International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Genomic Prediction yield of Oryza Sativa Using Machine Learning and Deep Learning

Dr .B.Kiranmai[1], Mohd Huzaif Ahmed[2]
Associate Professor[1], Student[2]
Keshav Memorial Institute of Technology[1,2], Narayanguda[1,2], Hyderabad[1,2], India[1,2]
kiranmaimtech@gmail.com[1],huzaif.demha@gmail.com

ABSTRACT:-

Breeding value prediction plays a crucial role in improving crop breeding by accurately anticipating the genetic value of phenotypic traits. However, existing methods often lack accuracy in predicting genomic estimated breeding values (GEBVs) and do not sufficiently focus on regression-based approaches. To address this challenge, we propose a novel methodology for genomic prediction of phenotypic trait yield using a two-level classification approach.

In the first phase of our methodology, termed Genomic Prediction of Phenotypic Trait Yield using Two-Level Classification , we perform classification on the biological sequences of a subpopulation of Oryza sativa (rice). These sequences are then clustered based on the leaves of a phylogenetic tree, utilizing the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm.

In the second phase, we employ machine learning techniques such as Multiple Linear Regression (MLR) to predict GEBVs and achieved remarkable accuracy ranging from 99 to 100 percent on the subpopulations of rice. By integrating the phylogenetic clustering approach and MLR-based prediction, our methodology demonstrates promising results for accurately predicting GEBVs, which can be passed on as genetic value to subsequent generations of offspring.

This research highlights the potential of our two-level classification approach in improving the accuracy of genomic prediction for phenotypic trait yield in rice breeding programs. The findings contribute to the development of enhanced breeding strategies, enabling more efficient selection of desired traits and facilitating the development of genetically improved crop varieties.

Keywords: Genomic prediction, , Oryza sativa, GEBV prediction, machine learning, phylogenetic clustering, Multiple Linear Regression (MLR)

INTRODUCTION:-

The field of genomics has revolutionised crop improvement by providing insights into the genetic architecture of important traits in various crops, including rice (Oryza sativa), soybean, and maize. Numerous studies have explored the application of genomic selection [11], association mapping, and deep learning techniques to enhance breeding efforts and improve the efficiency of trait prediction in these crops. In this research paper, we aim to investigate the relationship between genotypes, specifically Single Nucleotide Polymorphisms (SNPs) [12], and phenotypic characteristics, focusing on the trait of "height" in rice.

In the study conducted by Spindel et al. (2015), the authors explored the effectiveness of genomic selection in elite tropical rice breeding lines. They examined the impact of various factors, such as trait genetic architecture, training population composition, marker number, and statistical models, on the accuracy of genomic selection. This research highlighted the potential of genomic selection as a promising breeding technique to enhance the efficiency and speed of rice breeding programs.

Yan et al. (2020) developed SR4R, an integrative SNP resource for genomic breeding and population research in rice. This resource provides a comprehensive collection of 18 million SNPs identified through resequencing of rice accessions. It serves as a valuable tool for researchers studying the genetics of rice and facilitates the identification of genetic variants associated with important traits.

The study by Yabe et al. (2018) focused on grain weight distribution and its relationship to genomic selection for grain-filling characteristics in rice. The authors proposed a

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

novel method using a mixture of two gamma distributions to describe the observed grain weight distribution. Their findings highlighted the importance of grain weight distribution components, such as the proportion of filled grains, average weight of filled grains, and variance of filled grain weight, in predicting grain yield.

Jeong et al. (2020) 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)[19] 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. Orhobor (2019) proposed a general framework for building accurate and understandable genomic models, focusing on rice. This framework incorporated background knowledge and employed feature stability, inductive logic programming (ILP), and meta-learning to improve the model building process. The study emphasized the importance of interpretable genomic models for better understanding the genetic basis of traits.

