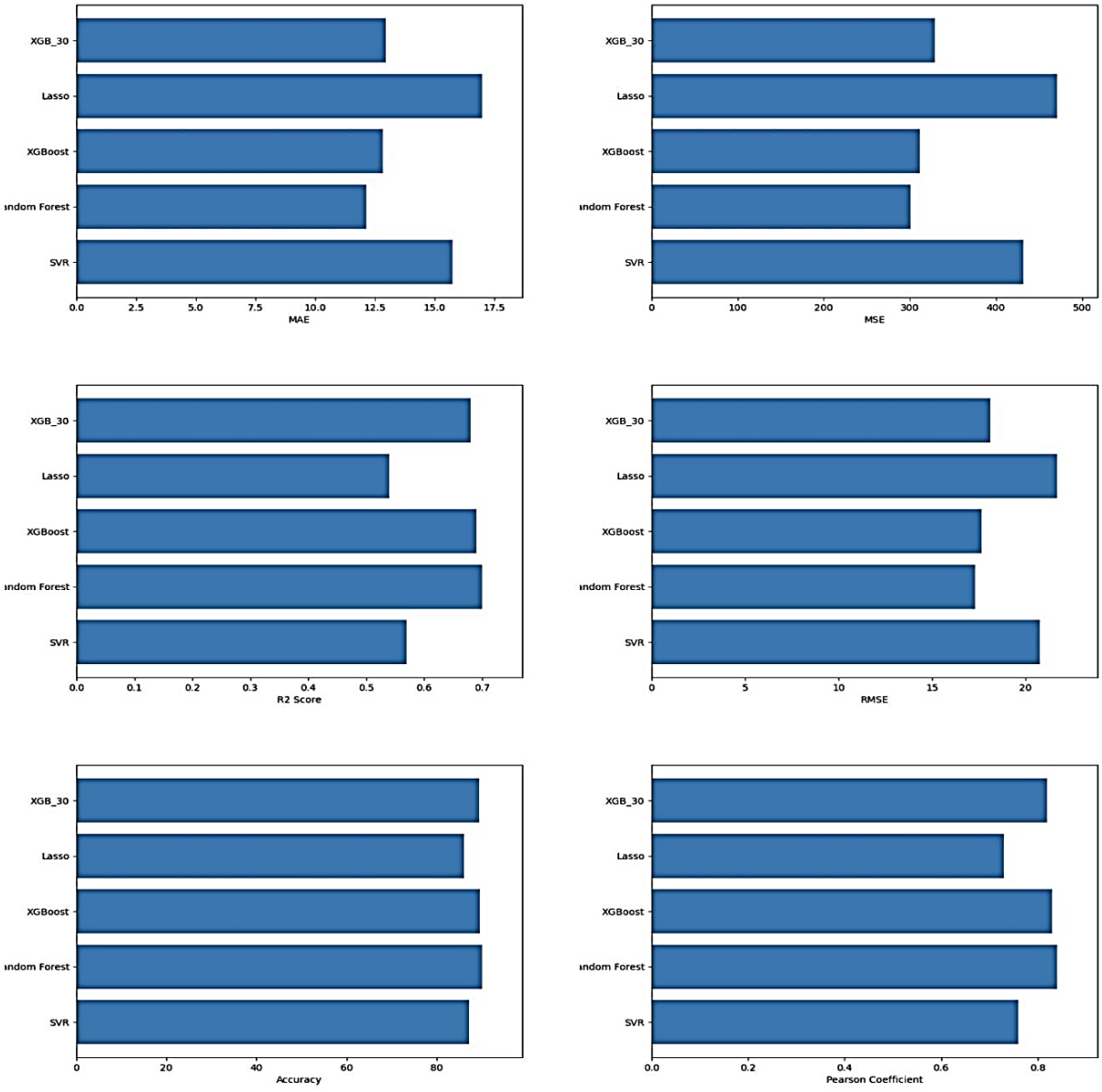

Fig. 6: Graphs on Evaluation metrics.

