Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 1, April 2024, pp. 575~583
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i1.pp575-583

DNA based phenotype optimization of oryza sativa using machine learning and MolCNN

Nikita Soren, Paramasivan Selvi Rajendran
Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India

Article Info

Article history:

Received Nov 20, 2023
Revised Jan 9, 2024
Accepted Jan 14, 2024

Keywords:

DNA
Genotype
Machine learning
MolCNN
Oryza sativa
Phenotype

ABSTRACT

The prediction of phenotype from the genotype of oryza sativa (rice) is very crucial for optimizing the crop management. utilizing molecular convolutional neural networks (MolCNNs) and machine learning for crop management in oryza sativa provides a data-driven method for phenotype prediction based on DNA data, improving farming techniques. Data gathering, preparation, and integration of phenotypic and DNA data are all part of this process. Meaningful DNA features are extracted by MolCNN, and phenotypic qualities are predicted by a machine learning algorithm. Making educated decisions is ensured by assessing the model's effectiveness, applying it to crop management, and updating it frequently. Choosing crop varieties, planting schedules, and management techniques are guided by molecular insights, which support sustainable agriculture and increase yields and quality. In the proposed research we are calculating pearson correlation coefficients between anticipated and actual trait values and the model's performance on a test set. Additionally, it determines the (PCC) for every characteristic in the model and we have received a binary accuracy of 0.9998 in 139 seconds.

Corresponding Author:

Nikita Soren
Department of Computer Science and Engineering, Hindustan Institute of Technology and Science
Chennai, India
Email: nikitashyren@gmail.com

1. INTRODUCTION

With the help of DNA data analysis, machine learning models and molecular convolutional neural networks (MolCNN) can predict rice plant phenotypic features with high reliability and accuracy. Compared to more conventional approaches, this can more accurately identify desired features like growth patterns, disease resistance, and yield potential. throughput screening of rice plants is made possible by machine learning models in conjunction with DNA-based phenotyping. This results in time and resource savings since numerous plants can be examined at once. Even before visual signs develop, machine learning algorithms can identify disease markers in rice plants. Early diagnosis minimizes crop loss and the requirement for chemical treatments by enabling prompt management. It is possible to optimize resource allocation by precisely anticipating plant features. For instance, waste and the impact on the environment can be decreased by using the exact amount of water, fertilizer, and pesticides. Machine learning algorithms can be used to identify certain genes and DNA markers linked to desired attributes. Crop breeding programs can use this knowledge to create new rice varieties with enhanced traits. Machine learning models can choose rice varieties that are more suited to shifting environmental factors like temperature, humidity, and soil quality based on the examination of DNA data. This is required in light of climate change in order to sustain crop yield. Agronomists and farmers that want to make data-driven decisions can benefit from using machine learning-based solutions. The optimal periods to plant, rotate crops, and apply other management techniques may be ascertained by examining both historical and current data. Naturally disease- and pest-resistant plants can be identified using machine learning methods. Planting these types can help the environment and human health by lowering the demand for chemical pesticides. Using real-time data, machine learning algorithms can optimize planting density, irrigation schedules, and fertilizer levels, so increasing agricultural yield. By employing data-driven insights, farmers can optimize crop management by cutting expenses associated with labour, resources, and inputs. Crop management based on machine learning encourages environmentally friendly farming practices. This entails conserving natural resources, cutting waste, and making the most use of available territory. Since rice is a staple crop for millions of people, machine learning -based crop management can increase productivity and efficiency, improving global food security. In conclusion, by applying machine learning and molecular clustering to optimize DNA-based phenotyping of oryza sativa (rice), crop management can be significantly altered. Enhancing phenotypic forecasts, boosting resource efficiency, and promoting global food security can all help achieve this. This technology has the power to change the rice growth process and open the door for data-driven, sustainable agriculture.

