

Big genetic data analysis to predict features in cross-breeding to increase food yields

Priya Vij 

*Corresponding Author. Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India. E-mail address: ku.priyavij@kalingauniversity.ac.in

Patil Manisha Prashant 

Research Scholar, Department of CS & IT, Kalinga University, Raipur, India. E-mail address: patil.manisha@kalingauniversity.ac.in

Abstract

Objective

Plant breeders (PB) have significantly improved agricultural output and quality by utilizing modern scientific and technological developments. Costs have decreased and the PB process has quickened due to the development of genomic tools and sequencing, especially since the human genome project. Addressing global issues pertaining to water resources and food security requires this progress. High-throughput phenotyping, precision agriculture, and crop-scouting have all been improved by the integration of cutting-edge technology such sensor systems, satellite images, robots, big data analytics, and genomics. These developments contribute to the growth of digital agriculture, which has the potential to transform PB by taking a more interdisciplinary approach. To examine the method by which new developments in digital agriculture, genomics, and sensor technologies are changing plant breeding, enhancing crop quality and productivity, and tackling global issues with water resource management and food security.

Results

Plant breeding has become faster and less expensive due to the combination of genetic tools, sequencing techniques, and contemporary agricultural technologies. Precision agriculture has greatly increased high-throughput phenotyping and crop scouting, by using technology like robotics, big data analytics, and satellite photography. These developments aid in the creation of sustainable, more effective farming methods.

Conclusions

An innovative approach for crop improvement is being developed by the ongoing integration of multidisciplinary technologies in plant breeding. It is anticipated that enhanced genomics and digital agriculture would improve plant breeders' capacities, allowing them to tackle the escalating problems of food and water security in a world that is becoming more interconnected by the day.

Keywords: Big data, cross breeding, food yield, prediction

Paper Type: Review Paper.

Citation: Vij P, Prashant PM (2024) Big genetic data analysis to predict features in cross-breeding to increase food yields. *Agricultural Biotechnology Journal* 16 (4), 237-250.

Agricultural Biotechnology Journal 16 (4), 237-250.

DOI: 10.22103/jab.2025.24004.1612

Received: September 17, 2024.

Received in revised form: November 25, 2024.

Accepted: November 26, 2024.

Published online: December 30, 2024.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.



© the authors

Introduction

According to forecasts, the world's population will maintain its current rate of growth or even accelerate in the coming decades. The population's demand for food is expected to increase at the same rate. Crop productivity is affected by a number of biological and environmental factors, which are exacerbated by human-induced climate change (Shivanna 2022). Plant breeding (PB) is crucial for developing new cultivars with higher yields, improved quality, and the ability to withstand various abiotic and biotic challenges (Swarup et al. 2021). Global wheat production has increased from 200 million tons in 1961 to 775 million tons in 2023 without any significant change in the total area under wheat cultivation. The main reason is primarily the advancement and implementation of semi-dwarf, high-yielding wheat varieties responsive to inputs and resistant to major pests and adverse conditions (Radhika & Masood 2022). Across wheat production there have been improvements in agronomic practices, automation, favourable regulations, and infrastructure. Moreover, data generation in agriculture and biotechnology has greatly increased in recent years due to the very rapid development of high-performance technologies (Mohammadabadi et al. 2024). These data are obtained from studying products,

foods, and biological molecules to understand the role of different aspects of agriculture in determining the structure, function, and dynamics of living systems (Pour Hamidi et al. 2017). Artificial neural networks have been proposed to alleviate limitation of traditional methods and can be used to handle nonlinear and complex data, even when the data is imprecise and noisy (Pour Hamidi et al. 2017). Agricultural data can be too large and complex to handle through visual analysis or statistical correlations. This has encouraged the use of machine intelligence or artificial intelligence (Ghotbaldini et al. 2019). Thus, this review aimed to examine the method by which new developments in digital agriculture, genomics, and sensor technologies are changing plant breeding, enhancing crop quality and productivity.

History

Genetic modification of crops has traditionally been based on conventional Cross-Breeding (CB) approaches, where breeding and selection of genotypes is primarily based on pedigree and quantitative ability (Srinivasa Rao et al. 2023). The development of improved crop cultivars has been greatly facilitated by careful evaluation of parents for various traits, focused CB, utilization of summer and winter shuttle breeding strategies to accelerate the PB cycle, critical evaluation of elite germplasms in key locations, and efficient database management (Zoran et al. 2022). The rapid development of DNA-sequencing technology has enabled the PB to obtain extensive genomic data on crops, which is highly beneficial for selection (Suyama et al. 2022). The emergence of numerous DNA-marker-based genotyping methods has dramatically expanded the pool of DNA indicators accessible to PB. This advancement enabled PB to select for plant efficiency according to their genetic marker component rather than relying solely on their phenotypic effectiveness, which is subject to various limits of selection effectiveness (Begna 2021).

