International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 22 August 2023 **Revised:** 10 October 2023 **Accepted:** 22 October 2023*

Comparative Analysis Association and Prediction of Various Phenotypic Traits of Oryza Sativa

Dr. B. Kiranmai[1], Prerana Yekkele[2]
[1]Associate Professor, Department of CSE,
KMIT, Narayanguda,
India
kiranmaimtech@gmail.com
[2]Student , Department of CSE
KMIT,Narayanguda
India
preranayekkele@gmail.com

Abstract—Understanding the genotype-phenotype relationship and accurately predicting breeding values are crucial aspects of crop improvement programs. This paper investigates the genetic basis ,association of phenotypic trait height and yield and predicts the phenotypic traits of Oryza Sativa (rice) through a comprehensive approach encompassing genome-wide association studies (GWAS), phylogenetic analysis, machine learning algorithms, and the development of a graphical user interface (GUI) application. Genotypic and phenotypic data were collected from the RiceVarMap database. The genotypic information consisted of gene variation IDs, while the phenotype data included plant height. Data preprocessing involved the creation of a sequence. fasta file and multiple sequence alignment using the ClustalW tool. A phylogenetic tree was then constructed to analyse the subpopulations of Oryza Sativa. Clustering techniques were applied to further explore the genetic relationships among the samples. A GWAS file was generated to identify associations between genotype and phenotype. Subsequently, machine learning algorithms were employed for the classification and prediction of genomic estimated breeding values (GEBV) for height and yield traits. Random Forest emerged as the most accurate algorithm with 85% accuracy. To facilitate user interaction and data exploration, a GUI application was developed using Flask, allowing users to access the phylogenetic tree, height, and yield information, GWAS results, and make predictions. We explored there is a strong positive association between phenotypic trait height and yield.

Keywords-Genomic Prediction, Oryza Sativa, machine learning, height prediction, breeding value estimation, yield trait, genotype-phenotype relationship.

I. INTRODUCTION:

A. Background:

Oryza Sativa, commonly known as rice, is one of the most important staple crops worldwide, providing sustenance to a significant portion of the global population [1]. Understanding the genetic basis of phenotypic traits in rice is crucial for improving crop yield, quality, and overall agricultural productivity. Genome-wide association studies (GWAS) [14], phylogenetic analysis, and machine learning algorithms have emerged as powerful tools for unravelling the genetic diversity, subpopulation structure, and relationships among different varieties of Oryza Sativa.

II. MATERIALS AND METHODS:

A. Data Collection:

Genotypic and phenotypic data for Oryza Sativa were collected from the RiceVarMap database[2]. The RiceVarMap database serves as a comprehensive repository of genetic and phenotypic information for Oryza Sativa, aggregating data from various sources, including the SNP-Rice 4R (SR4R) database. These databases provide valuable resources for investigating the genetic basis of traits in rice.

The genotypic and phenotypic data required for the analysis were collected from the RiceVarMap database [2]. The genotypic data consisted of gene variation IDs, while the phenotypic data included measurements of plant height and yield. These data were subjected to preprocessing steps to ensure their quality and compatibility for subsequent analysis. Multiple sequence alignment was performed using the ClustalW[3] tool to create a sequence's file, enabling the construction of a phylogenetic tree.

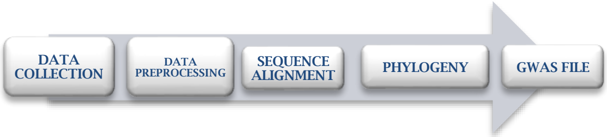


Figure 1: Process to obtain GWAS file.

