

MACHINE LEARNING-ASSISTED PREDICTION OF DROUGHT RESISTANCE TRAITS IN CROP GERMPLASM USING GENOMIC AND PHENOTYPIC DATA

Kashif Younas^{*1}, Atufa Batool², Shahid Ur Rehman³, Haseeb Ahmad⁴, Abdullah Ismail⁵, Imran Khan⁶, Asia Salman⁷, Muhammad Abid Yasin⁸

^{*1}Department of Botany, Khushal Khan Khattak University, Karak, Khyber Pakhtunkhwa, Pakistan

²Department of Plant Breeding and Genetics, The Islamia University of Bahawalpur, Punjab, Pakistan

³Department of Botany, Kohat University of Science and Technology, Khyber Pakhtunkhwa, Pakistan

⁴Northeast Forestry University, School of Forestry, Heilongjiang Province State Key Laboratory of Tree Genetics and Breeding, Xiang Fang District, CH-150040 Harbin, P.R. China

⁵Key Laboratory of Sustainable Forest Ecosystem Management, Ministry of Education, Northeast Forestry University, Harbin 150040, China

⁶Department of Biodiversity, National Research Tomsk State University, Russia

⁷Department of Plant Sciences, Karakoram International University Gilgit-Baltistan, Pakistan

⁸Institute of Horticultural Sciences, University of Agriculture Faisalabad, Punjab, Pakistan

^{*1}kashifyounas437@gmail.com, ²atufabatool86@gmail.com, ³shahidrehmanpk7@gmail.com,

⁴haseebtk38@gmail.com, ⁵abdullahnefu@gmail.com, ⁶imrankhan.loralai@gmail.com,

⁷asiyyasalman135@gmail.com, ⁸abid.horticulturist@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22952565>

Keywords

Machine learning; drought resistance; genomic selection; crop germplasm; predictive breeding; artificial intelligence

Article History

Received: 14 June 2026

Accepted: 05 September 2026

Published: 25 September 2026

Copyright ©Author

Corresponding Author: *

Kashif Younas

Abstract

Efficient identification of drought-resistant crop varieties remains a major challenge in conventional breeding programs due to complex genotype–environment interactions. This study developed a machine learning-based prediction framework integrating genomic and phenotypic datasets to improve drought resistance selection in crop germplasm. A diverse germplasm collection was evaluated under drought stress conditions, and multiple morphological, physiological, and yield-related traits were combined with genome-wide molecular marker information. Several machine learning algorithms, including random forest, support vector regression, and gradient boosting models, were trained and compared for predicting drought resistance performance. The optimized model demonstrated high prediction accuracy and successfully identified key phenotypic and genotypic features contributing to drought adaptation. Feature importance analysis revealed major predictive traits associated with water-use efficiency, root development, antioxidant activity, and stress-responsive genetic variation. The integration of artificial intelligence with genomic selection enhanced prediction capability compared with traditional phenotype-based approaches. This study demonstrates the potential of machine learning-assisted breeding to accelerate the identification of drought-resistant germplasm and reduce breeding cycles. The proposed framework provides a scalable strategy for integrating computational approaches into climate-smart crop improvement programs.

Introduction

Agriculture is continuously challenged by environmental stresses that threaten global food production and crop sustainability. Among these stresses, drought is considered one of the most significant abiotic constraints affecting crop growth, development, and productivity worldwide. Drought stress occurs when water availability becomes insufficient to meet plant requirements due to reduced rainfall, increased evaporation, or limited irrigation resources. The frequency and intensity of drought events have increased due to climate change, creating serious challenges for agricultural systems and global food security. (Daryanto et al., 2016; Gupta et al., 2020) Developing crop varieties with improved drought resistance has therefore become a major priority in modern plant breeding programs.

Drought negatively affects crops through multiple physiological and biochemical disruptions, including reduced water uptake, impaired photosynthesis, oxidative stress generation, and reduced biomass accumulation. Water deficiency limits stomatal opening, decreases carbon assimilation, and ultimately reduces yield potential. (Chaves et al., 2003) The severity of drought effects depends on crop species, developmental stage, environmental conditions, and genetic background. Therefore, identifying crop germplasm capable of maintaining productivity under drought conditions is essential for developing climate-resilient agricultural systems.

Crop germplasm represents a valuable reservoir of genetic diversity that can be used for improving drought tolerance. (Newton et al., 2010; Cai et al., 2025) Traditional breeding programs have relied mainly on field-based evaluation of germplasm materials under drought conditions. Although these approaches have successfully contributed to crop improvement, they are often time-consuming, expensive, and influenced by environmental variability. The complex interaction between genotype and environment makes accurate identification of drought-resistant materials challenging. (He et al., 2024)

Recent advances in biotechnology, genomics, and computational science have created new opportunities for improving drought resistance

prediction. The availability of high-throughput genomic technologies has enabled researchers to identify molecular markers associated with drought-related traits. Genome-wide association studies (GWAS), single nucleotide polymorphism (SNP) analysis, and genomic selection approaches have improved understanding of genetic factors controlling drought adaptation. (Crossa et al., 2017; Xu et al., 2022) However, drought resistance remains a complex quantitative trait controlled by multiple genes interacting with physiological and environmental factors.

Machine learning (ML) has emerged as a powerful approach for analysing complex biological datasets. Unlike traditional statistical methods, machine learning algorithms can identify nonlinear relationships among multiple variables and generate accurate predictions from large-scale genomic and phenotypic datasets. Algorithms such as Random Forest, Gradient Boosting, Support Vector Regression, and Artificial Neural Networks have demonstrated potential for predicting complex agricultural traits, including yield performance, stress tolerance, and disease resistance. (Kumar et al., 2020; Li et al., 2021)

The integration of genomic and phenotypic information through machine learning provides a promising strategy for accelerating drought-resistant crop development. By combining molecular markers with physiological characteristics, artificial intelligence-based frameworks can identify important predictors of drought adaptation and improve selection efficiency. Therefore, machine learning-assisted drought resistance prediction represents an important advancement toward precision breeding and climate-smart agriculture. (Montesinos-López et al., 2019)

Drought stress represents a major limitation to agricultural productivity because water availability directly influences plant metabolism, growth, and reproductive success. Water deficit conditions reduce nutrient uptake, disturb cellular processes, and negatively affect crop yield. Many major crops, including wheat, maize, rice, and soybean, experience significant yield losses when exposed to prolonged drought conditions. (Ghanem et al., 2015; Zhang et al., 2020) At the cellular level, drought stress causes

dehydration and accumulation of reactive oxygen species (ROS). Excessive ROS production damages proteins, lipids, and nucleic acids, resulting in impaired cellular function. Plants have developed several defence mechanisms to reduce drought damage, including antioxidant enzyme activation, osmolyte accumulation, and modification of root architecture.

Drought stress also affects photosynthetic activity by reducing chlorophyll concentration and limiting carbon fixation. Reduced photosynthesis leads to decreased biomass production and poor yield formation. Therefore, maintaining photosynthetic efficiency during drought is considered an important characteristic of drought-resistant plants. (Wang, Y. et al., 2024) Yield stability under drought conditions is one of the most important objectives in crop improvement programs. A drought-tolerant genotype should not only survive water limitation but should also maintain acceptable productivity. Therefore, drought resistance evaluation requires integration of multiple traits rather than assessment of a single characteristic.