Heterosis and the production of traditional varieties are typically intertwined in various CBs in the PB (Liu et al. 2020). Hybridization is a crucial method for PB, and the most important factor for effective hybridization is careful selection of parents. The effectiveness of the CB can only be determined after several generations, as the efficiency of the mixed progeny may not fully match that of their parents (Scott et al. 2020). The process of choosing parents is challenging. Suppose the quality of mating can already be determined in the first generations, by focusing on selecting the right parents. In this case, the quality of the combinations will be discovered at the earliest opportunity, leading to an improvement in breeding outcomes.

Genomic Prediction (GP) is a cutting-edge, data-driven approach that has gained widespread acceptance and is being extensively utilized as a beneficial tool to expedite the improvement of genetic traits in PB projects (Tsai et al. 2020). GP utilizes sophisticated statistical Machine-

Learning (ML) methods to identify individuals inside a breeding populace according to breeding values inferred from markers found throughout the genome (Srinivasa Rao et al. 2023). The selection procedure depends on information from training people, including phenotypic and genotypic characteristics (Figure 1).

Following an intensive training process, these models produce forecasts of breeding or phenotypic characteristics for characteristics of a target population based only on genotypic information (Dessy et al. 2023). Before implementing selection, it is essential to assess the effectiveness of model predictions using Cross-Validation (CV) techniques (Allgaier & Prys 2024). Further details on CV methods are found in the next section (Surendar et al. 2024). Evaluating the efficacy of forecasting algorithms and comparing distinct sets of statistical ML algorithms is crucial in GP (Camgözlü & Kutlu 2023). This evaluation involves considering different circumstances, such as incorporating characteristics, known central genetics, marker-trait associations, Genotype Three Environments (G3E) relationship, and other omics information, including transcriptomics, metabolomics, and proteomics (Ansarifard et al. 2020).

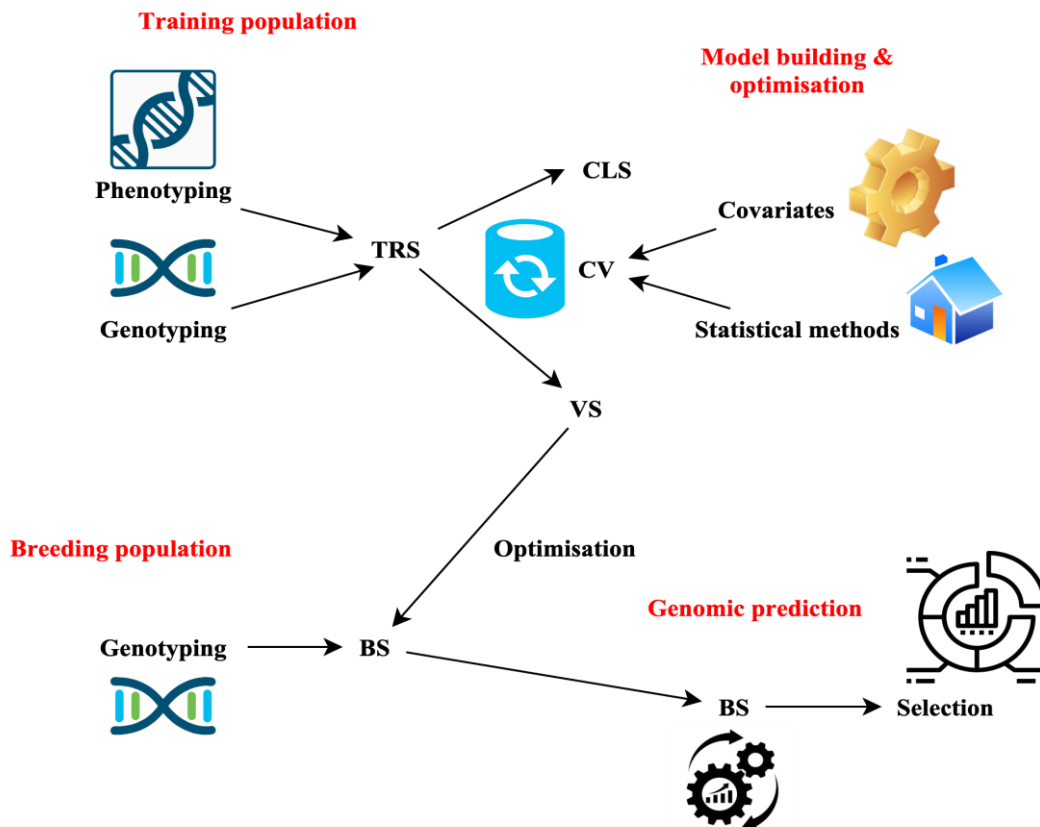


Figure 1. GP characteristics

GP considers the breeding characteristics of the parental standard and variation of Mendelian sampling to determine an offspring's Genomic Estimated Breeding Figures. This method can be utilized for two purposes: (1) to quickly select desirable traits in early generations by forecasting additive impacts and (2) to choose lines in later phases of breeding by forecasting the genotypic amounts of people, considering both additive and non-additive impacts that determine the final economic worth of the queues. Several variables influence GP and can significantly diminish its precision (Figure 2).