2. RELATED STUDIES

While there is a chance of overfitting with the new data, Liu *et al.* [1] implemented advanced computer algorithms to identify genetic links, predict soybean traits, and create automated pattern recognition. Combining DNA regulatory circuits with molecular convolutional neural networks suggests an innovative approach Xiong *et al.* [2] for molecular biology and computational modelling. Wang *et al.* [3] produced deep learning enables sophisticated analysis of complex plant genetic data. The deep learning approaches are more predictive than conventional methods Piecyk *et al.* [4] they may be able to increase the accuracy of 3D chromatin interaction predictions. Deep learning algorithms are highly effective at capturing intricate non-linear correlations within data. Qu *et al.* [5] implemented deep learning for the prediction of DNA binding protein. Without the requirement for human feature engineering, deep learning algorithms are able to automatically learn from and extract characteristics from raw DNA sequences. Predictions made as a result may become more automated and efficient. Namin *et al.* [6] implemented deep learning which works well with high-dimensional data, which includes the intricate and multi-dimensional datasets that are frequently used in phenotyping research. For jobs containing spatial relationships, like genetic data, CNNs are a good fit Ma *et al.* [7] this may be useful in identifying inter-regional relationships throughout the genome. Deep CNN training can be computationally demanding, particularly when using huge genomic datasets. Kalra [8] predicted plant genotype-phenotype correlations using sophisticated modeling capabilities provided by machine learning and deep learning approaches. Automation of feature extraction: by eliminating the need for manual feature engineering and perhaps capturing complex patterns, these methods can automatically learn and extract pertinent features from genomic data. Singh *et al.* [9] plant stress phenotypes is automatically and effectively analysed by deep learning, which eliminates the need for labour-intensive manual evaluations. Sperschneider [10] addresses the use of machine learning methods to comprehend the relationships between pathogens and plants. The power of machine learning to generate predictions at different scales from genomic data to field observations is highlighted in the essay. Danilevicz *et al.* [11] looks at how genetic data may be used to predict plant attributes using machine learning, which allows for more accuracy and efficiency in breeding. The work by Guo and Li [12] explores the use of machine learning to predict traits based on environmental and genetic information. Although biotechnology tactics can be tailored and agriculture optimized, there are still obstacles to overcome, including data complexity, model reliance, validation requirements, and ethical considerations with data privacy and equitable access. Das-Choudhury *et al.* [13] explores the use of neural networks for time series modeling with the goal of forecasting phenotypic outcomes and connecting them to genotypes. Poland and Rife, [14] paper addresses the use of genotyping-by-sequencing (GBS) in genetics and plant breeding. The GBS technology provides a high-throughput, low-cost means of producing genetic markers, enabling faster genetic research and improvements in plant breeding programs. It is likely that the integration of machine learning in plant science and breeding is the subject of van Dijk *et al.* [15] study. Through the analysis of complicated plant data, improved breeding accuracy, and more accurate trait prediction, machine learning can provide revolutionary insights. This technology has the potential to maximize resilience and agricultural productivity. The study by Crain *et al.* [16] combines genomic data and high-throughput phenotyping to improve selection and prediction accuracy in wheat breeding. Combining these cutting-edge methods allows scientists to find desired features more quickly and shorten the breeding process, which could result in improved crop types with higher yields and greater resilience. Shah and Wu [17] analyze crop and soil management techniques meant to increase crop yield while preserving environmental sustainability. Sustainable agriculture goals can be aligned with greater yields, preservation of soil health, and decreased environmental impact through the adoption of appropriate management strategies. The study conducted by Raja and Rajendran [18] compares several techniques for identifying and categorizing illnesses in banana leaves. The project intends to improve disease management by assessing several strategies, which could result in enhanced productivity and health of banana crops. By employing sophisticated detection techniques, crop losses could be reduced by enabling early disease detection and focused responses. The prophet algorithm is used in Soren and Rajendran [19] research to forecast and examine the growth of oryza sativa. Making use of these computational tools helps improve agricultural planning by predicting rice crop growth patterns and maximizing rice crop cultivation techniques. Precise growth forecasts have the potential to enhance yield approximations and resource administration, promoting more efficient and sustainable rice cultivation. Wang *et al.* [20] work, a deep neural network-based technique called DNNGP which makes use of multi-omics data is introduced for plant genome prediction. By enabling more accurate trait selection and crop improvement, this novel approach has the potential to revolutionize plant breeding and genetics research by improving the accuracy of genomic predictions. DNNGP may reveal intricate genetic linkages and offer thorough insights into plant biology by combining several data sources.

The study by Maldonado *et al.* [21] uses bayesian regularized neural network (BNN) and deep learning techniques to predict complex features across the genomes of two outcrossing plant species. To effectively utilize the potential of deep learning and BNN in plant genomics, however, a number of obstacles must be overcome, including interpretability of results, model optimization, and the integration of computational findings into practical breeding programs. Kono *et al.* [22] propose a structure-based method for predicting DNA target sites by regulatory proteins, enhancing accuracy compared to sequence-based methods. However, reliance on known protein structures limits its applicability, and computational complexity may hinder scalability. These challenges underscore the necessity of thorough validation and cooperative research. “G2PDeep” by Zeng *et al.* [23] provides a web-based deep-learning framework for genomic marker identification and phenotype prediction, which may improve the effectiveness of genetic research. Hou *et al.* [24] highlight the advantages of deep learning in genetic analysis and crop development as they talk about the technology’s promise in plant genomics and breeding. However, obstacles like model integration and data complexity highlight the necessity of ongoing study and modification for agricultural applications. The study by Montesinos-López *et al.* [25] uses densely constructed deep learners to predict plant features in many environments using genomics. This method may improve forecast accuracy in a variety of settings, supporting breeding and crop selection techniques that are more resilient. Nonetheless, obstacles including processing requirements, interpretability of the model, and incorporation into real-world agricultural environments could emerge, underscoring the significance of methodological improvement and cooperative study to maximize the advantages of deep learning in plant genomics.