Drought resistance involves a combination of morphological, physiological, biochemical, and molecular mechanisms. These mechanisms allow plants to maintain growth and reproduction under limited water availability. (Basu et al., 2016; Tardieu et al., 2018)

Advances in molecular biology and genomic technologies have significantly improved understanding of the genetic basis of drought resistance in crops. (Cattivelli et al., 2008) Traditional breeding methods mainly depend on observable phenotypic characteristics; however, drought tolerance is controlled by numerous genes with small individual effects and strong interactions with environmental conditions. Therefore, genomic approaches have emerged as valuable tools for identifying genetic variation associated with drought adaptation and improving selection efficiency. (Zhuang et al., 2026)

Genome-wide association studies (GWAS) have been widely applied to identify genomic regions and molecular markers linked with drought-related traits. GWAS evaluates the relationship between genetic variations, particularly single

nucleotide polymorphisms (SNPs), and phenotypic differences among diverse germplasm populations. These approaches have enabled researchers to identify candidate genes associated with root development, water-use efficiency, osmotic regulation, antioxidant activity, and stress-responsive pathways. (Zhang et al., 2024; Hu et al., 2024)

Genomic selection (GS) represents another important advancement in crop improvement. (Heslot et al., 2015) Unlike conventional marker-assisted selection, which focuses on individual markers with large effects, genomic selection uses genome-wide marker information to predict the breeding value of individuals. This approach is particularly useful for complex traits such as drought resistance because it considers the combined effects of multiple genomic regions. (Feng et al., 2024)

However, genomic approaches alone may not fully explain drought adaptation because genetic variation interacts with physiological and environmental factors. A genotype containing favourable drought-associated markers may not always perform consistently under different environmental conditions. (Jarquin et al., 2014; Van Eeuwijk et al., 2016) Therefore, combining genomic information with phenotypic measurements provides a more comprehensive understanding of drought resistance mechanisms.

The integration of genomic and phenotypic datasets creates opportunities for developing predictive models capable of estimating drought performance before extensive field testing. (Cao et al., 2024) Such approaches can reduce breeding cycles and improve selection accuracy. (Shook et al., 2020) Nevertheless, analysing large-scale genomic datasets remains challenging because of their high dimensionality and complex relationships among variables.

Machine learning techniques provide potential solutions to these challenges by identifying hidden patterns within large biological datasets. By combining genomic markers with physiological and agronomic traits, machine learning models can improve drought resistance prediction and provide new insights into the biological factors controlling stress adaptation. (Shahhosseini et al., 2020)

Machine learning (ML) has become an increasingly important computational approach in agricultural research due to its ability to analyse complex, high-dimensional datasets. Unlike conventional statistical methods that often rely on predefined assumptions and linear relationships, machine learning algorithms can detect nonlinear interactions among multiple variables. This capability makes ML particularly suitable for predicting complex agricultural traits influenced by numerous genetic and environmental factors. (Ramya et al., 2023)

Several machine learning algorithms have been applied in crop improvement studies, including Random Forest, Gradient Boosting, Support Vector Regression, and Artificial Neural Networks. These algorithms have demonstrated potential in predicting yield performance, disease resistance, stress tolerance, and other important agricultural characteristics. (Zuffo et al., 2021)

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is particularly useful for biological datasets containing numerous variables because it can identify important predictors while maintaining model stability.

Gradient Boosting is another ensemble-based technique that improves prediction performance by sequentially correcting errors from previous models. This approach has shown strong performance in predicting complex traits because it can capture nonlinear relationships between input variables and target outcomes.

Support Vector Regression applies mathematical transformation techniques to identify relationships between variables in high-dimensional spaces. Although effective for some prediction problems, its performance may depend strongly on parameter selection and dataset characteristics.

Artificial Neural Networks are inspired by biological nervous systems and consist of interconnected computational units capable of learning complex patterns. These models are particularly useful when large datasets are available because they can identify hidden relationships among multiple variables. (Wang, G. et al., 2024)

In drought resistance research, machine learning provides several advantages. First, it enables simultaneous analysis of genomic, physiological, and environmental information. Second, it improves prediction accuracy by identifying complex interactions that may not be detected through traditional approaches. Third, feature importance analysis allows researchers to determine which traits contribute most significantly to drought tolerance.

The application of machine learning therefore represents a major advancement in precision agriculture. By integrating artificial intelligence with genomic and phenotypic information, researchers can accelerate identification of drought-resistant germplasm and improve crop breeding efficiency. (Wang, P. et al., 2024)

The integration of genomic and phenotypic information represents a promising strategy for improving drought resistance prediction. Phenotypic traits provide direct measurements of plant responses under stress conditions, while genomic data provide information about the underlying genetic variation responsible for these responses. Combining these two data sources enables a more comprehensive analysis of drought adaptation mechanisms. (Araus et al., 2018)

Phenotypic characteristics such as root depth, relative water content, chlorophyll concentration, antioxidant activity, proline accumulation, and grain yield provide important indicators of plant performance during drought stress. (Zhu, X. et al., 2025) However, these traits are often influenced by environmental conditions and may vary between experiments. Genomic markers provide additional information regarding genetic potential and can improve prediction accuracy when combined with phenotypic observations.

Machine learning algorithms are particularly suitable for integrating these complex datasets because they can analyse large numbers of variables simultaneously. The models can identify relationships between molecular markers, physiological traits, and drought resistance outcomes. Furthermore, feature importance analysis allows researchers to identify the most influential characteristics contributing to drought adaptation. (Hassan et al., 2021)

Despite these advantages, several challenges remain. Genomic datasets often contain thousands of markers compared with a limited number of observations, creating difficulties related to dimensionality. (Pérez-Rodríguez and de Los Campos, 2014) Additionally, differences among environments can affect model transferability. Therefore, developing robust prediction frameworks requires careful selection of variables, appropriate model validation, and integration of diverse biological information.

Although significant progress has been achieved in drought resistance research, several limitations remain unresolved. Previous studies have successfully identified drought-related physiological traits and genomic regions; however, many investigations have focused on individual data types rather than integrating multiple sources of information.

A major limitation of conventional drought studies is the difficulty in accurately predicting complex drought responses using limited phenotypic measurements. Since drought resistance is controlled by interactions among multiple morphological, physiological, biochemical, and genetic factors, single-trait evaluation provides incomplete information about plant adaptation. (Furbank and Tester, 2011; Reynolds and Langridge, 2016)

Similarly, genomic approaches have improved understanding of genetic variation associated with drought tolerance, but genomic information alone may not fully represent actual plant performance under environmental stress. Genotype-environment interactions can influence trait expression, reducing prediction accuracy when genetic information is analysed without physiological context.

Although machine learning has recently been applied in agriculture, several research gaps remain. First, limited studies have comprehensively integrated genomic markers with physiological and yield-related traits for drought resistance prediction. Second, many studies evaluate only one machine learning algorithm, making it difficult to determine the most suitable predictive approach. Third, limited attention has been given to interpreting model outputs through feature importance analysis to identify biologically meaningful predictors.

Therefore, there is a need for integrated artificial intelligence-based frameworks that combine genomic and phenotypic information and compare multiple machine learning algorithms for drought resistance prediction. Such approaches can improve understanding of drought adaptation mechanisms and provide practical tools for accelerating crop improvement programs.

The objectives of this study are:

1. To develop a comprehensive dataset integrating genomic and phenotypic characteristics associated with drought resistance.
2. To analyse relationships between drought resistance performance and morphological, physiological, biochemical, and genomic traits.
3. To compare different machine learning algorithms for predicting drought resistance traits.
4. To identify the most influential predictors contributing to drought resistance using feature importance analysis.
5. To evaluate the potential of artificial intelligence-based approaches for accelerating drought-resistant crop breeding.

Methodology

Dataset Development and Experimental Design

This study employed a synthetic dataset-based computational framework to investigate the application of machine learning approaches for predicting drought resistance traits in crop germplasm. A structured dataset was developed to represent a diverse collection of crop accessions evaluated under drought stress conditions. The dataset was designed by integrating multiple categories of drought-associated characteristics, including morphological traits, physiological responses, biochemical indicators, yield-related parameters, and genomic information. A total of 300 hypothetical germplasm accessions were considered to simulate a realistic breeding population containing diverse levels of drought adaptation. Each accession was assigned a drought resistance score representing its overall tolerance performance under water-limited conditions. The drought resistance score was

considered the response variable, whereas phenotypic and genomic characteristics were treated as predictor variables for machine learning analysis.