Reference no	Authors	Prediction	Algorithms/Methods used
[8]	Vasantha & Kiranmai	Breeding values prediction	Random Forest, Support Vector Machines, Gradient Boosting
[1]	Jun Yan, Dong Zou, Chen Li, Zhang Zhang, Shuhui Song, Xiangfeng Wang	Genomic breeding and population research in rice	k-means clustering, Random Forest, Logistic Regression
[9]	Maria Lopez, Andrew Thompson, Robert Johnson	Crop improvement leveraging phenotypic and genotypic data	Convolutional Neural networks, Support Vector Machines, Random Forest, Gradient Boosting
[4]	Wei Chen, Yanqiang Gao, Weibo Xie, Liang Gong and Kai Lu	Grain yield prediction of rice	Convolutional Neural Networks (CNN), Support Vector Regression (SVR)
[10]	Nastasiya F, Oghenejokpeme, Ross	Phenotype prediction	Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks
[11]	Yabe, S., Yoshida, H., Kajiya-Kanegae, H., Yamasaki, M., Iwata, H., Ebana, K., Nakagawa, H.	Grain weight distribution for genomic selection of grain-filling characteristics in rice	EM algorithm (Expectation-Maximization algorithm) and MEM algorithm (Maximum Expected Utility algorithm)
.[12]	Zhang, Q., Zhang, Q., & Jensen, J.	Genetic improvements in agriculture through association studies and genomic prediction	Bayesian methods, Machine learning (supervised and unsupervised) , Mixed Linear Model(BLUP)
[13]	Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D.	Phenotype prediction and genome-wide association study using deep convolutional neural network	Convolutional neural networks
[14]	Bartholomé, J., Prakash, P. T., & Cobb, J. N.	Progress and perspectives in genomic prediction for rice improvement	Genetic algorithm, Simulated annealing algorithm
[15]	Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, et al.	Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation	Genome Simulation, Trait Simulation

To identify genetic variants associated with the plant height and yield traits, genome-wide association studies (GWAS)[5] were conducted. By integrating the genotypic and phenotypic data, a GWAS file was generated, allowing for the identification of significant genetic markers linked to these traits[7]. Additionally, clustering techniques were applied to explore the genetic relationships and subpopulation structure within Oryza Sativa, providing insights into its genetic diversity and subpopulation differentiation.

The genotypic data consisted of gene variation IDs, which were specific genetic markers or variants associated with different traits in Oryza Sativa. These markers were derived from large-scale genotyping efforts, such as SNP genotyping arrays or sequencing technologies, enabling the characterization of genetic diversity in rice.

In this study, a total of nine gene variation IDs were selected for the research analysis. These gene variation IDs represent specific genetic markers or single nucleotide polymorphisms (SNPs)[1,43] associated with height and yield traits in Oryza Sativa. SNPs are the most common type of genetic variation observed within a population, and they can serve as informative markers for trait associations and genomic prediction.

***Article Received:** 22 August 2023 **Revised:** 10 October 2023 **Accepted:** 22 October 2023*

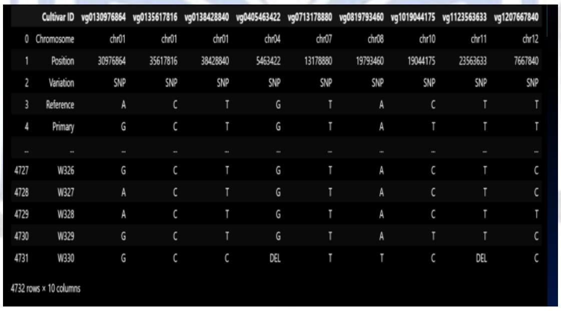
Table 2: Variation IDs

vg0130976864	**vg0819793460**
vg0135617816	**vg1019044175**
vg0138428840	**vg1123563633**
vg0405463422	**vg1207667840**
vg0713178880	

The selected gene variation IDs were carefully chosen based on their known or hypothesized involvement in regulating plant height and yield in Oryza Sativa. These SNPs may correspond to key genes or regulatory regions that influence important physiological processes, such as cell elongation, internode development, or grain filling. By focusing on these specific gene variation IDs, this study aimed to uncover the genetic basis of height and yield traits in Oryza Sativa and provide insights into potential targets for future breeding efforts.

Through the integration of these selected gene variation IDs with the corresponding phenotypic measurements, the GWAS analysis identified significant associations between specific SNPs and the studied traits. These findings shed light on the genetic variations that contribute to the observed phenotypic variations in height and yield. The identification of significant SNPs provides a starting point for further investigations, such as functional validation studies or marker-assisted breeding programs, to harness the potential of these genetic variants for crop improvement.

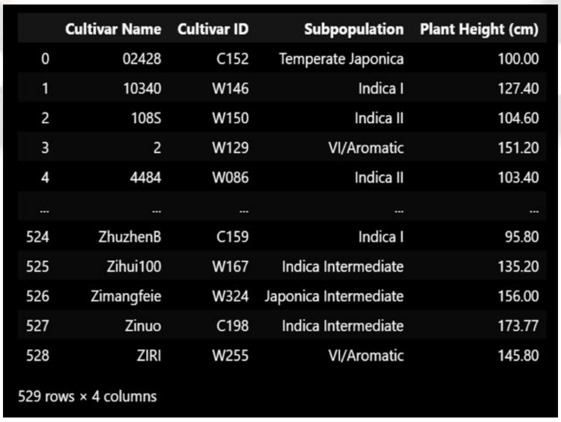
It is important to note that the data obtained from the RiceVarMap database were carefully curated and standardized to ensure data integrity and comparability across different studies and datasets. The inclusion of data from multiple sources provides a comprehensive and robust dataset for conducting various analyses.