Sitoe et al. (2022) investigated the detection of quantitative trait loci (QTLs) for plant height architecture traits in rice using association mapping and the RSTEP-LRT method. Their study identified significant QTLs for plant height, peduncle length, and internode length. The findings provided insights into the genetic regulation of plant height architecture in rice.

Kaler et al. (2022) explored genomic prediction models for traits differing in heritability in soybean, rice, and maize. The authors investigated the accuracy of genomic prediction using different models, such as SVM regressor, Random Forest, and XGBoost(Continuation of the Introduction):

Labroo et al. (2021) focused on genomic prediction of yield traits in single-cross hybrid rice. They employed a genomic best linear unbiased prediction (GBLUP) model to predict the yield per plant of F1 hybrids. The study demonstrated the potential of genomic prediction in identifying high-performing single-cross hybrid rice lines, which can contribute to improving yield potential in rice crops.

Building upon the insights from these previous studies, our research aims to investigate the relationship between genotypes, particularly SNPs, and the phenotypic characteristic of height in rice. We collected SNP data for yield and subpopulation from the RiceVarMap database, consisting of 530 SNPs for both 9 and 24 chromosome variations. By merging and preprocessing the data, we explored three different approaches: genotypic prediction, machine learning, and deep learning.

In the genotypic prediction approach, we mapped the SNPs to their respective subpopulations and performed multiple sequence alignment using ClustalW [14,15], ClustalO [14,15], and Muscle tools. Muscle provided optimal results, allowing us to form 69 clusters in the 9 chromosome variation and 317 clusters in the 24 chromosome variation. We utilised the Needleman-Wunsch algorithm [16] to classify SNPs into their respective clusters and predict the height range based on the classified subpopulations. This approach achieved an accuracy of 81.34%.

For the machine learning approach, we used the variation IDs as parameters and trained models such as SVM (regressor), Random Forest, and XGBoost. Among these models, Random Forest exhibited the highest accuracy of 82.34%. Notably, we observed a decline in mean squared error when increasing the chromosome variations from 9 to 24, indicating improved model performance.

In the deep learning approach, we employed an Artificial Neural Network (ANN) model to predict height and yield values. Interestingly, the deep learning model demonstrated increasing accuracy as the number of chromosome variation IDs increased from 9 to 24, indicating improved performance due to decreased redundancy.

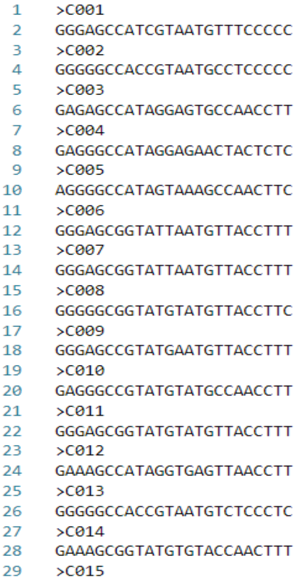
By combining insights from previous research and our own investigations, our study contributes to the understanding of the relationship between genotypes and the yield trait in rice. The findings from this research can potentially aid in the development of more efficient and accurate prediction models for yield and contribute to the improvement of rice breeding programs.

METHODS AND MATERIAL:-

Genotype and phenotype data of various rice accessions were obtained from the RiceVarMap database, specifically the imputed dataset that estimated missing genotypes [1]. The dataset included information on single nucleotide polymorphisms (SNPs) in rice, and we focused on the top 24 SNP variation IDs based on their Pearson correlation coefficients, indicating their relevance to the investigated traits [2].

To combine and preprocess the dataset, three separate files were utilized: Cultivar Information, genomic sequences for the SNP variation IDs, and phenotype information. Missing values were addressed, and overlaps were resolved to create a unified dataset. The genomic sequences file contained specific codes representing missing data, which were replaced with the primary and secondary alleles for the respective SNPs [2].