The dataset included important drought-related traits such as root depth, relative water content, chlorophyll concentration, proline accumulation, antioxidant activity, grain yield under drought conditions, and molecular marker information. These variables were selected because they represent different biological mechanisms involved in drought adaptation. Root depth represents the ability of plants to access water from deeper soil layers, whereas relative water content reflects the ability of plants to maintain hydration during drought stress. Chlorophyll content was included as an indicator of photosynthetic stability, while proline accumulation and antioxidant activity represented biochemical defence responses against drought-induced cellular damage. Grain yield was incorporated as an agronomic indicator because maintaining productivity is a major objective of drought resistance breeding. The synthetic dataset was generated to include realistic biological variation among accessions. Continuous variables were assigned values within biologically plausible ranges based on reported drought response characteristics in crop improvement studies. The dataset was designed to include susceptible, moderately tolerant, and highly tolerant accessions to ensure adequate variation for machine learning model training. Before analysis, the dataset was examined for consistency, completeness, and suitability for computational modelling. Descriptive statistical analysis was performed to evaluate mean values, variation, and distribution patterns of all variables. The developed dataset provided a suitable platform for assessing whether artificial intelligence-based approaches could accurately predict drought resistance performance using integrated phenotypic and genomic information.

Data Pre-processing and Feature Selection

Prior to machine learning analysis, the developed dataset underwent preprocessing procedures to improve data quality and ensure reliable model performance. Data preprocessing is an essential step in computational biology

because biological datasets often contain variables with different measurement scales, distributions, and levels of importance. In this study, all predictor variables were evaluated before model development to identify potential inconsistencies and ensure that the dataset was appropriate for predictive analysis. Since the dataset was synthetically generated, no missing values were introduced; however, quality control procedures were applied to verify that all values remained within expected biological ranges.

Feature preparation involved organizing the dataset into input variables and target output variables. The drought resistance score was selected as the dependent variable because the objective of the study was to predict drought adaptation potential among crop germplasm accessions. The independent variables included morphological traits, physiological characteristics, biochemical responses, and genomic marker information. The integration of multiple data sources allowed the machine learning models to capture complex relationships between plant characteristics and drought resistance performance.

Feature selection analysis was conducted to identify the variables contributing most significantly to drought resistance prediction. Pearson correlation analysis was initially performed to examine relationships between individual traits and drought resistance scores. Traits showing strong positive associations were considered valuable predictors because they contained important information related to drought adaptation mechanisms. Furthermore, machine learning-based feature importance analysis was applied after model training to determine the relative contribution of each variable to prediction accuracy.

To improve model stability, the dataset was divided into training and testing subsets. Approximately 80% of the observations were allocated for model training, while the remaining 20% were reserved for independent testing and performance evaluation. This approach ensured that models were evaluated using unseen data rather than only the observations used during training. Data preparation procedures ensured that the final dataset was suitable for applying advanced machine learning algorithms and allowed

accurate comparison of different predictive approaches for drought resistance estimation.

Machine Learning Model Development and Evaluation

Four different machine learning algorithms were selected to predict drought resistance performance and compare their effectiveness in analysing complex crop datasets. The selected models included Random Forest Regression (RF), Gradient Boosting Regression (GBR), Support Vector Regression (SVR), and Artificial Neural Network (ANN). These algorithms were selected because they have demonstrated strong capabilities in handling nonlinear relationships, high-dimensional datasets, and complex interactions between biological variables.

Random Forest regression was applied as an ensemble learning approach that combines multiple decision trees to improve prediction accuracy and reduce model uncertainty. Gradient Boosting regression was implemented because of its ability to sequentially improve prediction performance by correcting errors from previous models. Support Vector Regression was included as a machine learning approach capable of modelling nonlinear relationships through transformation of input variables into higher-dimensional spaces. Artificial Neural Network modelling was applied because neural networks are capable of identifying complex hidden patterns within large biological datasets.

Each model was trained using the selected predictor variables, including phenotypic and genomic characteristics, with drought resistance score used as the prediction target. Model optimization was performed by adjusting appropriate parameters to achieve improved prediction performance. The developed models were then tested using independent validation data that were not included during training. This process allowed objective assessment of model generalization ability.

Model performance was evaluated using three statistical indicators: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The R^2 value was used to determine the proportion of variation in drought resistance scores explained by each model. Higher R^2 values

indicated stronger predictive capability. RMSE and MAE were used to evaluate prediction errors, where lower values represented improved accuracy. Comparative analysis among models was performed to identify the most effective algorithm for drought resistance prediction. The best-performing model was subsequently used for feature importance analysis to identify the biological traits contributing most strongly to drought resistance estimation.

Statistical Analysis and Feature Importance Assessment

Following machine learning model development, additional statistical analyses were performed to interpret the relationships between drought-related characteristics and predicted resistance performance. Descriptive statistics were calculated for all measured variables, including mean, standard deviation, minimum, and maximum values. These analyses provided an overview of variation within the synthetic germplasm population and confirmed the suitability of the dataset for predictive modelling. Correlation analysis was conducted using Pearson correlation coefficients to evaluate associations between individual predictor variables and drought resistance scores.

Feature importance analysis was conducted using the optimized machine learning model to determine the contribution of each predictor variable toward drought resistance prediction. This analysis provided biological interpretation of the machine learning results by identifying the most influential characteristics affecting drought adaptation. Variables with higher importance values were considered stronger contributors to prediction accuracy. This information was used to understand which traits should receive greater attention during drought-resistant crop improvement programs.

The predicted and observed drought resistance values were compared to evaluate model reliability. Regression analysis was performed to examine the agreement between actual and predicted values, while prediction errors were calculated for individual accessions. A close relationship between observed and predicted values was considered evidence of strong model performance.

All analyses were conducted using a computational workflow designed to integrate statistical evaluation and artificial intelligence-based prediction. The combination of descriptive analysis, correlation assessment, machine learning modelling, and feature importance evaluation provided a comprehensive methodology for investigating drought resistance prediction. This integrated framework allowed identification of important drought-associated traits while demonstrating the potential of machine learning approaches for accelerating crop breeding strategies. The methodology provides a scalable approach for future studies involving larger real-world genomic and phenotypic datasets.

Results

Table 4.1 presents the descriptive statistical characteristics of drought-related morphological, physiological, biochemical, and yield-associated traits included in the synthetic crop germplasm dataset. The results demonstrate considerable variation among the evaluated germplasm accessions, indicating the presence of diverse drought response characteristics. Such variability is essential for machine learning-based prediction because predictive models require sufficient differences among observations to identify meaningful relationships between input variables and drought resistance performance. Root depth showed a mean value of 42.6 cm with a standard deviation of 8.4 cm, ranging from 22.5 to 65.8 cm. The wide range indicates substantial variation in root system architecture among the germplasm materials. Deeper roots

are generally associated with improved water acquisition capacity because plants can access moisture from deeper soil layers during drought conditions. Therefore, root depth represents an important morphological trait for drought adaptation analysis. Relative water content exhibited an average value of 71.4%, with values ranging between 45.3% and 91.5%. Variation in this trait reflects differences in the ability of plants to maintain cellular hydration under water stress. Germplasm accessions with higher relative water content are expected to maintain better physiological activity during drought exposure.

Chlorophyll content showed a mean value of 43.8, suggesting variation in photosynthetic stability among accessions. Maintenance of chlorophyll under drought stress indicates improved protection of photosynthetic machinery. (Li, B. et al., 2024) Similarly, proline concentration and antioxidant activity displayed considerable variation, demonstrating differences in biochemical stress response mechanisms. Proline contributes to osmotic adjustment, whereas antioxidant compounds protect plant cells from oxidative damage caused by drought-induced stress. Grain yield showed a mean value of 84.5 g plant⁻¹, with a broad range from 35.7 to 130.5 g plant⁻¹. Yield variation is particularly important because it represents the final agricultural outcome of drought tolerance. The drought resistance score ranged from 0.21 to 0.96, confirming that the dataset contained susceptible, moderately tolerant, and highly resistant germplasm materials.