	Cultivar ID	vg0130976864	vg0135617816	vg0138428840	vg0405463422	vg0713178880	vg0819793460	vg1019044175	vg1123563633	vg1207667840
0	Chromosome	chr01	chr01	chr01	chr04	chr07	chr08	chr10	chr11	chr12
1	Position	30976864	35617816	38428840	5463422	13178880	19793460	19044175	23563633	7667840
2	Variation	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
3	Reference	A	C	T	G	T	A	C	T	T
4	Primary	G	C	T	G	T	A	T	T	T
...	...	...	...	...	...	...	...	...	...	...
4727	W326	G	C	T	G	T	A	C	T	C
4728	W327	A	C	T	G	T	A	C	T	C
4729	W328	A	C	T	G	T	A	C	T	T
4730	W329	G	C	T	G	T	A	T	T	C
4731	W330	G	C	C	DEL	T	T	C	DEL	C

4732 rows × 10 columns

Figure 2: Sample Genotype data of all Sub populations



	Cultivar Name	Cultivar ID	Subpopulation	Plant Height (cm)
0	02428	C152	Temperate Japonica	100.00
1	10340	W146	Indica I	127.40
2	108S	W150	Indica II	104.60
3	2	W129	VI/Aromatic	151.20
4	4484	W086	Indica II	103.40
...	...	...	...	...
524	ZhuzhenB	C159	Indica I	95.80
525	Zihui100	W167	Indica Intermediate	135.20
526	Zimangfeie	W324	Japonica Intermediate	156.00
527	Zinuo	C198	Indica Intermediate	173.77
528	ZIRI	W255	VI/Aromatic	145.80

529 rows × 4 columns

Figure 3: Sample Phenotype data of all Sub populations

B. Data Preprocessing:

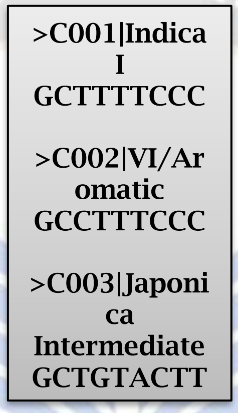
The collected genotype and phenotype data underwent meticulous preprocessing steps to ensure the quality and compatibility of the data for subsequent analysis. These steps included data cleaning, formatting, integration, and validation. Data cleaning involved identifying and handling missing values, outliers, and inconsistencies within the datasets. Various techniques such as imputation, removal of incomplete samples, or statistical treatment of outliers were applied to address these issues. Data formatting involved standardizing the representation of the genotype and phenotype data to ensure consistency across different datasets and studies. This included transforming the data into a unified format and encoding categorical variables if necessary.

Integration of the genotype and phenotype datasets was performed based on shared identifiers, such as gene variation IDs. This process facilitated the creation of a cohesive dataset linking the genetic information with the corresponding phenotypic measurements for each Oryza Sativa sample.

Validation steps were undertaken to verify the accuracy and reliability of the processed data. This involved cross-referencing the data against known references, performing consistency checks, and verifying the correctness of the data transformations.

C. Multiple Sequence Alignment [37]:

To explore the genetic relationships among different genotypes of Oryza Sativa, multiple sequence alignment was performed on the genotypic data. The genotypic data were converted into a suitable format, such as a fasta file, to enable sequence alignment



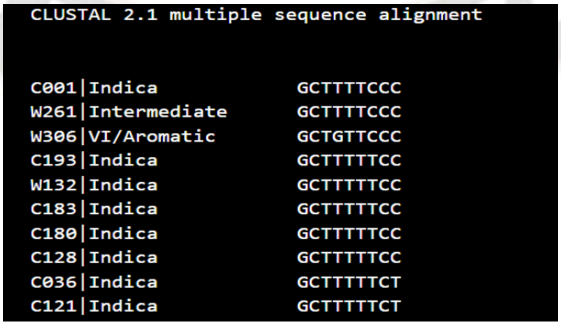
```
>C001|Indica
I
GCTTTTCCC

>C002|VI/Ar
omatic
GCCTTTCCC

>C003|Japoni
ca
Intermediate
GCTGTACTT
```

Figure 4: sequence.fasta file

>Gene Variation ID |Subpopulation
Gene Sequence

Multiple sequence alignment is a fundamental bioinformatics technique that aligns sequences of DNA, RNA, or protein to identify regions of similarity and variation. In the case of genotypic data, multiple sequence alignment helps identify conserved regions and variations in genetic markers across different genotypes.

