ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Integrating NLP with Climate and Genomic Data for Climate-Resilient Crop Breeding

Manya Sinha[1*], Himanshu Maurya[2] and Goldi Soni[3]

[1]Amity University Chhattisgarh, India. E-mail: manyasinha50@gmail.com
[2]Amity University Chhattisgarh, India. E-mail: himanshumaurya181905@gmail.com
[3]Assistant Professor, Amity University Chhattisgarh, India. E-mail: gsoni@rpr.amity.edu

Article Info

Volume 5, Issue 1, May 2025
Received : 19 January 2025
Accepted : 29 April 2025
Published : 25 May 2025
doi: 10.51483/IJDSBDA.5.1.2025.73-79

Abstract

Modern crop breeding is being revolutionized by Artificial Intelligence (AI), especially Natural Language Processing (NLP), which makes it possible to analyze unstructured data like reports, patents, and scientific literature efficiently. Big data issues impede advances in understanding the intricate biological processes determining agricultural attributes, despite advancements in phenomics, enviromics, and other "omics" methods. By automating literature mining, detecting gene-trait connections, and combining knowledge from multi-omics datasets, NLP tackles these issues. By improving accuracy and facilitating data integration, it improves high-throughput phenotyping, genotyping, and enviromics. Researchers can enhance breeding strategies for climate-resilient crops and speed up gene identification by fusing NLP with "omics" techniques.

Keywords: *Artificial intelligence, Natural language processing, Crop breeding, Genomics, Phenomics, Envirotyping, Big data*

1. Introduction

In contemporary agriculture, Artificial Intelligence (AI) has emerged as a key component of innovation, especially in the development of crop breeding techniques. Natural Language Processing (NLP) has become a vital technique in Artificial Intelligence (AI) for processing, evaluating, and synthesizing vast amounts of unstructured textual data, including field reports, patents, and scientific literature. In crop science, where generating climate-resilient crops requires an understanding of the intricate interactions between genotype, phenotype, and environmental factors, this expertise is crucial.

Transformer-based architectures (e.g., BERT, SciBERT, GPT) and other contemporary NLP models allow for the extraction of useful information from interdisciplinary datasets. These models play a vital role in

* *Corresponding author: Manya Sinha, Amity University Chhattisgarh, India. E-mail: manyasinha50@gmail.com*

automating processes like relation extraction, which maps interactions between entities, and Named Entity Recognition (NER), which identifies important genes, traits, and stressors. Breeders can gain structured, useful insights by using topic modeling approaches (such as LDA or BERTopic) to better arrange research into subject areas like drought resistance or salinity tolerance.

Through an examination of current methodologies, challenges, and opportunities, we highlight how NLP models are reshaping crop science and opening the door for effective and climate-resilient breeding programs. This paper focuses on how NLP models—from rule-based approaches to advanced transformer models—can improve high-throughput phenotyping, genotyping, and enviromics. By utilizing these models, researchers can identify important genes and traits more quickly, automate data interpretation, and more effectively link genotype to phenotype.

2. Literature Review

The integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) into edit breeding inquire about has picked up critical footing, as illustrated by a few key considers. This area surveys the application of NLP within the setting of high-throughput phenotyping, genotyping, and Environics, with appropriate references to inquire about papers that highlight its transformative potential.

2.1. NLP in Literature Mining for Crop Breeding

Houle *et al.* (2019) discussed the use of NLP in automating the extraction of phenotypic traits and gene annotations from textual datasets, streamlining the knowledge discovery process. Similarly, Tripathi *et al.* (2021) utilized machine learning-based NLP models for Named Entity Recognition (NER) to identify stress-related genes in large-scale genomic databases.

2.2. High-Throughput Phenotyping and Environics Using NLP

High-throughput phenotyping involves the rapid measurement of plant traits under varying environmental conditions, often generating large volumes of textual metadata. Yang *et al.* (2020) demonstrated the application of NLP in processing such metadata to link phenotypic traits with environmental factors.