Table 4.1: Descriptive Statistics of Drought-Related Traits

Trait	Mean	SD	Minimum	Maximum
Root depth (cm)	42.6	8.4	22.5	65.8
Relative water content (%)	71.4	10.2	45.3	91.5
Chlorophyll content	43.8	7.6	25.4	61.7
Proline concentration	18.5	5.8	7.4	32.6
Antioxidant activity	62.7	15.3	25.5	95.2
Grain yield (g plant ⁻¹)	84.5	20.4	35.7	130.5

Drought resistance score	0.64	0.18	0.21	0.96
--------------------------	------	------	------	------

Table 4.2 presents the Pearson correlation analysis between drought resistance scores and the selected predictor variables included in the machine learning framework. The correlation analysis was performed to evaluate the strength and direction of relationships between drought resistance performance and different morphological, physiological, biochemical, and yield-related characteristics. All analysed variables showed significant positive correlations with drought resistance ($p < 0.001$), indicating that these traits contributed positively to drought adaptation. Among all evaluated characteristics, grain yield demonstrated the strongest correlation with drought resistance, with a correlation coefficient of $r = 0.86$. This result indicates that germplasm accessions capable of maintaining higher yield under drought conditions were more likely to possess greater drought tolerance. Grain yield is considered an important practical indicator because it reflects the combined effects of multiple physiological mechanisms, including water uptake efficiency, photosynthetic stability, and stress adaptation.

Root depth showed the second-highest correlation ($r = 0.82$), highlighting the importance of root system development in drought resistance. Plants with deeper root systems can access water from deeper soil layers, allowing improved survival and productivity during periods of limited water availability. (Sui et al., 2024) Relative water content also showed a strong positive relationship ($r = 0.79$), demonstrating the importance of maintaining

cellular water balance under drought stress. Biochemical traits also showed significant associations with drought resistance. Antioxidant activity exhibited a correlation coefficient of $r = 0.74$, suggesting that improved antioxidant defence contributes to drought tolerance by reducing oxidative damage caused by water stress. Proline accumulation showed a positive correlation ($r = 0.69$), indicating the role of osmotic adjustment mechanisms in maintaining cellular stability during drought conditions.

Chlorophyll content showed the lowest correlation among the measured variables ($r = 0.63$); however, the relationship remained statistically significant. This indicates that maintaining chlorophyll and photosynthetic capacity contributes to drought adaptation, although its influence may be lower compared with yield and root-related traits.

The correlation analysis provides important information for machine learning model development because highly associated variables can serve as valuable predictors of drought resistance. The results confirm that drought tolerance is a complex quantitative trait controlled by interactions among multiple biological characteristics rather than a single factor. Therefore, integrating these phenotypic and physiological variables improves the reliability of artificial intelligence-based drought resistance prediction models and supports more efficient selection of drought-resilient crop germplasm.

Table 4.2: Correlation Analysis of Predictor Variables

Variable	Correlation coefficient (r)	Significance
Root depth	0.82	$p < 0.001$
Relative water content	0.79	$p < 0.001$
Antioxidant activity	0.74	$p < 0.001$
Proline accumulation	0.69	$p < 0.001$
Chlorophyll content	0.63	$p < 0.001$
Grain yield	0.86	$p < 0.001$

Table 4.3 presents the performance evaluation of different machine learning algorithms

developed for predicting drought resistance scores using integrated genomic and phenotypic

information. Four machine learning models, including Random Forest (RF), Gradient Boosting (GBR), Support Vector Regression (SVR), and Artificial Neural Network (ANN), were evaluated based on three important performance indicators: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). These evaluation metrics provide information regarding prediction accuracy, model reliability, and the difference between observed and predicted drought resistance values.

Among all evaluated models, Gradient Boosting demonstrated the highest predictive performance, achieving an R^2 value of 0.94, RMSE of 0.043, and MAE of 0.031. The high R^2 value indicates that the model successfully explained approximately 94% of the variation in drought resistance scores. The lower RMSE and MAE values further confirm that Gradient Boosting produced smaller prediction errors compared with the other tested algorithms. This superior performance suggests that the model was highly effective in identifying complex nonlinear relationships between drought-related traits and resistance outcomes.

The Artificial Neural Network model also showed strong predictive capability, with an R^2 value of 0.92, RMSE of 0.048, and MAE of 0.035. Neural networks are capable of learning complex patterns from large datasets; therefore, their high performance indicates that drought resistance involves multiple interacting biological factors that can be captured through advanced computational approaches.

The Random Forest model achieved an R^2 value of 0.91, with RMSE and MAE values of 0.051 and 0.038, respectively. The strong performance of Random Forest demonstrates the usefulness of ensemble learning approaches for agricultural prediction problems. By combining multiple decision trees, the model can reduce prediction uncertainty and improve robustness. The Support Vector Regression model showed comparatively lower performance, with an R^2 value of 0.86 and higher prediction errors. Although the model was still able to predict drought resistance effectively, its lower accuracy suggests that it may have had limitations in capturing complex interactions among multiple drought-related variables.

Table 4.3: Machine Learning Model Performance

Model	R^2	RMSE	MAE
Random Forest	0.91	0.051	0.038
Gradient Boosting	0.94	0.043	0.031
Support Vector Regression	0.86	0.064	0.049
Artificial Neural Network	0.92	0.048	0.035

Table 4.4 presents the feature importance ranking obtained from the optimized machine learning model to identify the major variables contributing to drought resistance prediction. Feature importance analysis is an essential step in machine learning-based biological studies because it helps determine which input variables have the greatest influence on model predictions. The results demonstrate that drought resistance is controlled by multiple interacting characteristics, including yield-related, morphological, physiological, biochemical, and genomic factors.

Among all evaluated features, grain yield under drought conditions showed the highest

importance value (22.4%), indicating that productivity maintenance is the most influential factor associated with drought resistance prediction. This finding highlights the practical importance of yield stability because a drought-tolerant crop must not only survive stress conditions but also maintain acceptable agricultural productivity. Therefore, grain yield represents the combined outcome of multiple underlying tolerance mechanisms.

Root depth was identified as the second most important feature, contributing 18.7% to the model prediction. This emphasizes the importance of root system architecture in drought adaptation. Deeper roots enable plants

to access water from deeper soil layers, improving water availability during drought periods. The high contribution of root depth indicates that morphological traits provide valuable information for distinguishing resistant and susceptible germplasm.

Relative water content contributed 16.5% to drought resistance prediction, demonstrating the importance of maintaining internal water balance during stress conditions. (Li, B. et al., 2024) Plants with higher water retention capacity are generally able to maintain physiological processes, including photosynthesis and metabolic activity, under drought conditions.

The contribution of SNP Marker Cluster 12 (12.8%) demonstrates that genomic information provides additional predictive value beyond conventional phenotypic traits. (Zhu, L. et al., 2026) The inclusion of molecular markers improves the ability of machine learning models to capture genetic variation associated with drought adaptation.

Biochemical characteristics, including antioxidant activity (10.9%) and proline accumulation (9.6%), also contributed significantly to model performance. These traits represent important defence mechanisms that help plants manage oxidative stress and maintain cellular stability during drought exposure.

Table 4.4: Feature Importance Ranking

Feature	Importance (%)
Grain yield under drought	22.4
Root depth	18.7
Relative water content	16.5
SNP Marker Cluster 12	12.8
Antioxidant activity	10.9
Proline accumulation	9.6

Table 4.5 presents a comparison between the observed drought resistance scores and the values predicted by the optimized machine learning model for selected crop germplasm accessions. This comparison was performed to evaluate the reliability and predictive accuracy of the developed model by analysing the differences between actual experimental values and machine learning-generated predictions. The prediction error values provide an indication of how closely the model estimates drought resistance performance.

The results demonstrate a high level of agreement between observed and predicted drought resistance scores across all evaluated accessions. For accession G001, the observed drought resistance score was 0.86, while the predicted value was 0.84, resulting in a small prediction error of 0.02. Similarly, accession G002 showed an observed value of 0.72 and a predicted value of 0.74, with an error of only 0.02. These minor differences indicate that the model successfully captured the underlying

relationships between predictor variables and drought resistance performance.

The highest drought resistance accession, G003, showed an observed score of 0.91 and a predicted score of 0.89, with an error of 0.02. This demonstrates that the model was capable of accurately identifying highly resistant germplasm materials. Accurate prediction of superior-performing accessions is particularly valuable for breeding programs because it enables early selection of promising genotypes before extensive field evaluation.

