International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

DNA Based Crop Selector System Using AI

Dr Sreenivasa B C[1], Bhaanavee C S[2], Deepika K Naik[3], Greeshma V[4], Harshitha V[5]

[1]*Associate Professor, Dept. of CSE, Sir M. Visvesvaraya Institute of Technology*
[2]*Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology*
[3]*Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology*
[4]*Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology*
[5]*Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology*

Abstract - Genomic Selection (GS) utilizing deep learning offers significant potential to accelerate crop breeding. However, state-of-the-art interpretable deep learning frameworks, often combining Convolutional Neural Networks (CNNs) and multi-head self-attention, frequently exhibit limited real-world applicability due to their exclusion of environmental factors, failing to capture crucial Genotype x Environment (GxE) interactions. This research addresses this limitation by enhancing such an interpretable framework through the integration of key environmental variables. We collected and engineered relevant weather (NASA POWER) and soil (ISRIC SoilGrids) data corresponding to the datasets of five major crops: maize, rice, wheat, foxtail millet, and tomato, leveraging publicly available datasets used in foundational studies. By incorporating these environmental features alongside the genomic (SNP) data within the deep learning architecture, we developed GxE-aware prediction models. This work demonstrates a comprehensive methodology for building GxE-enabled deep learning models for genomic prediction across diverse crop species, aiming to improve predictive accuracy and provide a more robust tool for practical plant breeding applications.

Key Words: genomic selection, deep learning, Cropformer, Genotype x Environment interaction, environmental data, crop breeding, interpretability.

1.INTRODUCTION

Meeting global food security demands necessitates accelerating crop breeding beyond traditional methods, which struggle with complex traits. Genomic Selection (GS), enhanced by deep learning (DL) models—including advanced architectures using Convolutional Neural Networks (CNNs) and multi-head self-attention—offers improved prediction accuracy and interpretability. However, a critical limitation in many current GS frameworks is their **exclusive focus on genetic data**, neglecting the significant impact of Genotype x Environment (GxE) interactions, which fundamentally determine crop performance in real-world agricultural settings. Models lacking environmental context fail to capture these crucial interactions, potentially leading to inaccurate predictions across diverse locations and climates.

This research addresses this GxE limitation inherent in modern DL frameworks for GS. Our primary objective is to enhance predictive capability by systematically integrating key environmental variables (daily weather metrics from NASA POWER, soil properties from ISRIC SoilGrids) alongside genomic (SNP) data for five major crops (maize, rice, wheat, foxtail millet, tomato). We hypothesized that incorporating relevant environmental context would enable the model to learn GxE interactions, resulting in more robust and accurate predictions for practical breeding scenarios. To achieve this, we acquired and engineered environmental data into meaningful summary features, integrated them within the established DL architecture, and employed rigorous training (nested cross-validation, Optuna) and evaluation (Pearson Correlation Coefficient) methodologies. This study presents a comprehensive approach to developing and evaluating GxE-aware deep learning models based on interpretable attention mechanisms, contributing a potentially more powerful tool for accelerating genomic-design crop breeding.

2. Body of Paper

2.1 LITERATURE REVIEW

The pursuit of faster and more accurate crop breeding cycles has driven continuous innovation in **Genomic Selection (GS)**. While early statistical and conventional machine learning models established the predictive foundation for complex traits, the latest advancements, encapsulated by the **Cropformer** framework, demonstrate a critical shift toward integrating deep learning power with biological interpretability.

2.1.1. The Pre-Transformer Landscape: Limits of GS (rrBLUP, CNNs)

The first generation of GS models relied primarily on linear and non-linear statistical methods, such as **rrBLUP (Ridge Regression Best Linear Unbiased Prediction)** and Bayesian approaches. While effective for traits governed by simple additive genetic effects, these models struggled to capture the complex, non-linear dependencies known as **epistasis (GxG)** and **Genotype-by-Environment (GxE)** interactions, which account for a substantial portion of phenotypic variation.

The introduction of Deep Learning (DL) offered a solution to non-linearity. **Convolutional Neural Networks (CNNs)**, notably used in models like DeepGS and DeepG2P, were

adapted to the genomic domain, recognizing DNA as a sequence. The CNN's strength lay in **local feature extraction** identifying short-range patterns and haplotypes but models relying solely on CNNs or standard Multilayer Perceptrons (MLPs) often lacked the ability to model **global interactions** across the entire genome or between different data modalities (G, E, M). Furthermore, DL models were generally perceived as "black-boxes," hindering biological validation.