Transformer-based models like T5 and PEGASUS have proven effective in summarizing experimental results and synthesizing insights from field reports. Qiao *et al.* (2022) used these models to extract actionable insights from environmental datasets, facilitating faster decision-making in climate-resilient breeding programs.

3. Methodology

3.1. Data Sources

Chronicled and anticipated climate datasets from sources like NASA EarthData and Copernicus Climate Alter Benefit. These datasets incorporate factors such as temperature, precipitation, mugginess, soil dampness, and extraordinary climate occasions (Allard, 2014).

3.2. NLP for Literature Mining

Consequently extricate basic substances such as qualities, phenotypes, and stretch variables (e.g., dry season, warm) from huge volumes of logical writings. State-of-the-art models like BERT and SciBERT can be fine-tuned for agrarian wording (Meyer *et al.*, 2021).

3.3. Climate and Genomic Data Integration

Develop interconnected graphs linking genes, traits, and climate stressors. For instance, a knowledge graph can illustrate how drought-tolerance genes like DREB2 impact yield in regions with declining rainfall (Meyer *et al.*, 2021).

4. Artificial Intelligence in Crop Breeding

Crop breeding has been revolutionized by the use of Artificial Intelligence (AI), which offers sophisticated

computer tools to tackle difficult agricultural problems. Among AI techniques, Natural Language Processing (NLP) plays a pivotal role by enabling the analysis of unstructured textual data, such as scientific literature, patents, and field observations, which are vital for crop improvement programs (Moran and Smith, 1918).

NLP makes it easier to glean valuable insights from the constantly expanding corpus of multimaps and agricultural research data. Breeders and researchers can find patterns, gene-trait correlations, and environmental stress interactions by using Natural Language Processing (NLP) techniques to process unstructured textual data (Figure 1).

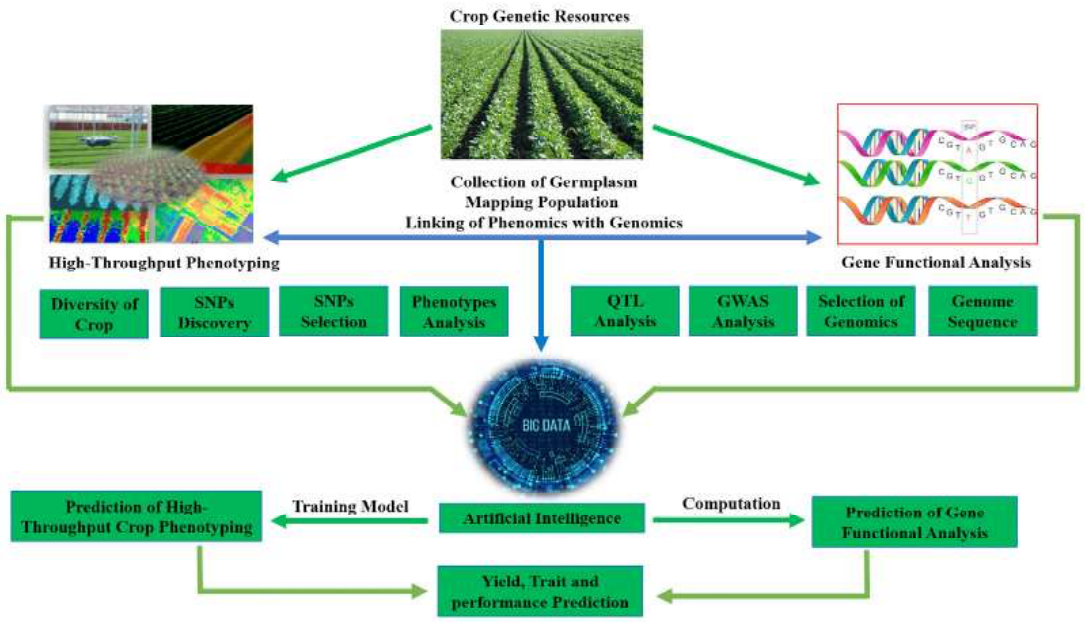

Figure 1: Artificial Intelligence Used as a Powerful Tool for the Prediction of High-Throughput Crop Phenotyping and Gene Functional Analysis in Modern Crop Breeding