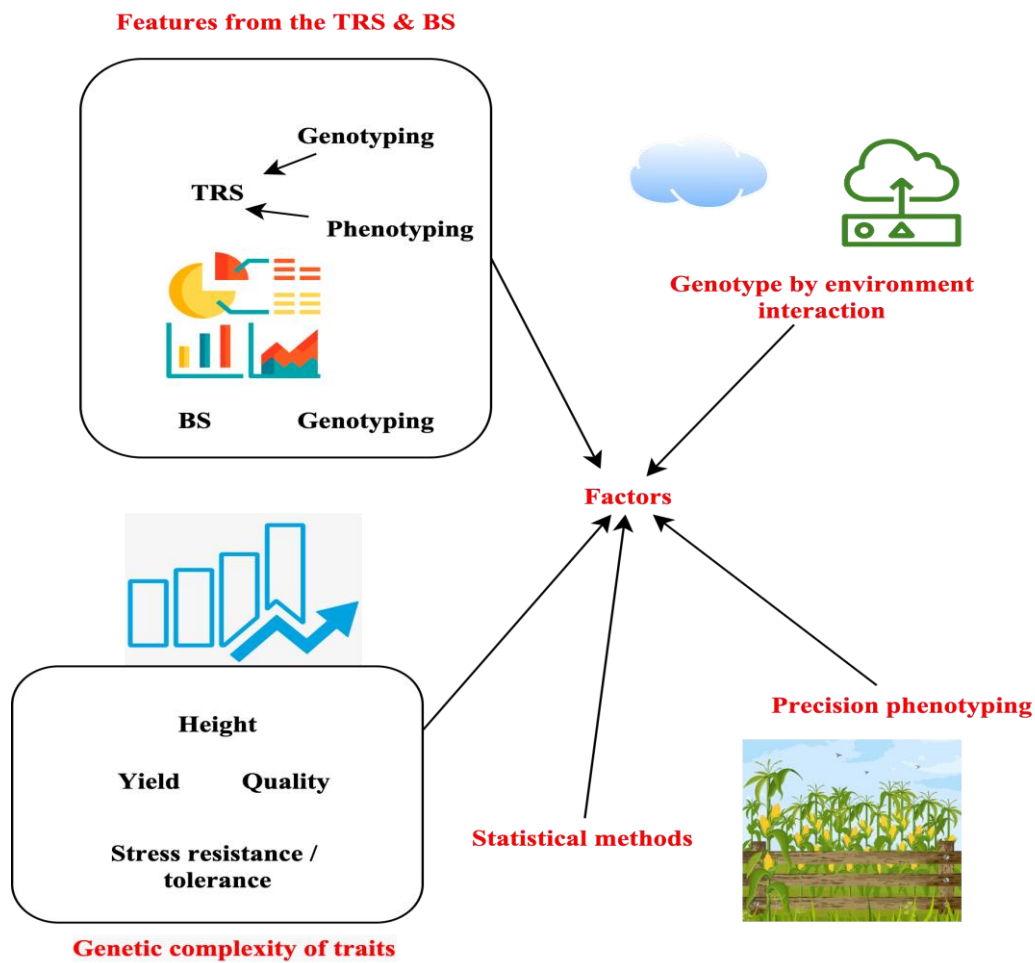


Figure 2. Influencing variables for GP

If not adequately dealt with, these issues can impede the efficient application of GP in PB projects. When optimizing the training populace, it is crucial to focus on essential parameters such as population count, genetic variation, and the genetic connection with the breeding community. Additional variables that make a more brutal genetic prediction in PB include the degree of linkage disequilibrium among indicators in both the development and evaluation of populations,

the genetic complexities and inheritance of target characteristics, the accuracy and reliability of phenotyping, the use of mathematical ML designs, the relationship between genotype and environment, and other non-additive variables (Guo & Li 2023).