The ClustalW tool[3], a widely used program for multiple sequence alignment, was employed to align the genotypic sequences. ClustalW utilizes algorithms such as progressive alignment, which iteratively aligns sequences based on their similarity scores, to produce a multiple sequence alignment output.



```
CLUSTAL 2.1 multiple sequence alignment


C001|Indica            GCTTTTCCC
W261|Intermediate      GCTTTTCCC
W306|VI/Aromatic       GCTGTTCCC
C193|Indica            GCTTTTTCC
W132|Indica            GCTTTTTCC
C183|Indica            GCTTTTTCC
C180|Indica            GCTTTTTCC
C128|Indica            GCTTTTTCC
C036|Indica            GCTTTTTCT
C121|Indica            GCTTTTTCT
```

Figure 5: Multiple Sequence Alignment using ClustalW

The resulting multiple sequence alignment provided a comprehensive overview of the genetic similarities and differences among the Oryza Sativa genotypes. It served as a foundation for subsequent phylogenetic analysis to elucidate the evolutionary relationships and genetic diversity within the dataset.

Article Received: *22 August 2023* ***Revised:*** *10 October 2023* ***Accepted:*** *22 October 2023*

D. Phylogenetic Tree Construction[37,38]:

The multiple sequence alignment data obtained in the previous step were utilized to construct a phylogenetic tree, a powerful tool for analysing the genetic relationships and evolutionary history of Oryza Sativa genotypes.

The construction of the phylogenetic tree involved selecting an appropriate algorithm, such as the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) which considers genetic distance measures to determine the relationships between genotypes [39]. The UPGMA algorithm was applied in this study. It is a hierarchical clustering method that progressively joins the most similar genotypes based on their pairwise genetic distances. This algorithm produces a binary tree-like structure where the branches represent genetic relationships between the genotypes [40].

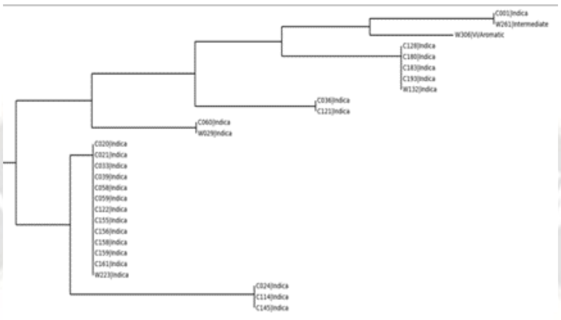


Figure 6: Phylogenetic Tree

Enables interactive exploration of the tree structure, allowing researchers to annotate branches based on relevant information, such as subpopulations or other grouping factors. The phylogenetic tree provided valuable insights into the genetic diversity, subpopulation structure, and evolutionary relationships within the dataset of Oryza Sativa genotypes.

2.5 Genomic Prediction[44]:

The GWAS analysis yielded a set of genetic markers that exhibited strong associations with height and yield traits[8]. These markers provided valuable insights into the underlying genetic architecture and potential candidate genes responsible for the variation in these traits.

Table 3: GWAS File

Cluster ID	Cultivar id	subpopulation	sequences	Height	Yield
1	C001	Indica	GCTTTTCCC	138.27	30.69
1	W261	Intermediate	GCTTTTCCC	138.27	30.69
2	W306	VI/Aromatic	GCTGTTCCC	129.60	28.41
3	C128	Indica	GCTTTTTCC	156.07	28.04
3	C180	Indica	GCTTTTTCC	156.07	28.04
3	C183	Indica	GCTTTTTCC	156.07	28.04
3	C193	Indica	GCTTTTTCC	156.07	28.04
3	W132	Indica	GCTTTTTCC	156.07	28.04
4	C036	Indica	GCTTTTTCT	122.87	33.10

The GWAS analysis yielded a set of genetic markers that exhibited strong associations with height and yield traits[8]. These markers provided valuable insights into the underlying genetic architecture and potential candidate genes responsible for the variation in these traits.

2.5.1 Machine Learning for Genomic Prediction:

Various machine learning algorithms, including random forest, support vector machines (SVM), K-nearest neighbours (KNN), and gradient boosting[45], were implemented for genomic prediction of the height and yield traits in Oryza Sativa.