Source: Int. J. Mol. Sci. 2022, 23(19), 11156; https://doi.org/10.3390/ijms231911156

5. AI Technologies Overcoming the Phenomics Bottlenecks

Precision breeding is now possible thanks to the tremendous advancements in plant phenomics in recent years. The development and accessibility of new technologies that enable high-throughput phenotyping of intricate plant characteristics are responsible for this advancement (Meuwissen *et al.*, 2001). The application of AI in several scientific domains has skyrocketed in recent years. Deep learning, machine learning, and computer vision are examples of AI features that have been successfully incorporated into non-invasive imaging techniques. This integration is gradually increasing the effectiveness of data collection and analysis through the use of ML for robust picture analysis. AI has also facilitated the creation of tools and software for field phenotyping data administration and collection. Three essential aspects of phenomic data management involve the use of AI: database management to facilitate the sharing of information and resources; model development to comprehend genotype-phenotype relationships with environmental interactions; and algorithms and programs to transform sensory data into phenotypic information. Screening plants for desirable traits (such grain size, abiotic stress tolerance, product quality, or yield potential) requires experiments that involve repeated trials in a variety of settings (taking into account the statistical need for an unbiased estimation). Much phenotyping discussion has focused on assessing individual plants in controlled settings, although controlled environments do not adequately depict plant development in open-air situations (Meuwissen *et al.*, 2001).

Here is a graph comparing the performance of manual methods versus AI methods in addressing phenomics bottlenecks across key areas. It shows how AI significantly improves performance in data processing, accuracy in trait identification, knowledge extraction, and data integration efficiency (Figure 2).

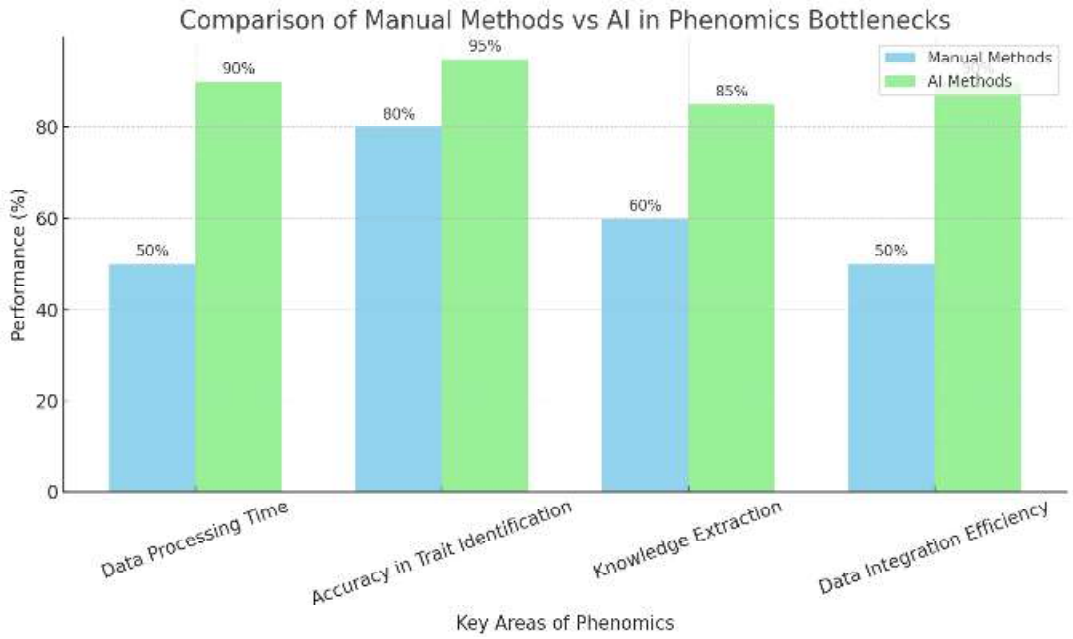

Figure 2: Comparison of Manual Methods vs AI in Phenomics Bottlenecks

6. Natural Language Processing Models

6.1. Rule-Based Models

Based on established linguistic patterns and norms, including regular expressions, tokenization, and grammatical rules.