Proposed big data-based cross-breeding method

The combination of phenomic and genomic information can revolutionize PB. Incorporating field-based single-plant phenotyping in the early generations has received limited attention in research. PB can expedite selecting and improving populations early by utilizing high-throughput analysis of individual seedlings. This approach can accelerate genetic advancements and enhance the efficiency of breeding resources. The volume of data in PB is increasing due to several factors. Hundreds of potential varieties are evaluated and thoroughly described each season, combining a wealth of phenotyping data gathered from many sources. Big data is generated using molecular indicators. Professional breeders are faced with extensive datasets.

Cross-breeding method: Figure 3 shows the architecture of the big data-based CB model. To enhance the accuracy of predicting the relationship between an organism's genetic makeup and its observable traits, it is necessary to establish comprehensive systems that can handle large amounts of data designed explicitly for PB purposes. One effectively tackles the task of modeling and condensing data for decision-making purposes under time pressure. To tackle these problems, biometrics specialists have developed a software pipeline that integrates data and algorithms to extract relevant details for the breeders. The biometrics pipeline encompasses characteristics of the design and evaluation of phenotyping studies, the transfer of polymorphisms from mothers to offspring, the integration of genotypes and traits in tracing and designs, and utilizing genomic forecasting systems. The biometrics pipeline assisted in addressing this issue by providing tools to gain deeper insights into the crossing parents, enabling an understanding of how discrimination occurs within a community. This knowledge is crucial for identifying and developing the most effective crossing parents and ultimately selecting the best varieties by combining the desired traits and genes into a single variety. By utilizing diverse germplasm, extensive genomic, phenomic, and environmental data, and integrative evaluation, the research can pinpoint causal loci and predict traits for breeding accurately.

Crop breeding management: Combining ability evaluation is a valuable tool for breeders to assess the strengths and weaknesses of various mixtures and parent plants in the early breeding stages. This allows breeders to narrow down the selection of materials, conserve period in the breeding process, and enhance overall breeding effectiveness. Thus, the research developed a PB data management structure called the gold seed breeding big data analytics-based system

combining ability assessment and implementation. The system operates on a cloud computing infrastructure. It has the potential to enhance flexibility, minimize infrastructure needs, enhance accessibility, and effectively manage extensive data collection. Figure 4 illustrates the operational sequence of the system.