3. PROPOSED METHOD

In our research we have implemented using MolCNN in conjunction with machine learning to maximize the DNA-based phenotype of oryza sativa. The process of optimizing the phenotype of oryza sativa by machine learning, notably MolCNN, and a DNA-based method makes use of genetic information to predict and enhance desired traits. Change these layers to efficiently process the molecular graph representation of DNA sequences. Specifically, we have developed layers that extract features from molecular graphs and encode them in a machine-learning style. We are using feature selection and pre-processing for DNA sequence data in order to improve the model’s performance and identify which DNA sequence features are most crucial for predicting a particular set of phenotypic traits.

The featured and pre-processed data is then sent to MolCNN for analysis. By integrating the strengths of genomics, bioinformatics, and machine learning, this approach maximizes the phenotype of oryza sativa based on DNA sequences. With MolCNN, we analyse learnt characteristics and determine which parts of the DNA sequence are most important for predicting particular phenotypic traits. The network is able to identify patterns linked to particular phenotypic outcomes, which leads to this optimization. The network is an effective tool for genotype-phenotype mapping in the fields of molecular biology and genetics because of its capacity to process molecular structures and capture complex correlations in genomic data. Our model Figure 1 is divided into two modules molecular and computational module. We are having the genome data and the phenotype data with the same target. We are integrating the genome and phenotype data. After integrating Figure 2 both the datasets we are proceeding with the preprocessing and feature selection. Then the featured and pre-processed data is fed to the CNN model with multiple hidden layers then we receive the output accordingly.

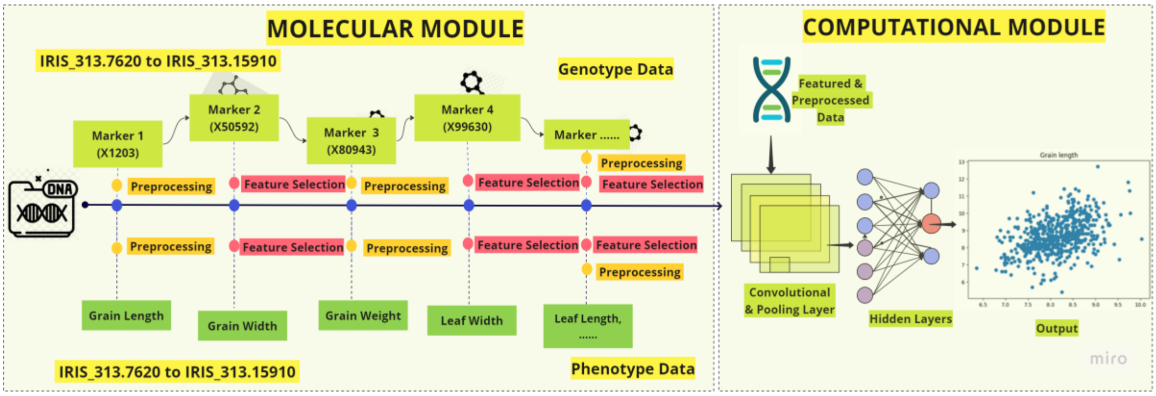
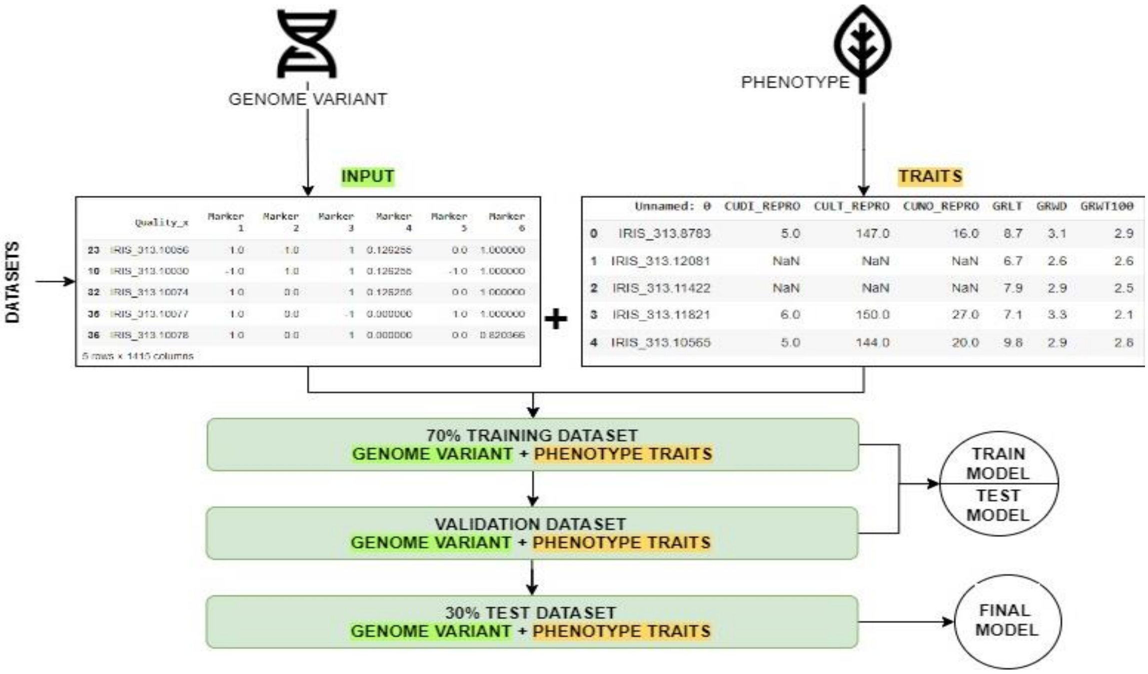