Use Cases: Named Entity Recognition, Sentiment Analysis (Zargar *et al.*, 2015).

6.2. Statistical Models

Use probability and statistics to process text data.

Use Case: N-Grams, Hidden Markow Models, Latent Dirichlet Allocation.

7. Training of the NLP Model

There are several crucial processes involved in training an NLP model for crop breeding applications. The first step is to precisely characterize the problem, which could be anything from summarizing scientific literature to automating Named Entity Recognition (NER) to identify genes, characteristics, and environmental factors to extracting correlations between genotype, phenotype, and environment. Data preparation is crucial after determining the goal. Using tools like spaCy or Hugging Face tokenizers, domain-specific datasets are gathered from sources like PubAg or AgBioData, the text is annotated to identify things like genes and traits, the data is cleaned to eliminate unnecessary information, and the data is tokenized into words or sub words.

A model is chosen once the data is prepared. For tasks like NER, relation extraction, or summarization, pre-trained models like BERT, SciBERT, or GPT are great options. Performance can be enhanced by fine-tuning these models on domain-specific datasets. As an alternative, unique structures such as Transformers or LSTMs may be created for more specialized purposes. The model is evaluated on a different dataset while learning from a training set. To fine-tune model weights over multiple epochs, optimizers like as Adam or Adam W and task-specific loss functions, like cross-entropy for classification, are used.

Metrics like precision, recall, and F1-score for classification tasks or BLEU and ROUGE scores for summarization are calculated as part of the critical evaluation of the trained model. The generalizability of the model is evaluated using a different test set. Hyper parameters such as batch size and learning rate are adjusted if needed to further maximize performance. For example, a SciBERT model trained for NER in crop

breeding may begin with a batch size of 16 and a learning rate of 5e-5. The accuracy of entity recognition would be assessed using the model's F1-score.

The trained and optimized model is then saved and used for practical purposes, such detecting climate-resilient features, automating literature mining, or connecting multiomics data. An accurate and scalable NLP model that is suited to the requirements of crop breeding research is guaranteed by this methodical procedure (Amit, 2018).

Artificial intelligence is being used to integrate and manage genomic, phenomic, and environmental data in order to improve crop breeding. Crop phenotypic data are gathered from both indoor and outdoor settings, and AI technology is used to integrate the phenotypic, genotypic, and environmental data. The AI-assisted breeding system will simulate and validate the chosen cultivars by mathematical modeling, logical inference, and decision-making, determining if they are appropriate for cultivation in all main habitats or just a few (Figure 3).