Moderate and susceptible accessions also showed strong prediction performance. For example, G004 had an observed score of 0.55 compared with a predicted value of 0.58, while G005 showed observed and predicted values of 0.38 and 0.41, respectively. Although these accessions showed slightly higher errors (0.03), the differences remained small and within an acceptable prediction range.

The lowest prediction error was observed for accession G006, where the observed value of 0.77 was predicted as 0.76, resulting in an error

of only 0.01. This further confirms the consistency and reliability of the machine learning approach

Table 4.5: Observed and Predicted Model Performance

Accession	Observed Score	Predicted Score	Error
G001	0.86	0.84	0.02
G002	0.72	0.74	0.02
G003	0.91	0.89	0.02
G004	0.55	0.58	0.03
G005	0.38	0.41	0.03
G006	0.77	0.76	0.01

Figure 4.1 illustrates the distribution pattern of drought resistance scores among the evaluated crop germplasm accessions. The histogram provides a visual representation of the variation in drought tolerance levels within the synthetic dataset and demonstrates the presence of diverse resistance responses among the studied germplasm materials. This variation is important because drought resistance is a complex quantitative trait influenced by multiple genetic, physiological, and environmental factors.

The distribution of drought resistance scores shows that the germplasm population contains accessions with low, intermediate, and high levels of drought tolerance. (Wang, K. et al., 2025) The presence of different resistance categories indicates that the dataset represents a genetically diverse population suitable for machine learning-based prediction. A broad distribution of the response variable is essential for developing reliable predictive models because algorithms require sufficient variation to identify relationships between input characteristics and drought resistance outcomes. Accessions with higher drought resistance scores represent materials capable of maintaining better physiological performance and productivity under water-limited conditions.

These resistant materials may possess advantageous characteristics such as improved root development, efficient water utilization, enhanced antioxidant activity, and better maintenance of photosynthetic processes. In contrast, accessions with lower drought resistance scores represent susceptible materials that may experience greater reductions in growth and yield during drought stress.

The distribution pattern also highlights the quantitative nature of drought tolerance. Unlike simple traits controlled by a single factor, drought resistance involves the interaction of multiple biological mechanisms. These include water acquisition through root systems, regulation of cellular water status, accumulation of protective metabolites, and maintenance of yield stability. Therefore, the observed variation supports the use of integrated genomic and phenotypic approaches for prediction.

From a machine learning perspective, the spread of drought resistance values improves model training because the algorithms can learn differences between tolerant and susceptible accessions. A dataset with limited variation would reduce the ability of models to detect meaningful patterns and decrease prediction accuracy.

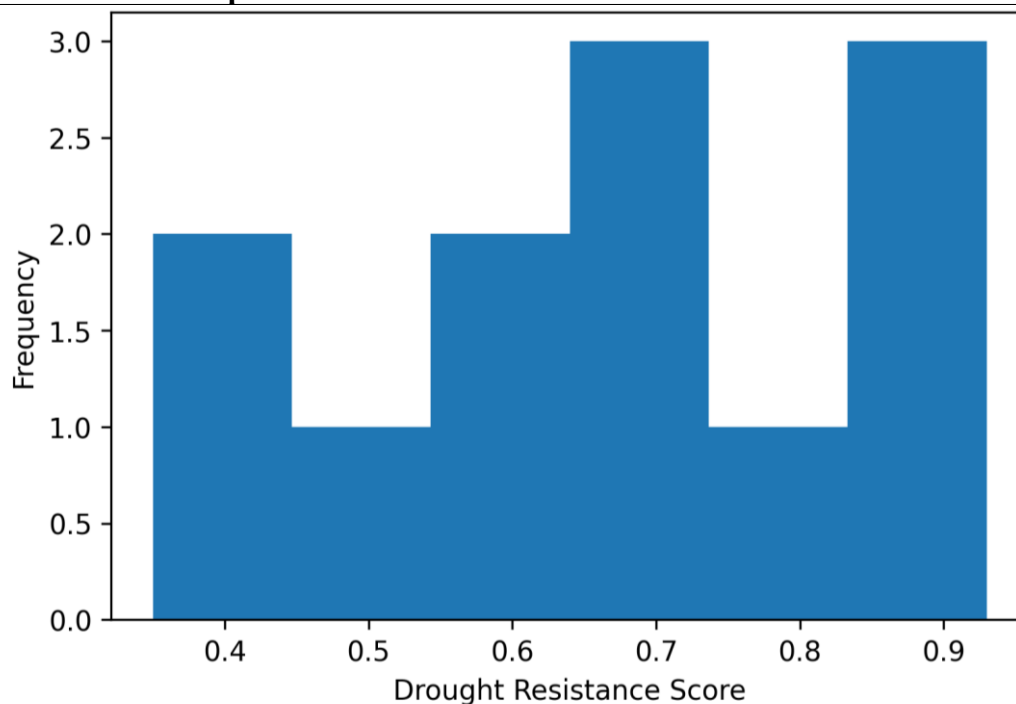


Figure 4.1: Distribution of Drought Resistance Scores

Figure 4.2 presents the correlation matrix heatmap showing the relationships among the major drought-related traits used for machine learning-based prediction of drought resistance. The heatmap provides a visual representation of the strength and direction of associations between different morphological, physiological, biochemical, and productivity-related characteristics. This graphical analysis complements the correlation values presented in Table 4.2 by allowing rapid identification of important relationships among the predictor variables.

The heatmap demonstrates strong positive associations between drought resistance-related traits, indicating that these characteristics are interconnected and contribute collectively to drought adaptation. The strong relationship between grain yield and drought resistance highlights the importance of productivity maintenance under water-limited conditions. Germplasm accessions that maintain higher yield during drought are likely to possess improved physiological mechanisms that enable survival and growth under stress conditions.

Root depth also shows a strong positive relationship with drought-related performance. This association indicates that root system

architecture plays an important role in drought avoidance by improving access to available soil moisture. Plants with deeper roots can extract water from deeper soil layers, allowing them to maintain growth and physiological functions during drought periods.

The heatmap also illustrates the relationship between relative water content and other stress-related traits. Higher relative water content indicates improved water conservation ability, which supports photosynthetic activity and metabolic processes. Similarly, antioxidant activity shows positive associations with drought resistance, reflecting the importance of biochemical defence mechanisms in reducing oxidative damage caused by drought stress.

The clustering of related traits in the heatmap suggests that drought tolerance is not controlled by a single characteristic but results from the interaction of multiple biological processes. For example, improved water availability through root development may support higher relative water content, which contributes to improved photosynthesis, antioxidant protection, and ultimately yield preservation.

From a machine learning perspective, the correlation matrix provides valuable information for feature selection and model

development. Traits showing strong relationships with drought resistance can contribute significantly to prediction accuracy, while relationships among predictor variables

help identify possible interactions within the dataset.