Figure 1. MOI-CNN architecture for phenotype prediction

4. METHOD

4.1. Data collection

IRIS datasets are collected from the international rice information system. In this study Figure 2, we are examining two distinct gene data sets: imputed and unimputed. Imputed genotype data is derived from the original, observable genetic data, but unimputed genotype data is derived from statistical techniques to anticipate missing genetic information.



	Quality_x	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6
23	IRIS_313.10066	1.0	1.0	1	0.126256	0.0	1.000000
10	IRIS_313.10030	-1.0	1.0	1	0.126256	-1.0	1.000000
32	IRIS_313.10074	1.0	0.0	1	0.126256	0.0	1.000000
36	IRIS_313.10077	1.0	0.0	-1	0.000000	1.0	1.000000
36	IRIS_313.10078	1.0	0.0	1	0.000000	0.0	0.820366

5 rows × 1415 columns

	Unnamed: 0	CUDI_REPRO	CULT_REPRO	CUNO_REPRO	GRLT	GRWD	GRWT100
0	IRIS_313.8783	5.0	147.0	16.0	8.7	3.1	2.9
1	IRIS_313.12081	NaN	NaN	NaN	6.7	2.6	2.6
2	IRIS_313.11422	NaN	NaN	NaN	7.9	2.9	2.5
3	IRIS_313.11821	6.0	150.0	27.0	7.1	3.3	2.1
4	IRIS_313.10565	5.0	144.0	20.0	9.8	2.9	2.8

Figure 2. Dataset representation

4.2. Genotype variant representation

The whole collection of genes and genetic material found in a single plant is referred to as the genotype. Capturing and communicating information about the precise genetic components that dictate an organism’s characteristics is known as genotype representation. The DNA sequence of the plant is the most basic source of genotype representation. The adenine, thymine, cytosine, and guanine nucleotide sequences that make up each gene are repeated across the genome. It is now possible to decipher the complete genomic sequence of different plant species because to advancements in DNA sequencing technologies

4.3. Phenotype (traits) representation

The visible features or attributes of an organism that arise from the interplay of its genetic composition (genotype) and external stimuli are referred to as phenotypes. These observable characteristics must be described and categorized in order to represent phenotypes. For oryza sativa, characteristics are quantifiable. A numerical measurement of height, weight, length, or any other phenotypic trait, for instance, can be used to quantitatively describe it.

4.4. Integrating the genotype and phenotype of oryza sativa

Sparsely recorded phenotypic characteristics, such plant height and germination rate, are frequently subjected to genotype to phenotype models. These traits are frequently gathered by hand, which may introduce bias and increase the expense of the experiment. It is essential to determine the connection between particular genetic variants (genotype) and related observable traits (phenotype). Finding genetic markers, mutations, or other genomic characteristics linked to specific phenotypic outcomes is part of this process.

4.5. Machine learning

Based on the genetic data of rice plants, machine learning is used to study and forecast their phenotypic features. The following machine learning related tasks are carried out in the research which are data loading and preprocessing, trait feature extraction, feedforward neural network, model compilation, training and evaluation, correlation and visualization and overall performance evaluation. Machine learning approaches evaluate model performance, forecast quantitative qualities of rice plants based on their genetic data, and provide light on the connections between phenotypic and genetic variables in rice. By allowing the prediction of rice plant traits, the machine learning models are utilized to optimize crop management and can help with agricultural decision-making.