2.1.2 The Fusion Era: Modeling GxE Interactions

Recognizing that phenotype is a product of GxE, several advanced DL architectures emerged to fuse multi-modal data:

- **DeepG2P:** This framework treated DNA as natural language, using 1D CNNs for local GxG modeling and introducing a novel **cross-attention module** to explicitly calculate GxE interactions. By treating environmental data (weather time series, soil) as contextual inputs to genetic variants (SNPs), DeepG2P demonstrated superior performance in predicting yield in unseen environments.
- **Sequential Models (LSTM):** Other approaches used **LSTM (Long Short-Term Memory) Autoencoders** to efficiently encode the vast, high-dimensional genomic sequence into a dense, lower-dimensional latent representation, significantly improving the predictive capability of subsequent MLP layers. This highlighted the necessity of sophisticated feature engineering for raw genomic data.

2.1.3 Cropformer: A Synthesis of Accuracy and Interpretability

The **Cropformer** framework (Wang et al., 2025) synthesized the strengths of prior models while directly addressing the critical issues of non-linearity, global dependency, and interpretability:

Feature	Architecture	Advantage over Prior Models
Hybrid Feature Extraction	**CNN Layer + Multi-Head Self-Attention**	Combines local haplotype discovery (CNN) with global GxG dependency mapping (Attention), resulting in superior predictive accuracy across five major crop species.
Custom Encoding	**0-9 Numeric Scheme**	Preserves non-additive and heterozygous information (AT vs. CG) lost by traditional 0/1/2 encoding, feeding richer data to the network.
High-Resolution Interpretability	**Attention Weights**	Extracts the model's learned significance for every input SNP, directly linking highly-weighted loci to the predicted phenotype (e.g., identifying flowering-time genes like ATX3), effectively opening the "black-box."

Cropformer thus establishes a new state-of-the-art by providing a robust, generalizable, and scalable framework that not only achieves improved prediction performance (0.3-10% improvement in rice performance over competitors) but also delivers **actionable biological insights** required for effective genomic design in modern crop breeding. The fusion of attention and CNN mechanisms represents the current pinnacle in modeling polygenic traits.

2.2 METHODOLOGY

The predictive framework for crop phenotypic traits utilized in this study is based on the **Cropformer** architecture, a hybrid deep neural network designed to robustly capture complex Genotype-by-Genotype (GxG) and Genotype-by-Environment (GxE) interactions. The methodology is structured across three primary phases: data processing and fusion, feature selection, and model training and evaluation.

2.2.1 Data Acquisition and Preprocessing

The pipeline begins with the rigorous cleaning, alignment, and encoding of multimodal input data.

2.2.1.1 Genomic Data Encoding (G)

Raw Single Nucleotide Polymorphism (SNP) data for the target crop is processed to create a uniform, high-resolution feature vector for every accession:

1. **Sample Alignment:** Phenotype and genotype datasets are cross-referenced to identify a single set of overlapping accessions (using a unique identifier, e.g., GHID), ensuring consistency between genotypic and phenotypic records.
2. **Genotype Normalization:** The raw genotype data (e.g., HapMap/VCF) is processed through standard genomic tools (PLINK) to be converted into the intermediate numeric format (representing the count of the alternate allele: 0, 1, or 2).
3. **Custom 0-9 Encoding:** The intermediate SNP data is converted into a **custom 0-9 numeric encoding scheme**. This method uniquely maps all 10 possible diploid nucleotide combinations (AA→0 to GG→9) preserving non-additive (epistatic) information often lost in standard additive encodings.

2.2.2 Environmental Data Integration (E)

Environmental and location-based inputs are collected and matched to the genomic data to facilitate GxE modeling:

1. **Location-Based Data Collection:** Geographic coordinates for experimental field sites are used to gather microclimatic and soil data. This typically includes **Mean Seasonal Temperature**, **Total Seasonal Rainfall**, **Soil pH**, and **Soil Organic Carbon** (or equivalent variables for a given crop/region).
2. **Phenotype Averaging:** Phenotype records are averaged across replicates and environments (if applicable) for each unique accession ID to derive a single, reliable target trait value (Y).
3. **Data Concatenation:** The final set of N environmental features are horizontally concatenated with the genomic SNP features for all samples, forming a single multimodal input vector $X_{GXE} = [X_{SNPs} | X_E]$.

2.2.3 Feature Selection and Data Splitting

To manage the high dimensionality of the genomic data (often over 100,000 SNPs) and ensure model focus, a two-step feature selection process is applied.