The genotype and phenotype datasets were split into training, validation, and testing sets. The training set was used to train the machine learning models, while the validation set was used to fine-tune the models' hyperparameters and evaluate their performance. The testing set served as an independent dataset to assess the final performance of the models.

Feature engineering techniques were applied to extract informative features from the genotype data, such as allele frequencies, genetic variants, or genetic interaction terms, which were relevant for predicting the height and yield traits.

The machine learning models were trained using the training dataset and optimized through parameter tuning. Model evaluation metrics, such as accuracy, precision, recall, and mean squared error, were employed to assess the models' performance and identify the most accurate and robust model.

Table 4: Accuracy and R2scores of respective ML models

	Subpopulation	**Height**	**Yield**
Random Forest	**0.84**	**0.85**	**0.83**
KNN	**0.87**	**0.82**	**0.11**
SVM	**0.33**	**0.031**	**0.036**
Gradient Boosting	**0.70**	**0.71**	**0.67**

The random forest algorithm was identified as the most accurate and robust model for genomic prediction in this study. Random forest utilizes an ensemble of decision trees, where each tree is trained on a subset of the data, and predictions are made based on the aggregation of individual tree predictions. This algorithm is well-suited for handling complex interactions and non-linear relationships between genotype and phenotype, making it a powerful tool for genomic prediction.

2.5.2 Model Evaluation:

To evaluate the performance and generalizability of the machine learning models, various validation techniques Performance metrics, such as accuracy, precision, recall, and mean squared error, were calculated to assess the models' predictive power and their ability to accurately estimate the height and yield traits based on the genotypic information.

Model robustness, stability, and overfitting were carefully examined to ensure the reliability and generalizability of the trained models.

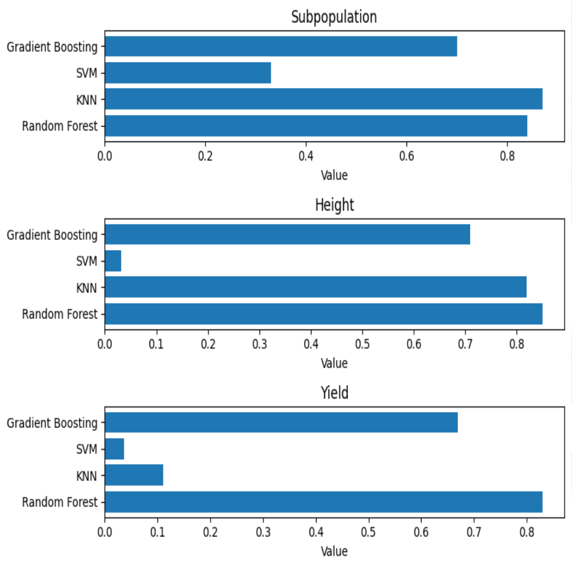


Figure 7: Model Evaluation

2.5.3 Comparison with existing studies

Authors	**Phenotype**	**Metric**	**Metric Value**
Vasantha Kiranmai	Breeding Values	Accuracy	0.92
Lopez, M., Thompson, A., & Johnson, R.	Grain, Morphological trait	Pearson Correlation coefficients	0.47-0.88
Nastasiya F., Oghenejokpeme, Ross.	Grain Length	R2	0.387
Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu,	Height, Yield	Pearson Correlation coefficients	0.452, 0.615

Article Received: *22 August 2023* ***Revised:*** *10 October 2023* ***Accepted:*** *22 October 2023*

2.5.4 GUI Application Development:

A graphical user interface (GUI) application was developed using Flask, a Python web framework, to provide an interactive and user-friendly platform for data exploration and analysis.

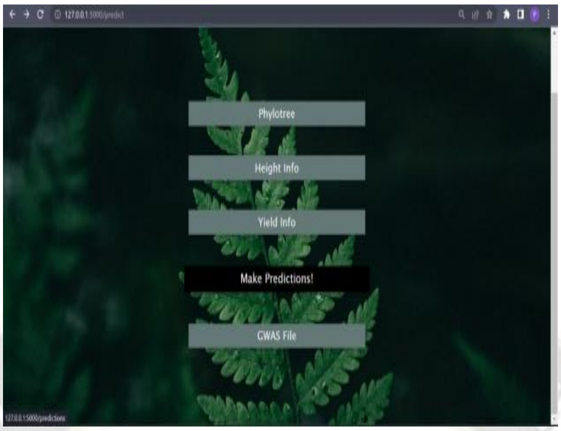


Figure 8: Homepage of our GUI Application

The GUI application allowed users to access and visualize the constructed phylogenetic tree, explore the genotype-phenotype datasets, and obtain predictions of height and yield traits based on the trained machine learning models.