4.6. Molecular convolutional neural network

MolCNN has been developed to work with molecular data, like DNA sequences or chemical compounds. It offers a practical means of transforming the sequence or chemical structure into a machine-learning-friendly format. MolCNN is useful in representing the DNA sequences linked to different strains of oryza sativa. MolCNN uses convolutional layers to extract salient characteristics from the molecular data stream. It is able to extract structural information and hierarchical patterns from DNA sequences. Understanding the genetic differences and characteristics between various rice strains may depend on these characteristics. MolCNN is included into a machine learning process to produce prediction models once the features have been acquired. Numerous rice phenotypic characteristics or attributes can be predicted by the DNA sequences. MolCNN, for instance, can assist in the prediction of characteristics such as growth features, yield, resistance to disease, and reactions to environmental stimuli. MolCNN's ability to extract pertinent characteristics from DNA sequences and create prediction models is essential for optimizing DNA-based phenotypes in oryza sativa. It helps to better understand rice characteristics and makes genetic data analysis easier, both of which are essential for crop management and breeding programs.

5. RESULT

The MolCNN are proposed to optimize the DNA-based phenotype of oryza sativa. In phenotype traits we are considering 12 main attributes of oryza sativa which are culm diameter, culm length, culm number, grain length, grain width, grain weight, heading date, ligule length, leaf length, leaf width, panicle length, and seedling height. Missing genotypes can be computationally handled as a genotype category and coded using the one-hot binary coding approach with deep learning. We used a one-hot vector with four channels to code the genotype matrices, both raw and imputed, and then we applied the same deep learning architecture to them. In the Figure 3 we are predicting the phenotype with the imputed and unimputed genotype using MoI-CNN. Pearson score (1):

$$r = \sum i = 1n(Xi - X^{-})2\sum i = 1n(Yi - Y^{-})2\sum i = 1n(Xi - X^{-})(Yi - Y^{-}) \qquad (1)$$

where, number of data points is n, and the individual data points for variables X and Y are Xi and Yi, respectively. The means of variables X and Y are, respectively, X and Y.

On Un-imputed genotypes compared to imputed genotypes, deep learning-based approaches have greater pearson score see in Figure 3. About 15% of the genotypes in the quality-assured raw datasets for the oryza sativa dataset are missing. The imputation method fills in the majority of missing genotypes with reference alleles, which deflates the effects of different genotypes, which may be one reason why deep learning models perform better on raw datasets. The ability of some quantitative traits to be predicted may be compromised by imputation approaches, which incorporate missing genotype effects based on non-missing genotypes. By computing the covariance of the deviations of the two variables, X and Y, from their respective averages, normalized by the product of their standard deviations, this formula determines the linear relationship between the two variables. The resultant value, r, is a number between -1 and 1, where 0 denotes no linear correlation, -1 represents a perfect negative correlation, and 1 represents a perfect positive correlation. In Figure 4 the average pearson score of grain length, grain width, grain weight, heading date, leaf length, and leaf width features are predicted.