Table 3: Comparison of different model evaluations for prediction of height. XBG_30 - XGBoost regressor trained on the top 30 features

	MAE	MSE	RMSE	R2 Score	Accuracy	Pearson coefficient
Random Forest	12.13	300.54	17.33	0.70	90.02	0.84
XGBoost	12.83	311.69	17.65	0.69	89.53	0.83
XGB_30	12.94	328.77	18.13	0.68	89.42	0.82
SVR	15.74	432.23	20.78	0.57	87.14	0.76
Lasso	16.974	470.77	21.69	0.54	86.06	0.73

Table 4: Comparison of accuracy with existing studies

Authors	Accuracy percentage
Joen et al., 2020	0.52
Vasantha & Kiranmai, 2022	0.7 to 0.92
Grenier et al., 2015	0.54
Cui et al., 2020	0.50
Isdro et al., 2015	0.7
Yan et al., 2020	0.2 to 0.9
Wang et al., 2017	0.88
Proposed work	0.92

COMPARISON WITH EXISTING STUDIES

We compared our proposed work with some of the above-mentioned literature. Even though our procedure of predicting is hybrid involving biological and machine learning strategies, we compared the accuracy of predicting a phenotype with other works.

Here we adopted a hybrid method for predicting phenotypic traits that has better accuracy when compared with other models stated in the literature and depicted in Table 4. Most applications do not use phylogenetic trees and instead operate on pairwise sequence distances. In the proposed model, homologous allele sequences across different species or even within the same genome are clustered and classified using Phylogenetic trees.

Advantages of proposed method over existing studies

- Biological clustering (based on Genotype) is consistent and accurate when compared to other alternative approaches.
- Here we clustered and classified biological sequences, applied various Machine Learning Algorithms for the second level of classification and predicted phenotypic trait height.
- Most of the existing literature on prediction is based on genetic methods or machine learning techniques.

In our work, Random Forest has exhibited better predictions when compared with other Machine learning techniques. XGBoost has the second highest prediction rate after SVM and Lasso Regression.

REFERENCES

Ali, A., Zafar, M. M., Farooq, Z., Ahmed, S. R., Ijaz, A., Anwar, Z., ... & Maozhi, R. (2023). Breakthrough in CRISPR/Cas system: Current and future directions and challenges. *Biotechnology Journal*, 2200642.

Armstead, I. (2007). Identification and characterization of a key regulatory gene in Lolium perenne. Plant Biotechnol Journal, 5(3), 282-293.

Bartholomé, J., Prakash, P. T. and Cobb, J. N. (2022). Genomic Prediction: Progress and Perspectives for Rice Rice Improvement. Genomic Prediction of Complex Traits: Methods and Protocols, 569-617.

Baum, D. (2008). Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education*, *1*(1), 190.

Bloom, J. S., Kotenko, I., Sadhu, M. J., Treusch, S., Albert, F. W., & Kruglyak, L. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature communications*, *6*(1), 8712.

Cherlin, S. (2018). Applying machine learning methods to the analysis of genomic data. Comput Struct Biotechnol J, 16, 391-399.

Clark, S. A., & van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Genome-wide association studies and genomic prediction*, 321-330.

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., ... & Xu, S. (2020). Hybrid breeding of rice via genomic selection. *Plant biotechnology journal*, *18*(1), 57-67.

Desta, Z. A. and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. Trends Plant Science, 19(9), 592-601.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 111-139.

Gianola, D. (2006). A unified view of genomic prediction. Genetics, 182(2), 753-755.

Grenier, C., Cao, T. V., Ospina, Y., Quintero, C., Châtel, M. H., Tohme, J., ... & Ahmadi, N. (2015). Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PloS one*, *10*(8), e0136594.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat, 5(3), 1780-1815.

Haroon, M., Afzal, R., Zafar, M. M., Zhang, H., & Li, L. (2022). Ribonomics approaches to identify RBPome in plants and other eukaryotes: current progress and future prospects. *International Journal of Molecular Sciences*, *23*(11), 5923.

Haroon, M., Tariq, H., Afzal, R., Anas, M., Nasar, S., Kainat, N., ... & Zafar, M. M. (2023). Progress in genome-wide identification of RBPs and their role in mitigating stresses, and growth in plants. *South African Journal of Botany*, *160*, 132-146.

Haroon, M., Wang, X., Afzal, R., Zafar, M. M., Idrees, F., Batool, M., ... & Imran, M. (2022). Novel plant breeding techniques shake hands with cereals to increase production. *Plants*, *11*(8), 1052.

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.

Jannink, J. L., Lorenz, A. J. and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics, 9(2), 166-177.

Jarvis, P., Holland, B., & Sumner, J. (2017). Phylogenetic invariants and Markov invariants.

Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

Jeong, S., Kim, J. Y. and Kim, N. (2020a). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Scientific Reports, 10(1), 1-12.