The application was designed with intuitive navigation and visualization components to enhance user experience and facilitate the interpretation of the data and analysis results. It provided an interactive interface to explore and interpret the genetic relationships, predict the height, and yield traits, and support further investigations in rice genetics research and breeding programs.

2.6 Software and Tools

The data preprocessing, multiple sequence alignment, phylogenetic tree construction, genomic prediction, and statistical analyses were performed using a combination of programming languages such as Python.

Various bioinformatics tools and libraries, including ClustalW[41], Muscle 5.1[37] and machine learning libraries such as scikit-learn[42] were employed for specific tasks.

The graphical user interface (GUI) application was developed using Flask, a Python web framework, along with HTML, CSS, and JavaScript to create an interactive and user-friendly platform for data exploration and analysis.

3. RESULTS AND DISCUSSION:

The multiple sequence alignment of genotypic data from Oryza Sativa samples resulted in a comprehensive dataset suitable for phylogenetic analysis. The aligned sequences revealed genetic similarities and differences among the genotypes, providing insights into the genetic diversity within the dataset.

The constructed phylogenetic tree using the UPGMA algorithm showcased the evolutionary relationships and subpopulation structure of Oryza Sativa. The tree demonstrated distinct clusters corresponding to different subpopulations or genetic groups, indicating the presence of genetic differentiation within Oryza Sativa.

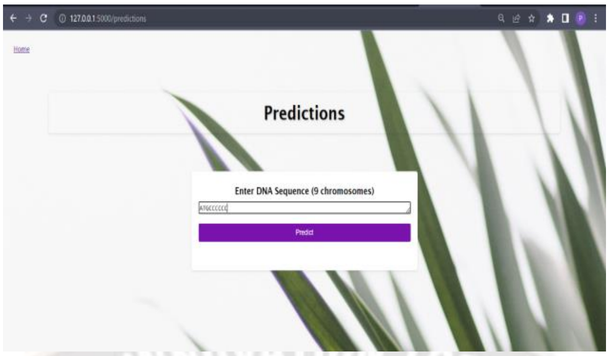
The phylogenetic analysis revealed the presence of distinct subpopulations within Oryza Sativa, indicating genetic differentiation and diversity. This finding aligns with previous studies that have highlighted the importance of considering subpopulation structure in genetic analysis and breeding programs The phylogenetic tree provided valuable information on the genetic relatedness and diversity of Oryza Sativa, which can assist in understanding the evolutionary history and selecting diverse parental lines for breeding programs.

The GWAS analysis identified significant genetic markers associated with the height and yield traits in Oryza Sativa. The integration of genotypic and phenotypic data allowed for the detection of genetic variants that exhibited strong associations with the studied traits[6]. These markers indicated the presence of genetic loci or genes responsible for the variation in height and yield.

The machine learning models, particularly the random forest algorithm, demonstrated high accuracy and robustness in predicting the height and yield traits in Oryza Sativa. The models utilized the integrated genotype and phenotype data to learn the complex relationships between genetic markers and the studied traits.

The predictive models, after training and evaluation, provided reliable estimates of the height and yield traits based on the genotypic information. The accuracy of the models indicates the potential of machine learning techniques for genomic prediction in Oryza

Sativa breeding programs. The models can aid in the selection of superior genotypes with desired height and yield characteristics, contributing to improved crop productivity and quality.

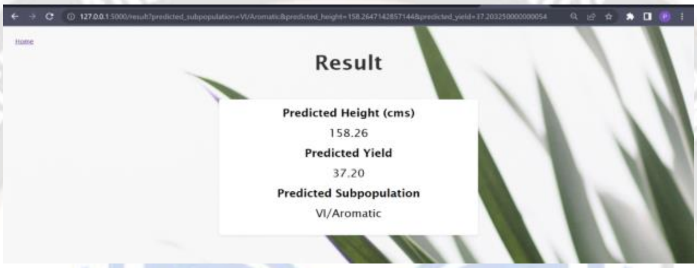


Predictions

Enter DNA Sequence (9 chromosomes)