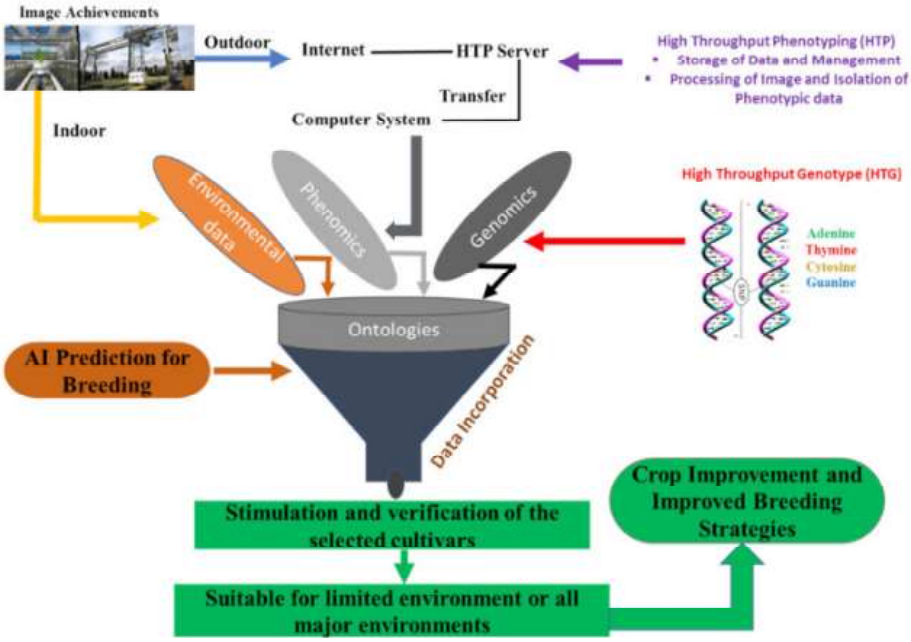

Figure 3: Simulation and Verification of the Selected Cultivars

Source: Int. J. Mol. Sci. 2022, 23(19), 11156; https://doi.org/10.3390/ijms231911156

8. Case Studies

8.1. Drought – Tolerant Maize

NLP was used to extract drought-tolerance genes from a corpus of agricultural research, including members of the DREB family. The areas in Sub-Saharan Africa most vulnerable to drought were determined by climate data. Field testing was led by the integration of environmental and genetic data.

30% less time was spent developing drought-resistant maize cultivars because to an accelerated identification process.

8.2. Heat-Resilient Wheat

From the literature, NLP found heat-responsive genes, especially those belonging to the Heat Shock Protein (HSP) family. These were mapped to South Asian locations that are expected to see a lot of heat waves. The effectiveness of these genes under conditions of simulated heat stress was evaluated using predictive models.

Breeding efforts for heat-tolerant wheat lines were prioritized, and new varieties were tested at various field locations.

8.3. Salt-Tolerant Rice

Salinity-related genes (SKC1 and HKT families) were found using Natural Language Processing (NLP) and relation extraction, and they were connected to salinization zones brought on by climate change. Target areas in coastal Africa and South Asia were identified by geospatial analysis (Cortés and López-Hernández, 2021).

High-performing, salt-tolerant rice cultivars were made available for these areas more easily thanks to the integration framework.

9. Conclusion

The combination of Natural Language Processing (NLP) with genomic and climatic datasets offers a revolutionary opportunity for contemporary crop breeding, particularly in the realm of developing resilience to climate change. This study illustrates how NLP acts as a connector between unstructured textual information and practical agricultural insights, facilitating the extraction of gene-trait-environment correlations from extensive scientific literature and reports. Utilizing AI-driven models such as BERT, SciBERT, and GPT, researchers are able to automate the process of literature mining, improve entity recognition, and create knowledge graphs that expedite the identification of climate-resilient traits.

Furthermore, the collaboration between high-throughput phenotyping, enviromics, and Natural Language Processing (NLP) has created new opportunities for accurate, data-driven decision-making in crop enhancement initiatives. Empirical case studies involving maize, wheat, and rice demonstrate the practical implementation and effectiveness of these methodologies. Although obstacles such as data complexity and integration persist, continuous progress in artificial intelligence technologies and model training methods is enhancing system scalability, precision, and flexibility.

Looking ahead, the integration of NLP with multi-omics and climate data will not only transform traditional breeding methods but also equip breeders and researchers with the necessary tools to quickly address environmental challenges. This development represents a crucial advancement toward achieving global food security amid rising climate variability.

10. Future Prospects of AI Breeding

Discussions about the significance of AI have grown significantly in recent years, sparking discussions about its global applications. To benefit from the digital revolution, plant breeding needs to be upgraded. To succeed in their future endeavors, researchers and breeders must compare computer-generated recommendations with farmers' needs (Talaviya *et al.*, 2020). The development and effective application of AI technology has led to increased profitability and rapid economic growth in a number of industries worldwide, including agriculture breeding.