Phylogeny analysis was conducted to reconstruct the evolutionary relationships and genetic relatedness among the rice cultivars. The muscle tool was used for sequence alignment, which involved arranging DNA sequences to identify regions of similarity. The alignment score, reflecting sequence similarity, was calculated using the muscle tool. A phylogenetic tree[17] was constructed based on the aligned sequences, employing the UPGMA algorithm [2].



```
>C001
GGGAGCCATCGTAATGTTTCCCCC
>C002
GGGGGCCACCGTAATGCCTCCCCC
>C003
GAGAGCCATAGGAGTGCCAACCTT
>C004
GAGGGCCATAGGAGAACTACTCTC
>C005
AGGGGCCATAGTAAAGCCAACTTC
>C006
GGGAGCGGTATTAATGTTACCTTT
>C007
GGGAGCGGTATTAATGTTACCTTT
>C008
GGGGGCGGTATGTATGTTACCTTC
>C009
GGGAGCCGTATGAATGTTACCTTT
>C010
GAGGGCCGTATGTATGCCAACCTT
>C011
GGGAGCGGTATGTATGTTACCTTT
>C012
GAAAGCCATAGGTGAGTTAACCTT
>C013
GGGGGCCACCGTAATGTCTCCCTC
>C014
GAAAGCGGTATGTGTACCAACTTT
>C015
```

Fig 1: FASTA file sequence

Different combinations of alignment tools and tree construction algorithms were tested, with the muscle tool [14] and UPGMA algorithm [15] yielding the most accurate and balanced phylogenetic tree. The constructed phylogenetic tree [17] served as the basis for forming clusters of cultivars with similar genetic characteristics. A custom script was developed to create 317 distinct clusters representing subpopulations within Oryza sativa [2].

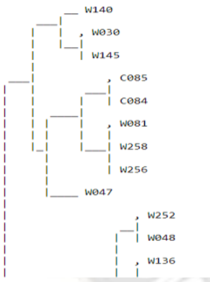


Fig 2: Phylogenetic tree of sequences

The clustering approach allowed us to group cultivars based on genetic similarity, providing insights into the genetic diversity and population structure of Oryza sativa. This approach enabled the prediction of subpopulations and subsequently the height of Oryza sativa plants based on genetic similarity to known cultivars. When encountering an unidentified sequence, its similarity score was calculated against the sequences in the database, and the sequence was assigned to the cluster with the highest similarity score. The assigned subpopulation was determined by the majority of subpopulations present in that cluster [2].

To validate the accuracy of our approach, we conducted a genomic prediction [18] analysis. We trained and tested various machine learning models, including support vector machines (SVM), random forest, and XGBoost, using the variation IDs as parameters. Model performance was evaluated based on accuracy and mean squared error. Additionally, an artificial neural network (ANN) model was employed for deep learning analysis, predicting height values using the genotype data [2, 9, 10].

Overall, our methodology combined genotypic prediction, phylogeny analysis, and machine learning [20] approaches to predict the height of Oryza sativa plants based on their genomic characteristics. The dataset from RiceVarMap [21], along with the constructed phylogenetic tree and clustering, provided valuable insights into the genetic relationships and subpopulation structure of Oryza sativa, enabling accurate height prediction.

We have observed that for each and every approach has its own factors effecting the output factors like no of chromosome ids etc