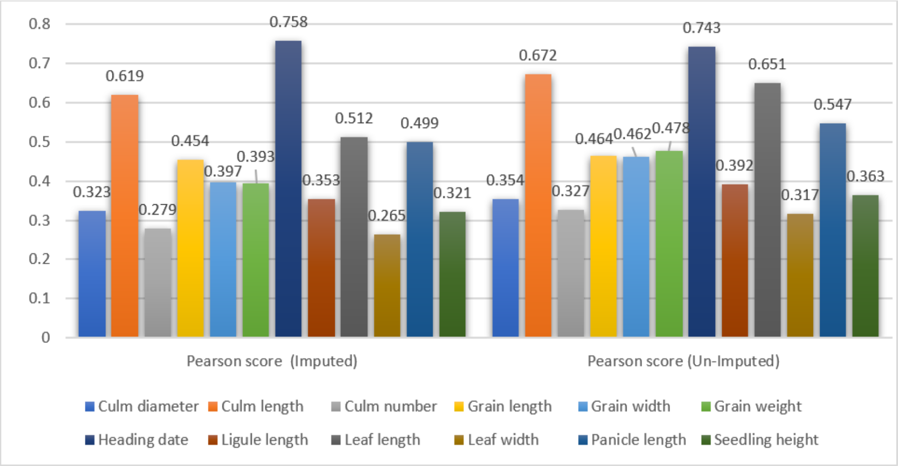


Figure 3. Principle attributes average PS

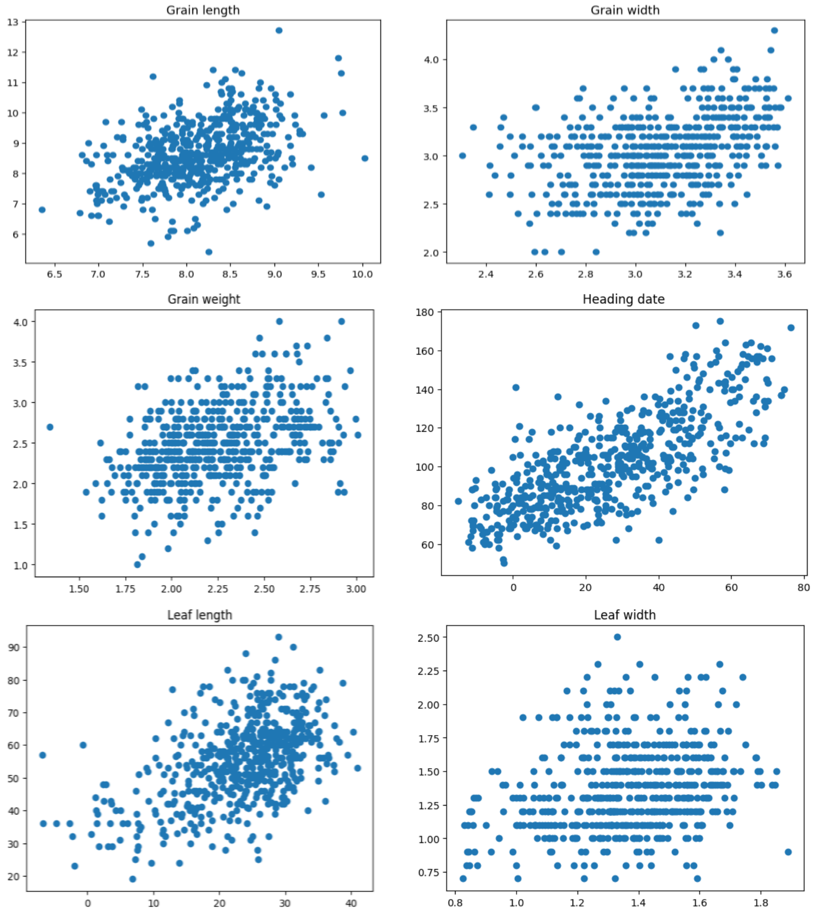


Figure 4. Average pearson score prediction

In Figure 5 while training accuracy evaluates the percentage of correct predictions made during the training phase, training loss gauges prediction accuracy. To guarantee efficient learning, the objective is to reduce loss and improve accuracy. In the Figure 6 during model training, validation loss quantifies the difference between expected and actual values on a different dataset with the goal of reducing discrepancies and improving the model’s generalization. On the testing set, trained with varying training set sizes, is displayed in Figure 5 and Figure 6. The x axis represents the predicted index and the y axis represents the actual index.

The percentage of accurate predictions on this validation dataset, or validation accuracy, measures the model’s overall efficacy on fresh, untested data and offers insights into how well it performs outside of the training set. It is essential to keep an eye on both measures to make sure the model is reliable and to avoid overfitting to the training set. The MolCNN outperforms with lowest training loss on validation set. After setting the epoch to 10, we were able to attain 0.9998 binary accuracy in the 7th epoch with a 0.0052 loss in 139 seconds.