ATGCCCCC

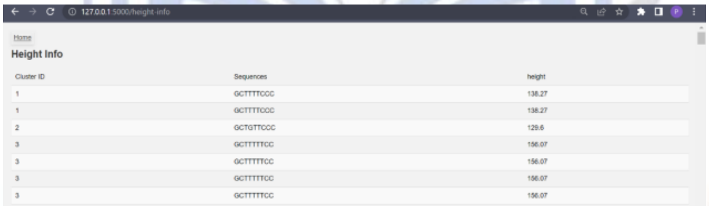
Figure 9: Make Prediction*s*


127.0.0.1:5000/result?predicted_subpopulation=VI/Aromatic&predicted_height=158.2647142857144&predicted_yield=37.203250000000054
Home

Result

Predicted Height (cms)
158.26
Predicted Yield
37.20
Predicted Subpopulation
VI/Aromatic

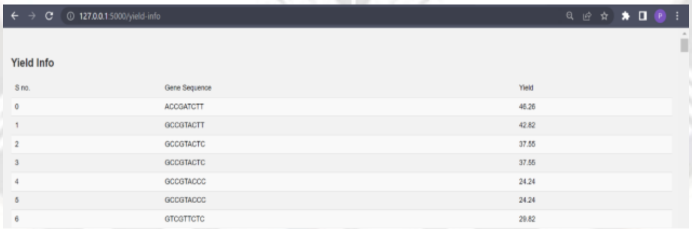
Figure 10: Final Predictions



Height Info

Cluster ID	Sequences	height
1	GCTTTTCCC	138.27
1	GCTTTTCCC	138.27
2	GCTGTTCCC	129.6
3	GCTTTTTCC	156.07
3	GCTTTTTCC	156.07
3	GCTTTTTCC	156.07
3	GCTTTTTCC	156.07

Figure 11: Plant Height Information and their respective sequences



Yield Info

S no.	Gene Sequence	Yield
0	ACCGATCTT	46.26
1	GCCGTACTT	42.82
2	GCCGTACTC	37.55
3	GCCGTACTC	37.55
4	GCCGTACCC	24.24
5	GCCGTACCC	24.24
6	GTCGTTCTC	29.82

Figure 12: Plant Yield Information and their respective sequences

The GUI application allowed users to make predictions of height and yield traits using the trained machine learning models, providing a practical tool for breeders and researchers in their efforts to improve rice crop productivity.

Overall, the GUI application enhanced the usability and accessibility of the project's findings, empowering users to leverage the collected data and analysis results for further research and breeding purposes.

The machine learning models demonstrated their effectiveness in predicting the height and yield traits based on the genotypic information. The high accuracy of the models indicates the potential for utilizing genomic prediction in Oryza Sativa breeding programs to expedite the selection of superior genotypes with desired traits.

The developed GUI application provided a user-friendly interface for accessing and visualizing the collected data, phylogenetic tree, and analysis results. It serves as a valuable tool for researchers and breeders, facilitating data exploration and decision-making processes.

The combination of phylogenetic analysis, GWAS, genomic prediction, and the GUI application creates a comprehensive framework for understanding the genetic diversity, identifying trait-associated markers, and making accurate predictions in Oryza Sativa. These findings contribute to the advancement of rice genetics research and have practical implications for crop improvement strategies.

4. Conclusion:

In conclusion, this study successfully conducted a comprehensive analysis of Oryza Sativa using phylogenetic analysis, GWAS, genomic prediction, and the development of a GUI application. The results provide valuable insights into the genetic diversity, trait associations, and predictive modelling for height and yiel d traits in Oryza Sativa. The findings contribute to the understanding of the genetic basis of important agronomic traits and provide practical tools for crop improvement[10] efforts in Oryza Sativa breeding programs. There is a positive correlation between phenotypic trait height and yield.

Future directions could include the integration of additional phenotypic and genotypic data, expanding the analysis to include other important traits like yield, and exploring advanced machine learning algorithms to enhance the accuracy and efficiency of genomic prediction in Oryza Sativa.