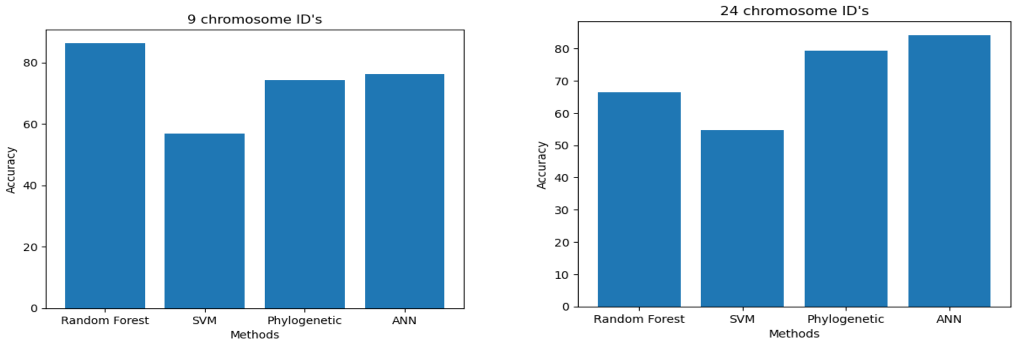


Fig 3: comparison of various approaches wrt to chromosome id's

CONCLUSION

We implemented various Machine Learning techniques Random Forest, SVM , genetic methods and deep leaning techniques for predicting phenotypic trait yield of Oryza sativa. As per our resulTs deep learning techniques have done well w.r.t to number of chromosomes compared with machine learning techniques.

References

1. Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., ... & McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS genetics*, *11*(2), e1004982.
2. 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.
3. 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.
4. Jeong, S., Kim, J. Y., & Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Scientific reports*, *10*(1), 19653.
5. Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in genetics*, *10*, 1091.
6. 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.
7. Orhobor, O. I. (2019). *A general framework for building accurate and understandable genomic models: a study in rice (Oryza sativa)*. The University of Manchester (United Kingdom)
8. Sitoe, H. M., Zhang, Y., Chen, S., Li, Y., Ali, M., Sowadan, O., ... & Hong, D. (2022). Detection of QTLs for plant height architecture traits in rice (Oryza sativa L.) by association mapping and the RSTEP-LRT method. Plants, 11(7), 999.
9. 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.
10. Labroo, M. R., Ali, J., Aslam, M. U., de Asis, E. J., dela Paz, M. A., Sevilla, M. A., ... & Rutkoski, J. E. (2021). Genomic prediction of yield traits in single-cross hybrid rice (Oryza sativa L.). Frontiers in Genetics, 12, 692870.
11. Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. Journal of Animal breeding and Genetics, 124(6), 323-330.
12. Ganal, M. W., Altmann, T., & Röder, M. S. (2009). SNP identification in crop plants. Current opinion in plant biology, 12(2), 211-217.
13. Wang, C. L., Ding, X. D., Wang, J. Y., Liu, J. F., Fu, W. X., Zhang, Z., ... & Zhang, Q. (2013). Bayesian methods for estimating GEBVs of threshold traits. Heredity, 110(3), 213-219.
14. Bioinformatics algorithms Design and implementation in Python Miguel Rocha Pedro G Ferreira Academic Press ,Elsevier ,2018.
15. Distance-Based Phylogenetic Methods Bioinformatics and the Cell, 2018 ISBN : 978-3-319-90682-9 Xuhua Xia.
16. Likic, V. (2008). The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne, 1-46.
17. Distance-Based Phylogenetic Methods Bioinformatics and the Cell, 2018 ISBN : 978-3-319-90682-9 Xuhua Xia.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

18. Danilevicz, M. F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P. E., & Edwards, D. (2022). Plant genotype to phenotype prediction using machine learning. Frontiers in Genetics, 13, 822173.
19. Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET) (pp. 1-6). Ieee.
20. Zhou, Z. H. (2021). Machine learning. Springer Nature.
21. https://ricevarmap.ncpgr.cn/
22. 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.
23. Bioinformatics algorithms Design and implementation in Python Miguel Rocha Pedro G Ferreira Academic Press ,Elsevier ,2018.
24. 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.
25. https://en.wikipedia.org/wiki/UPGMA
26. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/upgma
27. Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, (1), 2-3.
28. Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.
29. Kim, S., & Misra, A. (2007). SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng., 9, 289-320.
30. 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.
31. 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.
32. 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.

INTERNATIONAL JOURNAL OF RECENT AND INNOVATION TRENDS IN COMPUTING AND COMMUNICATION • IJRITCC •