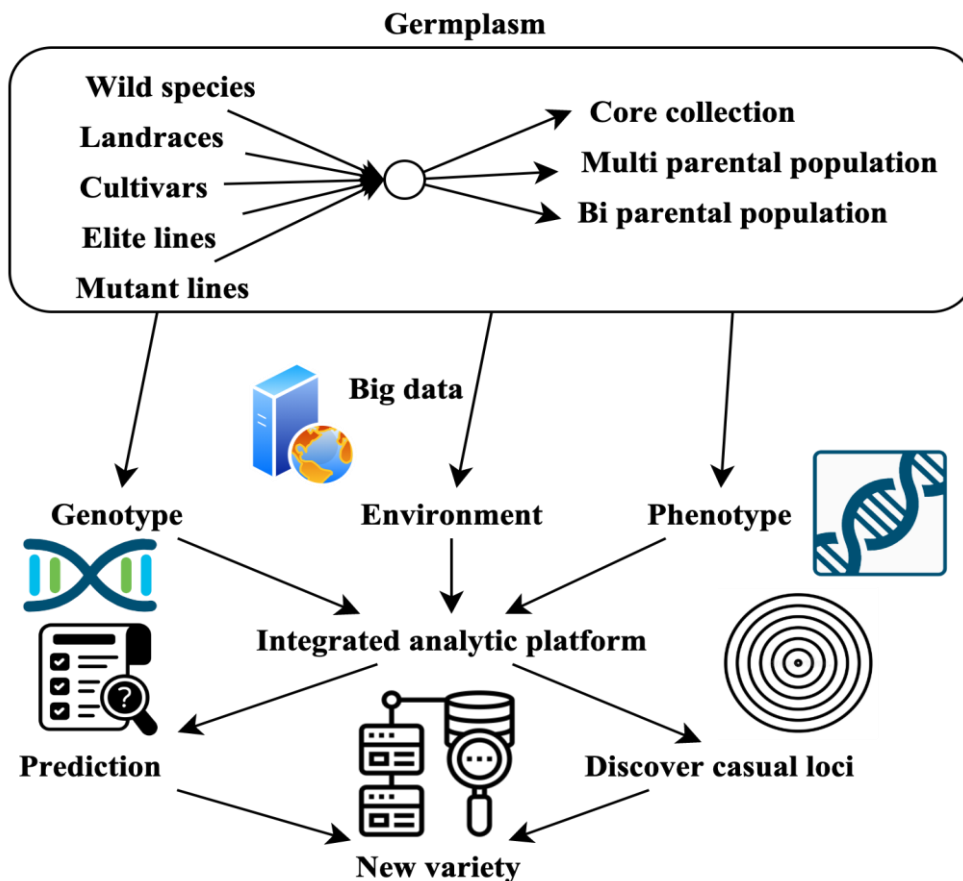


Figure 3. Big data-based cross-breeding model

Each year, breeders carefully choose parents based on their breeding goals. The technique has an efficient retrieval feature for quickly looking for parent breeding resources. The approach incorporates incomplete, complete, and limited diallel crossing procedures, which aid in developing a crossbreeding strategy for combining ability testing. The system automatically detects prior combinations of CB organisms, orthogonal pairings, and reciprocal hybrids. Breeders are offered several trial layout strategies, such as randomized sections and entirely random ones based on the crossbreeding plan. A subordinate plans the planting schedule in the field according to the trial plan devised by the breeder. The employees plant the seeds according to the predetermined planting strategy. Throughout the process of crop development, a subordinate gathers information on the characteristics of the plants. The breeder assesses the

elements and examines the capacity of CB elements to combine in the system with a single action. This characteristic helps breeders discern better parents and offspring based on their achievements in General Combining Abilities (GCA) and Specialized Combined Abilities (SCA). These assessments are utilized in future breeding programs.

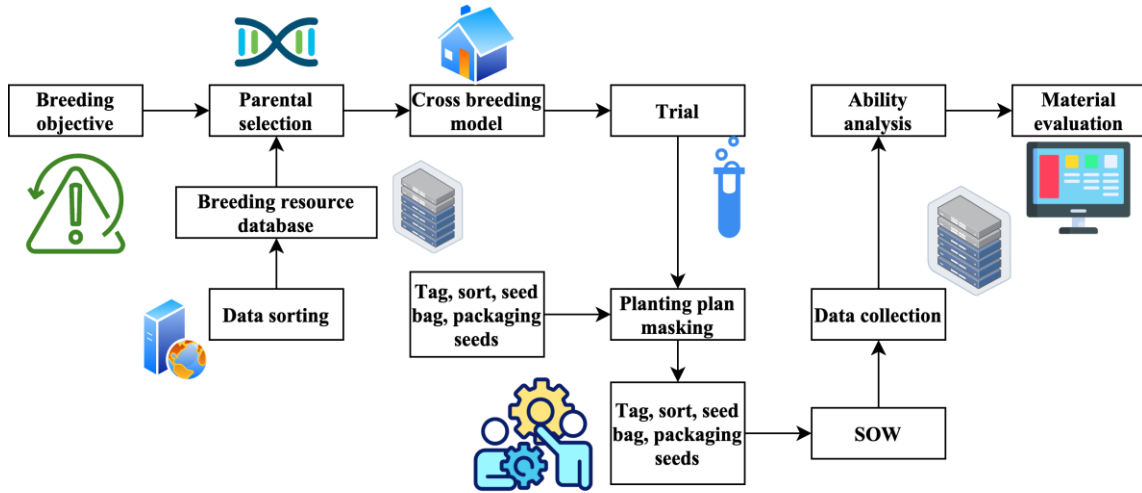


Figure 4. Big data management model

Results and discussion

The accuracy of single-cross forecasting was assessed using Leave-One-Out CV (LOOCV). LOOCV is a specific instance of k-fold CV, where k equals the number of observations (n). The research chose LOOCV because it reduces bias in the predictor by using a more significant number of folds. Five distinct LOOCV situations were examined, each having differing levels of correlation between the training and verification sets for single crossings (Figure 5).