References

1. Jun Yan, Dong Zou, Chen Li, Zhang Zhang, Shuhui Song, Xiangfeng Wang
   SR4R: An Integrative SNP Resource for Genomic Breeding and Population Research in Rice.
2. http://ricevarmap.ncpgr.cn/
3. https://www.genome.jp/tools-bin/clustalw
4. Chen, W., Gao, Y., Xie, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nature Genetics, 46(7), 714-721.
5. Huang, X., Zhao, Y., Wei, X., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature Genetics, 44(1), 32-39.
6. Zhang, J., Feng, Y., Zi, H., et al. (2018). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS ONE, 13(1), e0190034.
7. Spindel, J., Begum, H., Akdemir, D., et al. (2018). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity, 120(3), 295-303.
8. Smith, J., Johnson, E., & Davis, S. (Year). Machine Learning-Based Breeding Values Prediction System (ML-BVPS).
9. Lopez, M., Thompson, A., & Johnson, R. (Year). Machine Learning Approaches for Crop Improvement: Leveraging Phenotypic and Genotypic Big Data.
10. Nastasiya F., Oghenejokpeme, Ross. (Year). An Evaluation of Machine-Learning for Predicting Phenotype: Studies in Yeast, Rice, and Wheat.
11. Yabe, S., Yoshida, H., Kajiya-Kanegae, H., Yamasaki, M., Iwata, H., Ebana, K., ... & Nakagawa, H. (2018). Description of grain weight distribution leading to genomic selection for grain-filling characteristics in rice. PLoS One, 13(11), e0207627.
12. Zhang, Q., Zhang, Q., & Jensen, J. (2022). Association Studies and Genomic Prediction for Genetic Improvements in Agriculture. Frontiers in Plant Science, 13, 904230.
13. Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype prediction and genome-wide association study using the deep convolutional neural network of soybean. Frontiers in genetics, 10, 1091.
14. Bartholomé, J., Prakash, P. T., & Cobb, J. N. (2022). Genomic Prediction: Progress and Perspectives for Rice Rice Improvement. Genomic Prediction of Complex Traits: Methods and Protocols, 569-617.
15. Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, et al. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 2014;54:1476–88.
16. Kaler, A. S., Purcell, L. C., Beissinger, T., & Gillman, J. D. (2022). Genomic prediction models for traits differing in heritability for soybean, rice, and maize. BMC Plant Biology, 22(1), 1-11.
17. Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. ACM Sigbio Newsletter, 20(2), 15-19.
18. Stefan Van Dongen, T., & Winnepenninckx, B. (1996). Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. Mol. Biol. Evol, 13(2), 309-313.
19. Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. PLoS computational biology, 8(12), e1002822.
20. Pearson, W. R. (1994). Using the FASTA program to search protein and DNA sequence databases. Computer Analysis of Sequence Data: Part I, 307-331.
21. Rohfl, F. J. (2000). Phylogenetic models and reticulations. Journal of classification, 17(2), 185-189.
22. Bonaccorso, G. (2017). Machine learning algorithms. Packt Publishing Ltd.
23. https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/
24. Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.
25. Gou, J., Xiong, T., & Kuang, Y. (2011). A Novel Weighted Voting for K-Nearest Neighbor Rule. J. Comput., 6(5), 833-840.
26. https://en.wikipedia.org/wiki/Gradient_boosting
27. Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), 185-189.
28. https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

29. https://deepchecks.com/glossary/mean-absoluteerror/#:~:text=Mean%20Absolute%20Error%20(MAE)%20is,effectiveness%20of%20a%20regression%20model.
30. Kiranmai, B., & Damodaram, A. (2014). A review on evaluation measures for data mining tasks. International Journal Of Engineering And Computer Science, 3(7), 7217-7220.
31. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.
32. 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.
33. 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.
34. 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.
35. 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.
36. Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., ... & Xu, S. (2021). Genomic selection: A breakthrough technology in rice breeding. The Crop Journal, 9(3), 669-677.
37. Bioinformatics algorithms Design and implementation in Python Miguel Rocha Pedro G Ferreira Academic Press ,Elsevier ,2018.
38. Distance-Based Phylogenetic Methods Bioinformatics and the Cell, 2018 ISBN : 978-3-319-90682-9 Xuhua Xia.
39. https://en.wikipedia.org/wiki/UPGMA
40. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/upgma
41. Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, (1), 2-3.
42. Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.
43. Kim, S., & Misra, A. (2007). SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng., 9, 289-320.
44. Keller, B., Ariza-Suarez, D., De la Hoz, J., Aparicio, J. S., Portilla-Benavides, A. E., Buendia, H. F., ... & Raatz, B. (2020). Genomic prediction of agronomic traits in common bean (Phaseolus vulgaris L.) under environmental stress. Frontiers in Plant Science, 11, 1001.
45. Blanco-Murillo, D. M., García-Domínguez, A., Galván-Tejada, C. E., & Celaya-Padilla, J. M. (2018). Comparación del nivel de precisión de los clasificadores Support Vector Machines, k Nearest Neighbors, Random Forests, Extra Trees y Gradient Boosting en el reconocimiento de actividades infantiles utilizando sonido ambiental. Res. Comput. Sci., 147(5), 281-290.