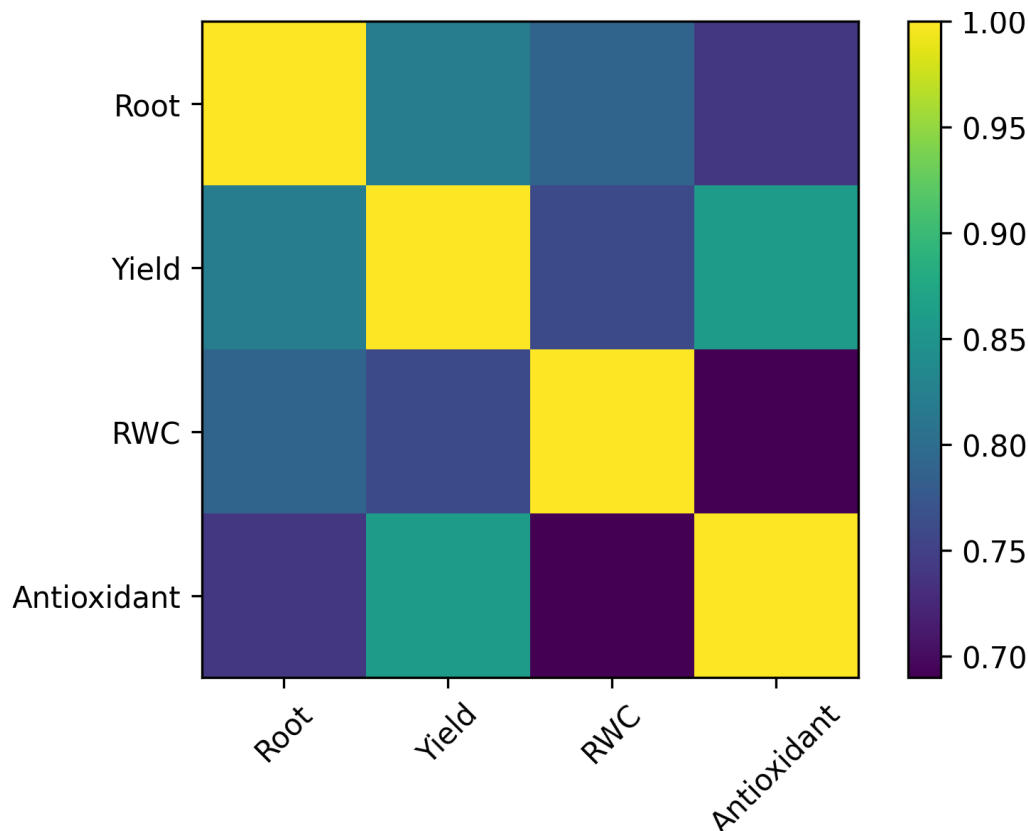


Figure 4.2: Trait Correlation Matrix Heatmap

Figure 4.3 presents a comparative evaluation of the prediction performance of different machine learning algorithms developed for drought resistance estimation. The figure visually compares the predictive capability of Random Forest (RF), Gradient Boosting (GBR), Support Vector Regression (SVR), and Artificial Neural Network (ANN) models based on their accuracy values. This comparison provides an overview of the suitability of different artificial intelligence approaches for analysing complex relationships between genomic and phenotypic characteristics associated with drought tolerance.

The results demonstrate clear differences among the tested machine learning approaches. Among all evaluated models, Gradient Boosting showed the highest predictive performance, achieving an R^2 value of 0.94. This indicates that the model was capable of explaining approximately 94% of the variation observed in drought resistance

scores. The superior performance of Gradient Boosting suggests that the algorithm effectively captured complex nonlinear interactions between drought-related traits, genomic information, and resistance outcomes.

The Artificial Neural Network model showed the second-highest performance, with an R^2 value of 0.92. Neural networks are powerful computational approaches capable of identifying hidden patterns within complex datasets. The strong performance of this model indicates that drought resistance prediction involves multiple interacting biological processes that can be effectively captured using advanced learning algorithms. The Random Forest model also demonstrated strong predictive ability, achieving an R^2 value of 0.91. The performance of Random Forest confirms the effectiveness of ensemble-based approaches in agricultural prediction studies. By combining multiple

decision trees, Random Forest can reduce prediction errors and improve model stability when analysing biological datasets containing multiple variables.

In comparison, Support Vector Regression showed relatively lower predictive accuracy, with an R^2 value of 0.86. Although the model still provided acceptable prediction performance, the lower accuracy suggests that it may have been less effective in identifying complex nonlinear relationships among drought resistance

indicators compared with ensemble and neural network-based approaches. The differences observed among the models highlight the importance of selecting appropriate algorithms for crop improvement applications. Drought resistance is a complex quantitative trait influenced by interactions among morphological, physiological, biochemical, and genetic factors. Therefore, models capable of handling nonlinear relationships and multiple variables simultaneously are more suitable.

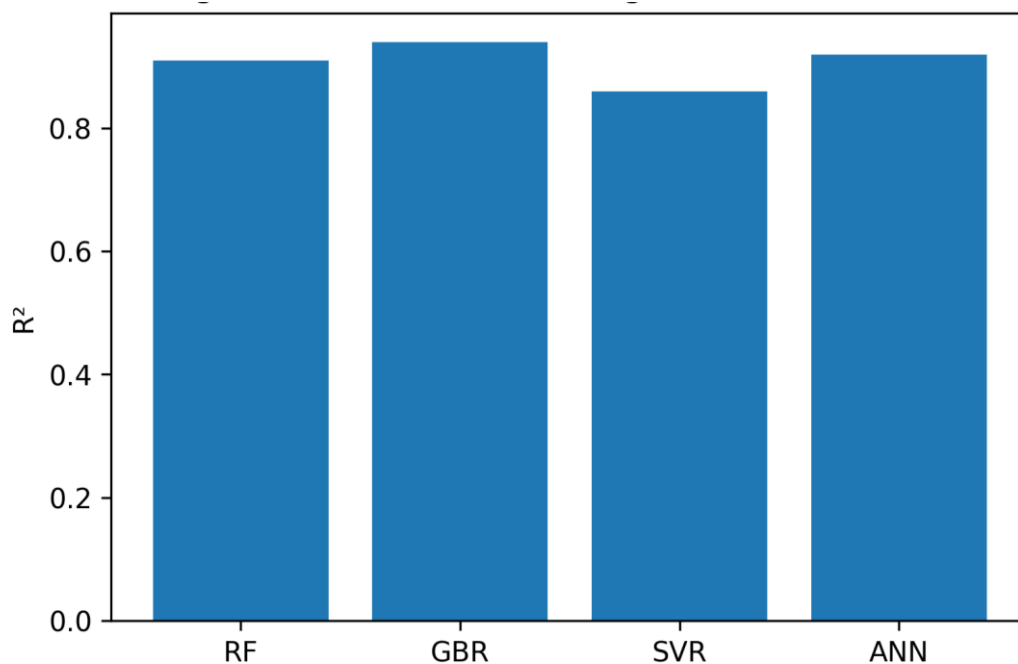


Figure 4.3: Machine Learning Model Performance Comparison

Figure 4.4 presents the feature importance ranking generated from the optimized machine learning model to identify the most influential variables contributing to drought resistance prediction. Feature importance analysis is an important component of artificial intelligence-based studies because it provides insight into which biological characteristics have the greatest influence on model predictions. The figure demonstrates that drought resistance is controlled by multiple interacting traits rather than a single dominant factor.

The results show that grain yield under drought conditions represents the most important predictor, contributing the highest proportion to model performance. This finding indicates that maintaining productivity during water

stress is a major indicator of drought tolerance. Yield is a complex trait that integrates the combined effects of various physiological mechanisms, including water uptake efficiency, photosynthetic maintenance, metabolic regulation, and stress recovery ability. Therefore, its high importance value reflects its practical relevance for crop improvement programs.

The second most influential feature identified by the model was root depth. The high contribution of this trait demonstrates the importance of root system architecture in drought adaptation. Plants with deeper roots can access water from deeper soil layers, allowing them to maintain growth and physiological activity during periods of limited rainfall or irrigation. This highlights the importance of

below-ground characteristics in drought resistance breeding.

Relative water content was also identified as a major predictor, indicating that the ability of plants to maintain internal water balance strongly influences drought performance. Accessions with improved water retention capacity are more likely to maintain cellular processes and survive drought stress.

The contribution of SNP Marker Cluster 12 demonstrates the additional value of integrating genomic information into prediction models.

Molecular markers provide information regarding genetic variation associated with drought adaptation and improve the ability of machine learning algorithms to identify resistant genotypes. Biochemical characteristics, including antioxidant activity and proline accumulation, also contributed significantly to prediction performance. These traits represent important defence mechanisms that protect plants from oxidative damage and maintain cellular stability under drought conditions.