Jeong, S., Kim, J. Y., & Kim, N. (2020b). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Scientific reports*, *10*(1), 19653.

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well.

Information Sciences, *585*, 609-629.

Li, M. and Sillanpää, M. J. (2012). Bayesian marker selection in high-dimensional generalized linear models. J Am Stat Assoc, 107(498), 565-576.

Lin, S. and Altman, R. B. (2004). Finding haplotype tagging SNPs by use of principal components analysis. Am Journal Hum Genetic, 75(5), 850-861.

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T. and Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Frontiers in Genetics, 10, 1091.

Lynch, M. and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, *39*(7), 906-913.

Meng, Z. B. (2003). Haplotype-based linkage disequilibrium mapping via direct haplotype sequencing. Ann Hum Genetic, 67(Pt 3), 261-273.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819-1829.

Ogutu, J. O., Piepho, H. P. and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC proceedings (Vol. 5, No. 3, pp. 1-5). BioMed Central.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S. and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. PLoS Genetics, 10(11), e1004754.

Razzaq, A., Zafar, M. M., Ali, A., Hafeez, A., Sharif, F., Guan, X., ... & Yuan, Y. (2022). The pivotal role of major chromosomes of sub-genomes A and D in fiber quality traits of cotton. *Frontiers in Genetics*, *12*, 642595.

Rocha, M., & Ferreira, P. G. (2018). *Bioinformatics algorithms: design and Implementation in Python*. Academic Press.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421-427.

Schaid, D. J. (2018). Machine learning to integrate human epigenomic annotations in GWAS. Nat Genet, 50(2), 220-231.

Song, S., Tian, D., Zhang, Z., Hu, S. and Yu, J. (2018). Rice genomics: over the past two decades and into the future. Genomics, Proteomics & Bioinformatics, 16(6), 397-404.

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H. and Sun, Y. V. (2009). Machine learning in genome-wide association studies. Genetic Epidemiology, 33(S1), S51-S57.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci, 91(11), 4414-4423.

Vasantha, S. V. and 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.

Vignal, A., Milan, D., SanCristobal, M. and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution, 34(3), 275-305.

Wang, X., Li, L., Yang, Z., Zheng, X,, Yu, S., Xu, C. and Hu, Z. () Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. aHeredity. 2017 Mar;118(3):302-10.

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.

Westfall, P. H., Zaykin, D. V. and Young, S. S. (2002). Multiple tests for genetic effects in association studies. Biostatistical Methods, 143-168.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S. and Kratzer, W. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics, 46(11), 1173-1186.

Xu, S. (2017). Predicted residual error sum of squares of mixed models: an application for genomic prediction. *G3: Genes, Genomes, Genetics*, *7*(3), 895-909.

Yan, J., Zou, D., Li, C., Zhang, Z., Song, S. and Wang, X. (2020). SR4R: An integrative SNP resource for genomic breeding and population research in rice. Genomics, Proteomics Bioinformatics, 18(2), 173-185.

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 & Bioinformatics*, *18*(2), 173-185.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R. and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565-569.

Zafar, M. M., Jia, X., Shakeel, A., Sarfraz, Z., Manan, A., Imran, A., ... & Ren, M. (2022a). Unraveling heat tolerance in upland cotton (Gossypium hirsutum L.) using univariate and multivariate analysis. *Frontiers in plant science*, *12*, 727835.

Zafar, M. M., Manan, A., Razzaq, A., Zulfqar, M., Saeed, A., Kashif, M., ... & Ren, M. (2021). Exploiting agronomic and biochemical traits to develop heat resilient cotton cultivars under climate change scenarios. *Agronomy*, *11*(9), 1885.

Zafar, M. M., Rehman, A., Razzaq, A., Parvaiz, A., Mustafa, G., Sharif, F., ... & Ren, M. (2022b). Genome-wide characterization and expression analysis of Erf gene family in cotton. *BMC plant biology*, *22*(1), 134.

Zafar, M. M., Shakeel, A., Haroon, M., Manan, A., Sahar, A., Shoukat, A., ... & Ren, M. (2022c). Effects of salinity stress on some growth, physiological, and biochemical parameters in cotton (Gossypium hirsutum L.) germplasm. *Journal of Natural Fibers*, *19*(14), 8854-8886.

Ziegler, A. (2007). Methods for meta-analysis of genetic data. Eur J Hum Genetic, 15(7), 740-746.