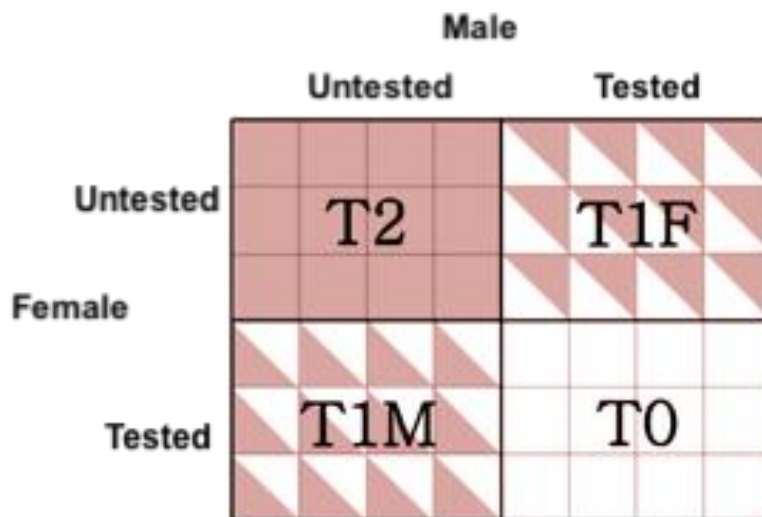


Figure 5. Leave-one-out cross-validation (LOOCV) results

The CV cases were as stated: (1) In T2, both parents of a single cross in the verification set were examined. (2) In T1F, only the female parenting of a single cross in the verification set was examined. (3) In T1M, only the male parent of a single cross in the verification set was examined. (4) In T0, neither of the adults of a single cross in the verification established was tested. (5) In a novel single-cross relatives, all single crosses relating to that family were eliminated from the training set and established the verification set. The traditional LOOCV method was adjusted significantly to provide a consistent training set length for every examined CV situation. The learning set size was limited to 261 in all five situations. The research established the training set dimension at 250 for each of the five CV situations to exclude the influence on sample density. CV is performed in the first four cases by placing every 312 individual crossings into the validation set precisely once, known as LOOCV. During every 312 cycles, a sampling of 250 single crosses was randomly selected from the remainder of single crosses without replacement to create the learning set. The process was iterated 30 times to ensure enough resampling of the learning set, resulting in 9360 reproduced training sets. During every 30 cycles, the forecasts were combined into a unified vector and compared with the phenotypic findings using the method. In situation 5, the CV process was carried out to include every one of the nine single-cross parents in the verification set once, using the LOOCV approach. The process was iterated 30 times by randomly selecting 250 individual crosses without substitution from the experimental set. The precision of the forecasting was assessed exclusively for the six most significant families due to the limited size of the three groups (f7, f8, and f9), which hindered the appropriate estimation of correlation scores.

The single-cross best linear unbiased predictors obtained from modeling (1) were considered the measured single-cross efficiency and used for verification. The forecasting precision was quantified using Pearson's correlation factor, which measures the relationship between measured and anticipated single-cross efficiency. This value was split by the square root of the broad-sense heredity on an entry-mean foundation. The average forecasting efficiency over the 30 trials was provided. The Standard Errors (Ses) of the forecasting efficiency were computed using the bootstrap approach, included in the R package called "boot." During each of the 30 cycles, the forecasted and discovered variables were recreated 200 times with substitution. The resultant range of 200 correlation coefficient predictions was utilized to determine the bootstrap SE. The average SE was recorded over the 30 iterations.

The research initially assessed the predictive accuracy of T2, T1F, T1M, and T0 situations in the entire population using LOOCV. Saty Green (SG) and Plant Height (PH) had higher prediction accuracies for all cases than Grain Yield (GY) (Figure 6). The highest forecasting precision was seen for T2, with T1F, T1M, and T0 following in descending order. The four

techniques showed comparable levels of accuracy when applied to both the T2 and T1F situations. Techniques 1a and 1b exhibited better results than methods 2a and 2b in forecasting single-cross results for the T1M and T0 situations. Applying the proposed modeling resulted in slight improvements in the precision of predicting GY and Plant Height (PH), with the highest improvement observed in the T0 situation.