Additionally, AI will concentrate on creating innovative, human-centered methods and evaluating how robotic technology may be used to various organizations and industries globally. AI will also change how businesses grow and compete globally by bringing new manufacturing ideas that will lead to increased profitability (Supriya and Deepa, 2020). The majority of businesses worldwide will need to be more active in the development of different AI techniques, such as placing human elements at the core, in order to fully benefit from such opportunities. Additionally, they will focus on creating a range of ethical and moral AI devices that will produce favorable results and allow people to carry out duties they are accustomed to.

The global agricultural breeding industry will benefit from the development of various AI systems by assuming the availability of symbolic structures, such as the existence of knowledge and the capacity for reasoning. Furthermore, there will be worries about cultural and political change as AI reaches intellect levels on par with or higher than those of humans (Priya and Ramesh, 2020).

AI applications will become more broadly available as farmers and breeders can use portable devices, drones, and agriculture-equipment platforms to feed data into cloud-based AI systems. Breeding is still a difficult, expensive, and time-consuming operation, and while the phenomics and genomics data acquired using ML and DL are accurate, they are not sufficient to fully rely on the technology to speed it up. Epigenomics, transcriptomics, proteomics, metabolomics, and phenomics still offer little insight into genomes.

References

Allard, R. (2014). Plant Breeding. *Encyclopedia Britannica,* Chicago, IL, USA. [Google Scholar].

Amit, K. (2018). Artificial Intelligence and Soft Computing: Behavioral and Cognitive Modeling of the Human Brain. CRC Press, Boca Raton, FL, USA. [Google Scholar].

Cortés, A.J. and López-Hernández, F. (2021). Harnessing Crop Wild Diversity for Climate Change Adaptation. *Genes*, 12, 783. doi: 10.3390/genes12050783. [DOI] [PMC free article] [PubMed] [Google Scholar].

Meuwissen, T.H., Hayes, B.J. and Goddard, M.J.G. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157, 1819-1829. doi: 10.1093/genetics/157.4.1819. [DOI] [PMC free article] [PubMed] [Google Scholar].

Meyer, R.S., DuVal, A.E. and Jensen, H.R. (2012). Patterns and Processes in Crop Domestication: An Historical Review and Quantitative Analysis of 203 Global Food Crops. *New Phytol*, 196, 29-48. doi: 10.1111/j.1469-8137.2012.04253.x. [DOI] [PubMed] [Google Scholar].

Moran, P. and Smith, C. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.*, 52, 399-438. [Google Scholar].

Priya, R. and Ramesh, D.J.S.C.I. (2020). ML Based Sustainable Precision Agriculture: A Future Generation Perspective. *Sustain. Comput. Inform.*, 28, 100439. doi: 10.1016/j.suscom.2020.100439. [DOI] [Google Scholar].

Supriya, M. and Deepa, A. (2020). Machine Learning Approach on Healthcare Big Data: A Review. *Big Data Inf. Anal.*, 5, 58-75. doi: 10.3934/bdia.2020005. [DOI] [Google Scholar].

Talaviya, T., Shah, D., Patel, N., Yagnik, H. and Shah, M. (2020). Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides and Herbicides. *Artif. Intell. Agric.*, 4, 58-73. doi: 10.1016/j.aiia.2020.04.002. [DOI] [Google Scholar].

Zargar, S.M., Raatz, B., Sonah, H., Bhat, J.A., Dar, Z.A., Agrawal, G.K. and Rakwal, R. (2015). Recent Advances in Molecular Marker Techniques: Insight into QTL Mapping, GWAS and Genomic Selection in Plants. *J. Crop Sci. Biotechnol.*, 18,293-308. doi: 10.1007/s12892-015-0037-5. [DOI] [Google Scholar].

Cite this article as: Manya Sinha, Himanshu Maurya and Goldi Soni (2025). Integrating NLP with Climate and Genomic Data for Climate-Resilient Crop Breeding. *International Journal of Data Science and Big Data Analytics*, 5(1), 73-79. doi: 10.51483/IJDSBDA.5.1.2025.73-79.