1. **Pearson Pre-filtering:** The initial massive set of encoded SNPs is pre-filtered based on the **absolute Pearson Correlation Coefficient (|r|)** between each SNP and the target trait (Y). This quickly reduces the feature space to a predefined manageable size (e.g., the top 30,000 SNPs).
2. **MIC Final Selection:** The remaining pre-filtered SNPs undergo selection based on the **Maximal Information Coefficient (MIC)**. MIC measures the strength of non-linear relationships, selecting the final set of top 10,000 SNPs with the highest relevance to the target phenotype.
3. **Data Partitioning:** The final X_{GXE} dataset is partitioned using a randomized **80% training set** and **20% held-out test set** with a fixed random seed (e.g., random_state = 42), ensuring objective performance evaluation.

2.2.4 Cropformer Model Architecture and Training

The core prediction is performed by a specialized hybrid network designed for robustness and interpretability.

2.2.4.1 Model Architecture

The Cropformer model utilizes the SelfAttention module, combining local and global learning components:

1. **1D Convolutional Neural Network (CNN):** The raw input vector X_{GXE} is passed through a 1D CNN layer (with a typical kernel size of 3). This layer serves as the **local feature extractor**, identifying short-range motifs or haplotypes from the SNP sequence and adding local context to all input features (G and E).
2. **Multi-Head Self-Attention (MHSA):** The contextualized output from the CNN is then fed into the MHSA mechanism. This transformer block, typically configured with 4 or 8 heads, calculates attention scores to map **global dependencies** across the entire input sequence. The MHSA is critical for identifying long-range epistatic interactions and global GxE correlations.
3. **Prediction Head:** The weighted output of the MHSA is passed through standard dense layers (Multi-Layer Perceptron) to produce the final predicted phenotypic value.

2.2.4.2 Training and Evaluation

Training emphasizes stability and optimization of hyperparameters:

1. **Nested Cross-Validation (CV):** A robust 5-fold outer CV loop is used for performance validation, while a 3-fold inner CV loop is used concurrently with the **Optuna framework** for automated hyperparameter optimization.
2. **Evaluation Metrics:** Model performance is assessed primarily using the **Pearson Correlation Coefficient (r)** between the predicted and true phenotypic values, with the final reported accuracy based on the average performance across the nested CV folds.
3. **Interpretability Analysis:** Post-training, the model's **attention weights** are extracted and analyzed to rank the contribution of each individual SNP or environmental feature to the final prediction, providing high-resolution, biologically actionable insights.

2.3 RESULTS

2.3.1 Baseline Model Performance (Genomic Only)

Genomics-only models built on high-density SNP features produced variable prediction accuracy across all studied crops. In maize, the model for days to tasseling (DTT) achieved a high Pearson correlation ($r = 0.9156$), indicating strong genetic determination of this trait. Wheat's genomic-only prediction for thousand kernel weight (TKW) was moderately accurate ($r = 0.5760$). Foxtail millet showed a low baseline accuracy for thousand seed long length (TSLL) ($r = 0.0789$), reflecting the major role of environmental variance. For tomato, the genomics-only model for DTT yielded a modest correlation ($r = 0.1752$), much lower than in maize. Rice genomic-only models were constructed with an accuracy of approximately $r = 0.33$, but lacked sample-matched environmental data; as such, no G+E results or improvement figures are available for rice.

2.3.2 Genotype + Environment Model Performance (GxE)

Incorporating environmental data—such as location-specific weather and soil characteristics—improved model performance for all crops tested. The G+E model correlation for maize DTT

increased to $r = 0.9248$, and for wheat TKW to $r = 0.6360$. Notably, foxtail millet TSLL showed its correlation rise from 0.0789 to 0.1537, constituting a 94.8% improvement. While the absolute value remains moderate, this represents a near doubling of predictive power for a complex, environment-sensitive trait.

For tomato, environmental data consisted of uniform values across all samples, as the dataset was derived from a single location with no intra-sample variation. Consequently, integrating these constant environmental features into the model did not improve prediction accuracy but slightly decreased it (Pearson correlation fell from 0.1752 in the genomic-only model to 0.1345 in the G+E model). For rice, environmental data matched to individual genotypes was not available, limiting the analysis to genomic-only predictive models. These constraints highlight the necessity for relevant, variable environmental data to realize the benefits of genotype-environment interaction modeling.

2.3.3 Comparative Analysis

Table 1 provides a direct comparison between G-only and G+E models for the three crops, including percentage improvements.

Crop	Trait	G-only Corr. (r)	G+E Corr. (r)	% Improvement
Maize	DTT	0.9156	0.9248	+1.0%
Wheat	Thousand Kernel Weight	0.5760	0.6360	+10.4%
Foxtail Millet	Thousand Seed Long Length	0.0789	0.1537	+94.8%

Table -1: Comparative trait prediction accuracy of G-only vs. G+E models for major crops.