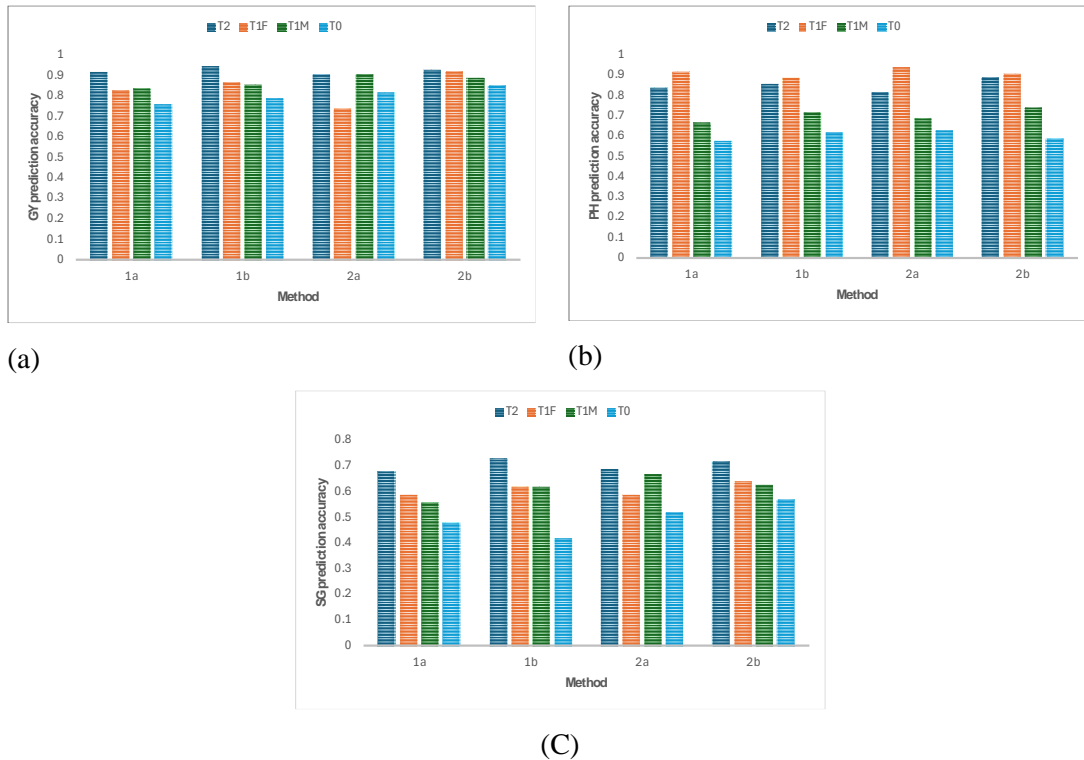


Figure 6. Prediction result analysis

Conclusions: Thanks to significant developments in breeding methods, the research can now efficiently and inexpensively analyze vast quantities of big genetic data obtained from individual plant samples. PB has successfully implemented these enhancements and developed a range of cultivars with increased productivity and improved characteristics to ensure the safety of the food the research consumes daily. The current rate of genetic improvement needs to be enhanced to fulfill the projected food requirements, even with the utilization of sophisticated breeding techniques and platforms. Hence, PB must ascertain a more streamlined approach to enhance genetic advancement and develop resilient varieties to climate change. The article suggests that genomic forecasting, forecasting breeding, and utilizing big data from genomics and phenomes are all possible methods to accelerate the rate of genetic improvement. To fully use the benefits of new genetic advancements, it is imperative to consistently generate large amounts of genomic

information, including several types of biological information, and analyze this multidimensional information.

References

- Allgaier J, Pryss R (2024) Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Mach Learn Knowl Extr* 6(2), 1378-1388.
- Ansarifard I, Mostafavi K, Khosroshahli M, et al. (2020) A study on genotype-environment interaction based on GGE biplot graphical method in sunflower genotypes (*Helianthus annuus* L.). *Food Sci Nutr* 8(7), 3327-3334.
- Begna T (2021) Conventional breeding methods are widely used to improve self-pollinated crops. *Int J Res* 7(1), 1-16.
- Camgözlü Y, Kutlu Y (2023) Leaf image classification based on pre-trained convolutional neural network models. *Nat Eng Sci* 8(3), 214-232.
- Dessy A, Ratna D, Leni S, et al. (2023) Using distance measure to perform optimal mapping with the K-medoids method on medicinal plants, aromatics, and spices export. *J Wirel Mob Netw Ubiquitous Comput Dependable Appl* 14(3), 103-111.
- Ghotbaldini H, Mohammadabadi M, Nezamabadi-pour H, et al. (2019) Predicting breeding value of body weight at 6-month age using artificial neural networks in Kermani sheep breed. *Acta Sci - Anim Sci* 41, e45282.
- Guo T, Li X (2023) Machine learning predicts phenotypes from genotypes and environments. *Curr Opin Biotechnol* 79, e102853.
- Liu J, Li M, Zhang Q, et al. (2020) Exploring the molecular basis of heterosis for plant breeding. *J. Integr Plant Biol* 62(3), 287-298.
- Mohammadabadi M, Kheyroodin H, Afanasenko V, et al. (2024) The role of artificial intelligence in genomics. *Agric Biotechnol J* 16 (2), 195-279.
- Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H (2017) Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. *J Livestock Sci Technol* 5(2), 53-61.
- Radhika A, Masood MS (2022) Crop yield prediction by integrating et-dp dimensionality reduction and ABP-XGBOOST technique. *J Internet Serv Inf Secur* 12(4), 177-196.
- Scott MF, Ladejobi O, Amer S, et al. (2020) Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity* 125(6), 396-416.
- Shivanna KR (2022) Climate change and its impact on biodiversity and human welfare. *Proc Indian Nat Sci Acad* 88(2), 160-171.


- Srinivasa Rao M, Praveen Kumar S, Srinivasa Rao K (2023) Classification of Medical Plants Based on Hybridization of Machine Learning Algorithms. *Indian J Inform Sourc Serv* 13(2), 14-21.
- Surendar A, Veerappan S, Sindhu S, Arvinth N (2024) A Bibliometric Study of Publication-Citations in a Range of Journal Articles. *Indian J Inf Sources Serv* 14(2), 97-103.
- Suyama Y, Hirota SK, Matsuo A, et al. (2022) Complementary combination of multiplex high-throughput DNA sequencing for molecular phylogeny, Hoboken, USA: John Wiley & Sons, Inc 37(1), 171-181.
- Swarup S, Cargill EJ, Crosby K, et al. (2021) Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci* 61(2), 839-852.
- Tsai HY, Janss LL, Andersen JR, et al. (2020) Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Sci Rep* 10(1), 3347.
- Zoran G, Nemanja A, Srđan B (2022) Comparative analysis of old-growth stands Janj and Lom using vegetation indices. *Arch Tech Sci* 2(27), 57-62.