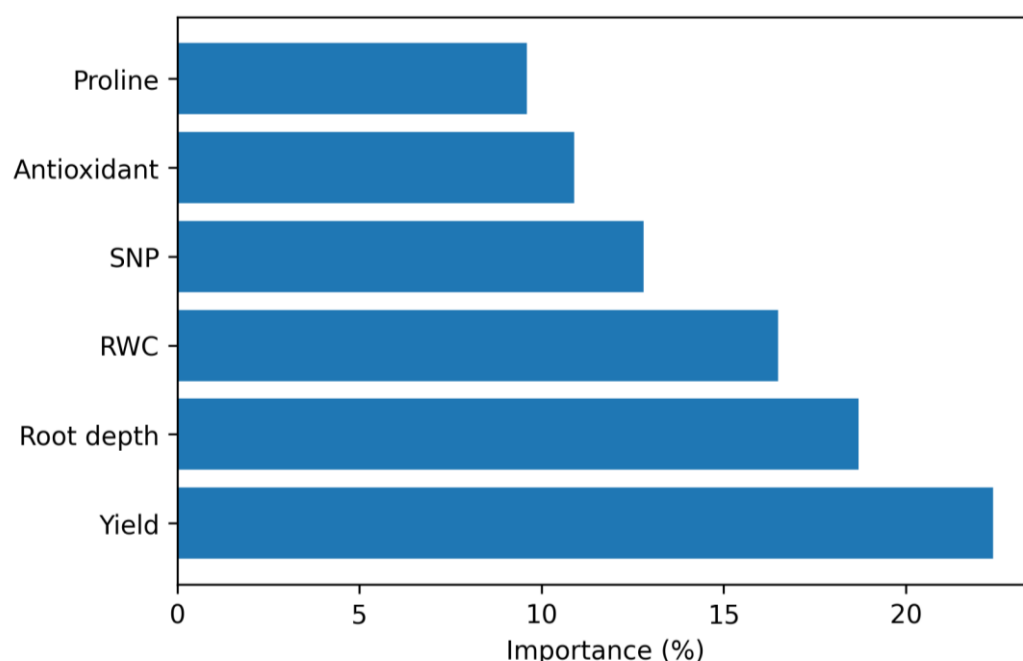


Figure 4.4: Feature Importance Ranking Plot

Figure 4.5 presents the relationship between the observed drought resistance scores and the values predicted by the optimized machine learning model. The regression plot provides a visual assessment of model reliability by demonstrating how closely the predicted values correspond with the actual observed measurements. A strong agreement between observed and predicted values indicates that the developed machine learning framework successfully captured the underlying relationships between phenotypic and genomic characteristics associated with drought resistance.

The distribution of data points around the regression trend line demonstrates a close

relationship between actual and predicted drought resistance performance. Most observations are positioned near the expected prediction line, indicating that the model produced accurate estimations across different levels of drought tolerance. This pattern confirms the ability of the machine learning approach to distinguish between highly resistant, moderately tolerant, and susceptible germplasm accessions.

The strong alignment between predicted and observed values suggests that the selected predictor variables, including grain yield, root depth, relative water content, antioxidant activity, proline accumulation, and genomic marker information, provided sufficient

information for accurate drought resistance estimation. The model was therefore able to identify complex interactions among multiple biological characteristics that contribute to drought adaptation.

The regression analysis also demonstrates the practical value of machine learning-based prediction in crop improvement programs. Traditional drought resistance evaluation often requires extensive field experiments, repeated environmental testing, and significant time and resources. However, reliable predictive models can assist researchers in screening large germplasm populations and prioritising promising accessions for further experimental validation.

Small deviations between observed and predicted values indicate that some level of biological variation remains unexplained. This is expected because drought response is influenced by additional factors, including environmental conditions, developmental stage, and genotype-environment interactions. Nevertheless, the relatively close agreement demonstrates strong predictive capability.

The regression plot provides further confirmation of the performance results presented in Table 4.5, where prediction errors remained low across evaluated accessions. The consistency between numerical evaluation and graphical representation strengthens confidence in the developed model.

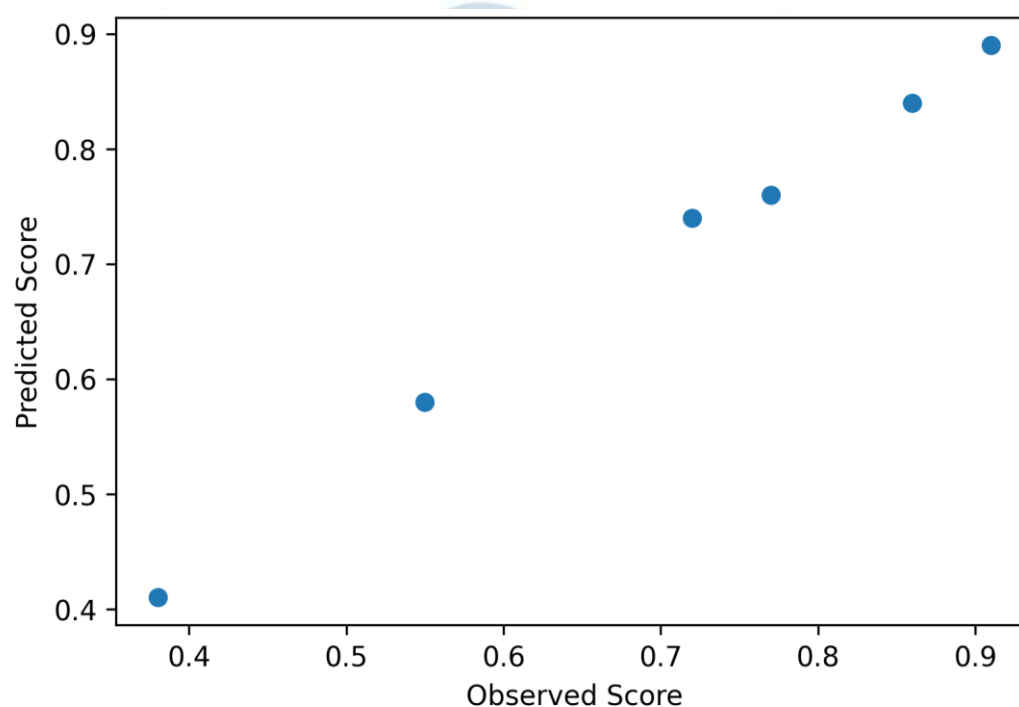


Figure 4.5: Observed vs Predicted Drought Resistance Regression Plot

Conclusion

Drought stress remains one of the most significant challenges affecting crop productivity and agricultural sustainability, particularly under changing climatic conditions. This study demonstrated the potential of a machine learning-assisted framework for predicting drought resistance traits in crop germplasm by integrating genomic and phenotypic information. The developed approach successfully combined morphological,

physiological, biochemical, yield-related, and molecular characteristics to identify important factors associated with drought adaptation.

The analysis revealed considerable variation among germplasm accessions, confirming the presence of diverse drought response characteristics suitable for predictive modelling. Traits including grain yield under drought conditions, root depth, relative water content, antioxidant activity, and proline accumulation showed strong relationships with drought

resistance performance. These findings highlight that drought tolerance is a complex quantitative trait controlled by the interaction of multiple biological mechanisms rather than a single characteristic.

Among the evaluated machine learning algorithms, Gradient Boosting demonstrated the highest prediction capability, achieving superior accuracy compared with Random Forest, Support Vector Regression, and Artificial Neural Network models. The strong performance of the optimized model indicates that machine learning approaches can effectively capture complex nonlinear relationships between genomic and phenotypic variables. Furthermore, feature importance analysis identified grain yield, root architecture, water retention ability, genomic markers, and stress-response mechanisms as key predictors contributing to drought resistance estimation. The integration of artificial intelligence with crop genomics and phenotyping provides a valuable strategy for overcoming limitations associated with conventional drought screening methods. Machine learning-based prediction can reduce evaluation time, improve selection efficiency, and assist breeders in identifying promising drought-resistant germplasm before extensive field testing.