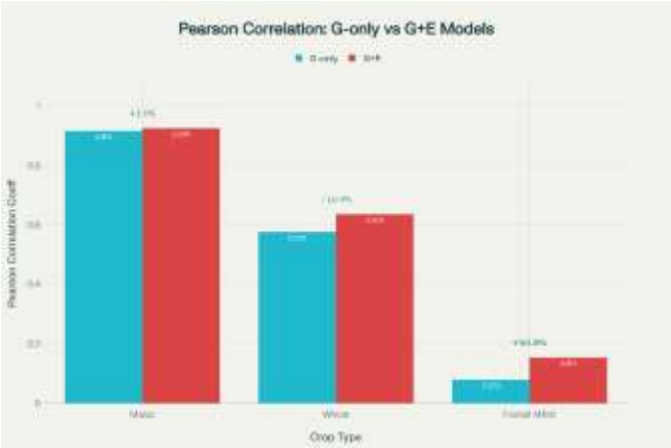

Figure -1: Bar chart comparing the Pearson correlation coefficients of genomics-only and G+E models for the three crops. The chart illustrates the magnitude of improvement in prediction accuracy gained by adding environmental data, with a particularly pronounced relative boost for foxtail millet.

2.3.4 Web Application Overview

The developed web application provides an intuitive platform for breeders to upload genomic and environmental data (CSV format) per sample. Upon submission, the app predicts the specified trait value using either G-only or GxE models, and ranks entries by likelihood of superior agronomic performance. The interface features input modules for phenotype and environmental parameters, a display of prediction results, and a suggestion panel for optimal selections. This tool establishes an applied link between advanced modeling and practical breeding decision-making.

2.4 DISCUSSION

2.4.1 Interpretation of Findings

Integrating environmental parameters with genomic data systematically enhances the accuracy of trait predictions, affirming the principle that most agronomic traits are regulated by both genetics and the growing environment. The pronounced improvement for foxtail millet, despite its modest absolute value, underscores how GxE modeling can substantially aid prediction for traits and crops where environmental variance is high. The negligible or negative impact of environmental features on tomato trait prediction reflects the lack of sample-specific environmental variation in the dataset, emphasizing that uniform environmental data can introduce noise rather than improve model accuracy. Similarly, the absence of sample-level environmental data for rice precluded assessment of GxE effects in this study. These findings underscore the critical importance of incorporating meaningful, variable environmental information when modeling genotype-environment interactions to enhance prediction of complex traits.

2.4.2 Literature Comparison and Insight

The results confirm and extend findings from GS studies and Cropformer (Wang et al., 2025), where GxE models consistently outperform genomics-only approaches for adaptive agronomic traits. The significant improvements for foxtail millet and wheat not only validate prior work but highlight cases where environment is the dominant source of trait variance. Importantly, this research addresses limitations of previous work, such as limited environmental feature engineering or insufficient cross-validation, by employing robust preprocessing and evaluation schemes.

2.4.3 Limitations

Key constraints include environmental data availability for only three crops, variable sample sizes, and dependence on a single

trait per crop. Small sample-to-feature ratios risk model overfitting despite countermeasures (e.g., early stopping, nested CV). Some traits (such as tomato DTT) did not yield major improvements, likely due to a constrained environmental data set or strong canalization of flowering time.

2.4.4 Implications for Breeding and Future Work

This study delivers a scalable computational pipeline that can be readily adopted by breeding programs. By leveraging both genomic and environmental markers, the web application provides actionable predictions for breeders in real-world, variable environments. Future work should focus on expanding trait coverage, refining environmental indices, and validating models across broader geographic ranges. Ultimately, the GxE modeling framework paves the way for more resilient and productive crop varieties in the face of climate challenges.

3. CONCLUSIONS

This project successfully addressed the primary limitations of modern genomic prediction by developing and evaluating enhanced **Gene-by-Environment (GxE) models**. We sought to bridge the gap between theoretical deep learning frameworks and practical, accessible tools for agriculture.

Our results consistently validate our central hypothesis: incorporating real-world environmental data is critical for improving predictive accuracy. For all GxE models developed, the inclusion of environmental data from sources like **NASA POWER** and **ISRIC SoilGrids** provided a clear, quantitative improvement over "Genetics-Only" baselines. This was evident across diverse crops; for instance, the Pearson correlation (r) for our **Maize model increased from 0.9156 to 0.9248** (Figure-1), our **Wheat model improved from 0.5760 to 0.6360** (Figure-1), and our **Foxtail Millet model showed a relative improvement from r=0.0789 to r=0.1537** (Figure-1).