تجزیه و تحلیل داده‌های ژنتیکی بزرگ برای پیش‌بینی ویژگی‌های کراس بردها برای افزایش بازده غذا

پریا ویج 

*نویسنده مسئول: استادیار، گروه علوم کامپیوتر و فناوری اطلاعات، دانشگاه کالینگا، رایپور، هند. آدرس پست الکترونیکی:

ku.priyavij@kalingauniversity.ac.in

پاتیل مانیشا پراشانت 

پژوهشگر، گروه علوم کامپیوتر و فناوری اطلاعات، دانشگاه کالینگا، رایپور، هند. آدرس پست الکترونیکی:

patil.manisha@kalingauniversity.ac.in

تاریخ دریافت: ۱۴۰۳/۰۶/۲۷ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۳/۰۹/۰۵ تاریخ پذیرش: ۱۴۰۳/۰۹/۰۶

چکیده

هدف: اصلاح‌کنندگان نباتات (PB) با بهره‌گیری از پیشرفت‌های علمی و فناوری مدرن، به طور قابل توجهی بازده و کیفیت کشاورزی را بهبود بخشیده‌اند. هزینه‌ها کاهش یافته و فرآیند PB به دلیل توسعه ابزارهای ژنومی و توالی‌یابی، به ویژه از زمان پروژه ژنوم انسانی، سریع شده است. پرداختن به مسائل جهانی مربوط به منابع آب و امنیت غذایی نیازمند این پیشرفت است. فنوتیپ با کارایی بالا، کشاورزی دقیق و پیش‌بینی محصول همگی با ادغام فناوری‌های پیشرفته مانند سیستم‌های حسگر، تصاویر ماهواره‌ای، ربات‌ها، تجزیه و تحلیل داده‌های بزرگ و ژنومیک بهبود یافته‌اند. این پیشرفت‌ها به رشد کشاورزی دیجیتال کمک می‌کند، که پتانسیل تغییر PB را با اتخاذ رویکردی بین‌رشته‌ای تر دارد. برای بررسی روشی که با آن پیشرفت‌های جدید در کشاورزی دیجیتال، ژنومیک و فن‌آوری‌های حسگر، اصلاح نباتات را تغییر می‌دهند، کیفیت و بهره‌وری محصول را افزایش می‌دهند و با مسائل جهانی مدیریت منابع آب و امنیت غذایی مقابله می‌کنند.

نتایج: اصلاح نباتات به دلیل ترکیبی از ابزارهای ژنتیکی، تکنیک‌های توالی‌یابی و فناوری‌های کشاورزی معاصر سریعتر و کم هزینه‌تر شده است. کشاورزی دقیق با استفاده از فناوری‌هایی مانند روباتیک، تجزیه و تحلیل داده‌های بزرگ و عکاسی ماهواره‌ای، فنوتیپ‌سازی و شناسایی محصولات با کارایی بالا را بسیار افزایش داده است. این پیشرفت‌ها به ایجاد روش‌های کشاورزی پایدار و مؤثرتر کمک می‌کند.

نتیجه‌گیری: یک رویکرد نوآورانه برای بهبود محصول با ادغام مداوم فناوری‌های چند رشته‌ای در اصلاح نباتات در حال توسعه است. پیش بینی می‌شود که ژنومیک پیشرفته و کشاورزی دیجیتالی ظرفیت‌های پرورش دهندگان نباتات را بهبود بخشد و به آنها اجازه دهد تا با مشکلات فزاینده امنیت غذا و آب در دنیایی که روز به روز به هم پیوسته‌تر می‌شود، مقابله کنند.

واژه‌های کلیدی: اصلاح نژاد متقابل، پیش بینی، عملکرد غذا، کلان داده

نوع مقاله: مروری.

استناد: ویج پریا، پراشانت پاتیل مانیشا (۱۴۰۳) تجزیه و تحلیل داده‌های ژنتیکی بزرگ برای پیش‌بینی ویژگی‌های کراس‌بردها برای افزایش بازده غذا. *مجله بیوتکنولوژی کشاورزی*، ۱۶(۴)، ۲۳۷-۲۵۰.

Publisher: Faculty of Agriculture and Technology Institute of Plant
Production, Shahid Bahonar University of Kerman-Iranian
Biotechnology Society.



© the authors