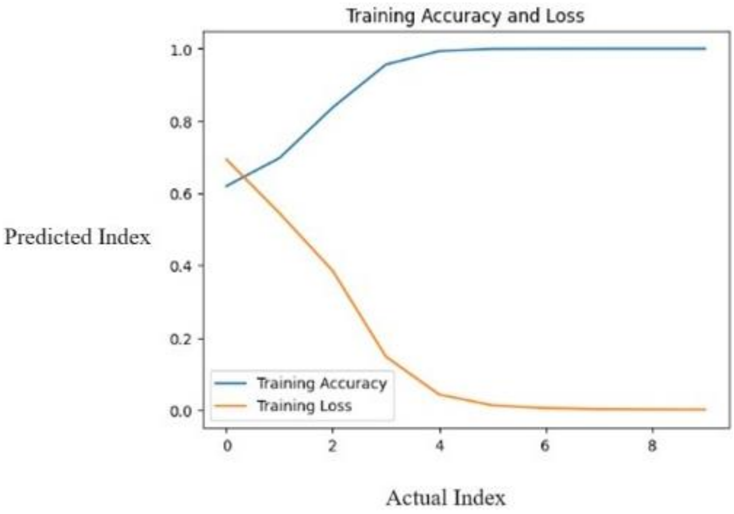


Figure 5. Training accuracy and loss

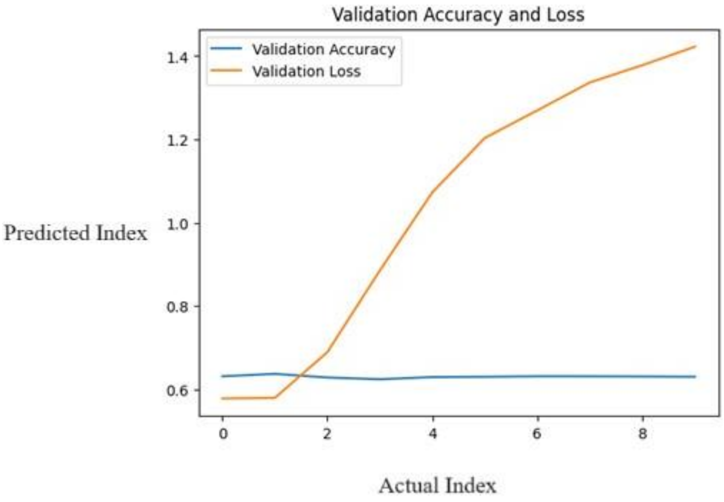


Figure 6. Validation accuracy and loss

6. CONCLUSION

In this paper we have proposed MolCNNs for identifying intricate connections in molecular data and predict the phenotype of oryza sativa. These networks are able to detect complex patterns and interactions between genes that lead to particular phenotypic features in the context of oryza sativa genotypes. MolCNNs are able to acquire hierarchical molecular structure representations. Because of this flexibility, the model can identify both large-scale genomic characteristics and minute changes that influence how many phenotypes in oryza sativa are expressed. MolCNNs are powerful tools for automatically identifying pertinent features in unprocessed genomic data. In order to get important insights into the genetic basis of observed differences, feature extraction is essential for finding important genetic markers or sequences linked to particular phenotypic features. The genetic landscape of oryza sativa is varied. Because MolCNNs can handle a wide range of genomic data, they are a good tool for examining genetic variation across populations and kinds of rice. The model's capacity to generalize across a range of genotypes is improved by this flexibility. In future work we would like to make a user-friendly architecture where we can use the image as an input and predict the phonotype for better crop management.

REFERENCES

[1] Y. Liu, D. Wang, F. He, J. Wang, T. Joshi, and D. Xu, "Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean," *Frontiers in Genetics*, vol. 10, Nov. 2019, doi: 10.3389/fgene.2019.01091.

[2] X. Xiong *et al.*, "Molecular convolutional neural networks with DNA regulatory circuits," *Nature Machine Intelligence*, vol. 4, no. 7, pp. 625–635, Jul. 2022, doi: 10.1038/s42256-022-00502-7.

[3] H. Wang, E. Cimen, N. Singh, and E. Buckler, "Deep learning for plant genomics and crop improvement," *Current Opinion in Plant Biology*, vol. 54, pp. 34–41, Apr. 2020, doi: 10.1016/j.pbi.2019.12.010.

[4] R. S. Piecyk, L. Schlegel, and F. Johannes, "Predicting 3D chromatin interactions from DNA sequence using deep learning," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 3439–3448, 2022, doi: 10.1016/j.csbj.2022.06.047.

[5] Y.-H. Qu, H. Yu, X.-J. Gong, J.-H. Xu, and H.-S. Lee, "On the prediction of DNA-binding proteins only from primary sequences: A deep learning approach," *PLOS ONE*, vol. 12, no. 12, p. e0188129, Dec. 2017, doi: 10.1371/journal.pone.0188129.

[6] S. T. Namin, M. Esmaeilzadeh, M. Najafi, T. B. Brown, and J. O. Borevitz, "Deep phenotyping: deep learning for temporal phenotype/genotype classification," *Plant Methods*, vol. 14, no. 1, p. 66, Dec. 2018, doi: 10.1186/s13007-018-0333-4.

[7] W. Ma *et al.*, "A deep convolutional neural network approach for predicting phenotypes from genotypes," *Planta*, vol. 248, no. 5, pp. 1307–1318, Nov. 2018, doi: 10.1007/s00425-018-2976-9.

[8] K. Kalra, "Predicting the genotype to phenotype relationship in plants using machine learning and deep learning," Doctoral dissertation, University of Saskatchewan, 2023.

[9] A. K. Singh, B. Ganapathysubramanian, S. Sarkar, and A. Singh, "Deep learning for plant stress phenotyping: trends and future perspectives," *Trends in Plant Science*, vol. 23, no. 10, pp. 883–898, Oct. 2018, doi: 10.1016/j.tplants.2018.07.004.

[10] J. Sperschneider, "Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale," *New Phytologist*, vol. 228, no. 1, pp. 35–41, Oct. 2020, doi: 10.1111/nph.15771.