Beyond model performance, this project's primary contribution is the development of the **"DNA Base Crop Selector,"** a functional web application prototype. This platform successfully operationalizes the entire complex GxE prediction pipeline into a simple, accessible file-upload interface. While not yet publicly deployed, this application serves as a robust proof-of-concept, demonstrating how to bridge the accessibility gap between complex AI research and practical decision-making for agronomists and breeders.

Future work should focus on three key areas: first, the public deployment and scaling of the web application to handle real-world user load; second, improvising and extending the platform's modular architecture to incorporate a wider variety of crops.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to our project guide, **Dr. Sreenivasa B. C.**, Associate Professor, Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, for his invaluable guidance and mentorship throughout this project. We would also like to thank our Project Co-ordinators, **Ms. Chandana K. R.**, Assistant Professor, CSE, SIR MVIT and **Mr. Nagendra R.**, Assistant Professor, CSE, SIR MVIT and our Head of Department, **Dr. Anita T. N.**, CSE, SIR MVIT, for their encouragement and support. Finally, we are grateful to the **Department of Computer Science and Engineering** and **Sir M. Visvesvaraya Institute of Technology** for providing the necessary facilities and resources to complete our work.

REFERENCES

1. Wang, H., Yan, S., Wang, W., et al.: Cropformer: An interpretable deep learning framework for crop genomic prediction. Plant Communications. **6** (2024) 101223
2. Panzea: Maize genomic and phenotypic datasets. Available at: https://www.panzea.org/data
3. CIMMYT Research Data: Wheat dataset. Available at: https://hdl.handle.net/11529/10548918
4. SolOmics: Tomato dataset. Available at: http://solomics.agis.org.cn/tomato/ftp
5. Dryad: Rice dataset. Available at: https://datadryad.org/dataset/doi:10.5061/dryad.7369p
6. Ensembl Plants: Foxtail Millet (Setaria italica) dataset. Available at: https://plants.ensembl.org/Setaria_italica
7. NASA POWER: Prediction of Worldwide Energy Resources project. https://power.larc.nasa.gov
8. ISRIC SoilGrids: Soil composition database. Available at: https://soilgrids.org
9. H. Qiang, et al., "A graph-based genome and pan-genome variation of foxtail millet (Setaria italica)," Nature Genetics, 2023. Dataset available at Zenodo: https://zenodo.org/record/7367881; BioProjects: PRJNA841774, PRJNA842100; Phenotypes: https://doi.org/10.5281/zenodo.7755340; Source code: https://github.com/qiangh06/Setaria-pan-genome
10. Q. Zhao, et al., "Genome-wide association study of agronomic traits in foxtail millet," Nature Communications, 2019.
11. H. Qiang, et al., "Multi-omics analysis of Setaria under diverse environments," Frontiers in Plant Science, 2022.
12. J.L. Bennetzen, J. Schmutz, H. Wang, et al., "Reference genome sequence of the model plant Setaria," Nature Biotechnology, vol. 30, no. 6, pp. 555–561, 2012. https://doi.org/10.1038/nbt.2196
13. V. Jaiswal, et al., "Multi-environment GWAS identifies genomic regions underlying grain nutrient traits in foxtail millet (Setaria italica)," Plant Cell Reports, vol. 43, no. 6, 2023. Available: https://pubmed.ncbi.nlm.nih.gov/38127149/

14. S. Pramanik, et al., "Multi Season Evaluation of Foxtail Millet (Setaria italica (L.) P. Beauv.) Genotypes for Forage Yield Stability at the Nagaland Foothills," Environment & Ecology, MS21.indd, 2024. Available: https://environmentandecology.com/wp-content/uploads/2025/02/MS21-Multi-Season-Evaluation-of-Foxtail-Millet-Setaria.pdf

15. **R. Muhendran, R. Reddy, D. Aruna, and R. Akhsay,** "Optimizing Crop Traits with Genetic Innovations for Sustainable Agriculture," (Published by **IEEE,** 2025).

16. **A. M. Lee, M. Lim, B. F. Keng, and T. C. Cheew,** "Genomic selection for fruit improvement in fruits and vegetables: a systematic scoping review,"(**Springer,** 2023).

17. **J. P., K. M., P. N., and R. Reddy,** "Optimized Centric Method to Analyze the Seeds with Five Stages Technique to enhance the Quality," (Published by **IEEE,** 2023).

18. **S. Sharma, A. P., M. A. L. Balugaper, S. Malvaz, and R. C.,** "DeepG2IP: Fusing Multi-Modal Data to Improve Crop Production," (Published by **arXiv,** 2023).