[11] M. F. Danilevicz *et al.*, "Plant genotype to phenotype prediction using machine learning," *Frontiers in Genetics*, vol. 13, May 2022, doi: 10.3389/fgene.2022.822173.

[12] T. Guo and X. Li, "Machine learning for predicting phenotype from genotype and environment," *Current Opinion in Biotechnology*, vol. 79, p. 102853, Feb. 2023, doi: 10.1016/j.copbio.2022.102853.

[13] S. Das-Choudhury, "Time series modeling for phenotypic prediction and phenotype-genotype mapping using neural networks," 2020, pp. 228–243. doi: 10.1007/978-3-030-65414-6_17.

[14] J. A. Poland and T. W. Rife, "Genotyping-by-sequencing for plant breeding and genetics," *The Plant Genome*, vol. 5, no. 3, Nov. 2012, doi: 10.3835/plantgenome2012.05.0005.

[15] A. D. J. van Dijk, G. Kootstra, W. Kruijer, and D. de Ridder, "Machine learning in plant science and plant breeding," *iScience*, vol. 24, no. 1, p. 101890, Jan. 2021, doi: 10.1016/j.isci.2020.101890.

[16] J. Crain, S. Mondal, J. Rutkoski, R. P. Singh, and J. Poland, "Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding," *The Plant Genome*, vol. 11, no. 1, Mar. 2018, doi: 10.3835/plantgenome2017.05.0043.

[17] F. Shah and W. Wu, "Soil and crop management strategies to ensure higher crop productivity within sustainable environments," *Sustainability*, vol. 11, no. 5, p. 1485, Mar. 2019, doi: 10.3390/su11051485.

[18] N. B. Raja and P. S. Rajendran, "Comparative analysis of banana leaf disease detection and classification methods," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Mar. 2022, pp. 1215–1222. doi: 10.1109/ICCMC53470.2022.9753840.

[19] N. Soren and P. S. Rajendran, "Growth prediction and analysis of oryza sativa using prophet algorithm," 2023, pp. 625–636. doi: 10.1007/978-981-19-9819-5_45.

[20] K. Wang, M. A. Abid, A. Rasheed, J. Crossa, S. Hearne, and H. Li, "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants," *Molecular Plant*, vol. 16, no. 1, pp. 279–293, Jan. 2023, doi: 10.1016/j.molp.2022.11.004.

[21] C. Maldonado *et al.*, "Genome-wide prediction of complex traits in two outcrossing plant species through deep learning and bayesian regularized neural network," *Frontiers in Plant Science*, vol. 11, Nov. 2020, doi: 10.3389/fpls.2020.593897.

[22] Kono, Hidetoshi, and A. Sarai, "Structure-based prediction of DNA target sites by regulatory proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, no. 1, pp. 114-131, doi: 10.1002/(SICI)1097-0134.

[23] S. Zeng, Z. Mao, Y. Ren, D. Wang, D. Xu, and T. Joshi, "G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers," *Nucleic Acids Research*, vol. 49, no. W1, pp. W228–W236, Jul. 2021, doi: 10.1093/nar/gkab407.

[24] X. Hou, Y. Cui, and J. Liu, "Applications and prospect analysis of deep learning in plant genomics and crop breeding," *Journal of Library and Information Science in Agriculture*, vol. 34, no. 8, pp. 4–18, 2022, doi: 10.13998/j.cnki.issn1002-1248.22-0101.

[25] A. Montesinos-López, O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez, "Multi-environment genomic prediction of plant traits using deep learners with dense architecture," *G3 Genes|Genomes|Genetics*, vol. 8, no. 12, pp. 3813–3828, Dec. 2018, doi: 10.1534/g3.118.200740.

BIOGRAPHIES OF AUTHORS

Nikita Soren received the B.Sc. Degree in Computer Science in 2018 and M.Sc. in Information Technology from SRM Institute of Science and Technology in 2021, and doing Ph.D. in Computer science and engineering at Hindustan Institute of Technology and Science, Chennai, India. Her research interests include artificial intelligence, image processing, deep learning, and machine learning. She can be contacted at email: nikitashyren@gmail.com.

Dr. Paramasivan Selvi Rajendran has 22 years of teaching experience and she received the B.E. degree and M.E. in Computer science and engineering from Madurai Kamaraj University, and Ph.D. in Computer science and engineering from the National Institute of Technology (NIT), Trichy, India. Her current research interests include natural language processing, deep learning, and machine learning. She is a Life Member of the Indian Society for Technical Education (ISTE), and Computer Society of India. She authored two books and published research papers in 59 international Journals. She can be contacted at email: selvir@hindustanuniv.ac.in.