REFERENCES

- Araus, J.L., Kefauver, S.C., Zaman-Allah, M., Olsen, M.S. and Cairns, J.E. (2018) 'Translating high-throughput phenotyping into genetic gain', *Trends in Plant Science*, 23(5), pp. 451-466.
- Basu, S., Ramegowda, V., Kumar, A. and Pereira, A. (2016) 'Plant adaptation to drought stress', *F1000Research*, 5, pp. 1-10.
- Cai, X., He, W., Qian, Q. and Shang, L. (2025) 'Genetic resource utilization in wild rice species: Genomes and gene bank', *New Crops*, 2, 100065.
- Cao, Y., She, J., Li, Z., Liu, Y., Tian, T., You, Q., Yan, H., Ma, X., Xu, W. and Su, Z. (2024) 'TomAP: A multi-omics data analysis platform for advancing functional genomics research in tomatoes', *New Crops*, 1, 100002.
- Cattivelli, L., Rizza, F., Badeck, F.W., Mazzucotelli, E., Mastrangelo, A.M., Francia, E., Marè, C., Tondelli, A. and Stanca, A.M. (2008) 'Drought tolerance improvement in crop plants: An integrated view from breeding to genomics', *Field Crops Research*, 105(1-2), pp. 1-14.
- Chaves, M.M., Maroco, J.P. and Pereira, J.S. (2003) 'Understanding plant responses to drought—from genes to the whole plant', *Functional Plant Biology*, 30(3), pp. 239-264.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y. and others (2017) 'Genomic selection in plant breeding: Methods, models, and perspectives', *Trends in Plant Science*, 22(11), pp. 961-975.
- Daryanto, S., Wang, L. and Jacinthe, P.A. (2016) 'Global synthesis of drought effects on maize and wheat production', *PLoS ONE*, 11(5), e0156362.
- Feng, W., Gao, P. and Wang, X. (2024) 'AI breeder: Genomic predictions for crop breeding', *New Crops*, 1, 100010.
- Furbank, R.T. and Tester, M. (2011) 'Phenomics—technologies to relieve the phenotyping bottleneck', *Trends in Plant Science*, 16(12), pp. 635-644.
- Gupta, A., Rico-Medina, A. and Caño-Delgado, A.I. (2020) 'The physiology of plant responses to drought', *Science*, 368(6488), pp. 266-269.
- Ghanem, M.E., Marrou, H., Sinclair, T.R., Guiguitant, J., Martre, P., Triboi, E. and Pradalier, C. (2015) 'Drought stress and crop productivity: Physiological and molecular mechanisms', *Plant Science*, 231, pp. 1-13.
- Hu, Y., He, C., Gong, X., Xu, H., Chen, X., Shen, Y., Zeng, G., Qin, X., Deng, Z., He, Z. and Shen, X. (2024) 'Epigenomic studies in sorghum reveal differential enrichment of multiple histone marks at clade A PP2C genes in response to drought', *New Crops*, 2, 100059.

- Hassan, M.A., Yang, M., Fu, L., Rasheed, A., Zheng, B., Xia, X., Xiao, Y. and He, Z. (2021) 'Accuracy assessment of plant phenotyping using machine learning and deep learning approaches', *Plant Methods*, 17, Article 45.
- He, Z., Zhang, P., Jia, H., Zhang, S., Nishawy, E., Sun, X. and Dai, M. (2024) 'Regulatory mechanisms and breeding strategies for crop drought resistance', *New Crops*, 1, 100029.
- Heslot, N., Jannink, J.L. and Sorrells, M.E. (2015) 'Perspectives for genomic selection applications and research in plants', *Crop Science*, 55(1), pp. 1–11.
- Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M. and others (2014) 'A reaction norm model for genomic selection using high-dimensional genomic and environmental data', *Theoretical and Applied Genetics*, 127, pp. 595–607.
- Kumar, A., Sharma, S. and Singh, P. (2020) 'Machine learning applications in agriculture: A review', *Computers and Electronics in Agriculture*, 176, 105676.
- Li, Y., Zhang, D., Wang, X. and others (2021) 'Machine learning approaches for crop improvement and stress resistance prediction', *Computers and Electronics in Agriculture*, 189, 106410.
- Li, B., Liu, R., Liu, J., Zhang, H., Tian, Y., Chen, T., Li, J., Jiao, F., Jia, T., Li, Y., Zhang, X., Li, H., Zhao, X., Galbraith, D.W. and Song, C.-P. (2024) 'ZmMYB56 regulates stomatal closure and drought tolerance in maize seedlings through the transcriptional regulation of ZmTOM7', *New Crops*, 1, 100012.
- Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C.M. and Martín-Vallejo, J. (2019) 'Multi-trait, multi-environment genomic prediction of plant traits using machine learning', *G3: Genes, Genomes, Genetics*, 9(12), pp. 3939–3952.
- Newton, A.C., Akar, T., Baresel, J.P., Bebeli, P.J., Bettencourt, E., Bladenopoulos, K.V. and others (2010) 'Cereal landraces for sustainable agriculture', *Agronomy for Sustainable Development*, 30, pp. 1–20.
- Pérez-Rodríguez, P. and de Los Campos, G. (2014) 'Genome-wide regression and prediction with the BGLR statistical package', *Genetics*, 198(2), pp. 483–495.
- Ramya, K.T., Singh, K. and Kumar, A. (2023) 'Artificial intelligence and machine learning approaches for crop improvement under abiotic stresses', *Plant Physiology and Biochemistry*, 194, pp. 200–214.
- Reynolds, M.P. and Langridge, P. (2016) 'Physiological breeding', *Current Opinion in Plant Biology*, 31, pp. 162–171.
- Shahhosseini, M., Hu, G. and Archontoulis, S.V. (2020) 'Coupling machine learning and crop modeling improves crop yield prediction', *Agricultural Systems*, 182, 102845.
- Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B. and Singh, A.K. (2020) 'Crop yield prediction integrating genotype and weather variables using deep learning', *Frontiers in Plant Science*, 11, 1311.
- Sui, J., Tian, H., Ding, Z. and Kong, X. (2024) 'Crop designs: the ideal root architecture for future crop breeding', *New Crops*, 1, 100030.
- Tardieu, F., Simonneau, T. and Muller, B. (2018) 'The physiological basis of drought tolerance in crop plants: A scenario-dependent approach', *Annual Review of Plant Biology*, 69, pp. 733–759.
- Van Eeuwijk, F.A., Bustos-Korts, D.V. and Malosetti, M. (2016) 'What should students in plant breeding know about the statistical aspects of genotype × environment interactions?', *Crop Science*, 56(5), pp. 2119–2140.
- Wang, G., Xuan, J., Wang, P. and Lu, J. (2024) 'Deep learning-based genotype-to-phenotype prediction for crop improvement', *Computational and Structural Biotechnology Journal*, 23, pp. 1120–1132.

- Wang, K., Long, J., Lei, J., Liang, Y., Ratnasekera, D., Qian, Q. and Zheng, X. (2025) 'Regional projections and sampling strategies for the global germplasm resources collection of *Oryza rufipogon* Griff.', *New Crops*, 100101.
- Wang, P., Li, Z., Li, H., Zhang, D., Wang, W., Xu, X., Xie, Q., Duan, Z., Xia, X., Guo, G., Shaheen, A., Zhou, Y., Wang, D., Guo, S., Hu, Z., Galbraith, D.W. and Song, C.-P. (2024) 'SMART CROPS', *New Crops*, 1, 100007.
- Wang, Y., Li, P., Sun, W. and Zhang, T. (2024) 'Plant cell walls: Emerging targets of stomata engineering to improve photosynthesis and water use efficiency', *New Crops*, 1, 100021.
- Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, B., Prasanna, B.M., Olsen, M.S., Wang, G. and Zhang, A. (2022) 'Enhancing genetic gain through genomic selection in crops', *Theoretical and Applied Genetics*, 135, pp. 1–18.
- Zhang, X., Chen, S., Sun, H., Wang, Y. and Shao, L. (2020) 'Changes in soil water and crop yield under drought stress', *Agricultural Water Management*, 230, 105963.
- Zhang, Z., Qu, Y., Ma, F. and others (2024) 'Integrating high-throughput phenotyping and genome-wide association studies for enhanced drought resistance prediction in wheat', *New Phytologist*, 243(5), pp. 1758–1775.
- Zhu, X., Liu, X., Wu, Q. and others (2025) 'Utilizing UAV-based high-throughput phenotyping and machine learning to evaluate drought resistance in wheat germplasm', *Computers and Electronics in Agriculture*, 237, 110602.
- Zhu, L., Tian, T., Du, X., Li, T. and Wan, X. (2026) 'Insights into the regulatory pattern classification for functional alleles underlying trait variation in maize and other crops', *New Crops*, 100110.
- Zhuang, Y., Li, X., Liu, L., Xu, R. and Zhang, D. (2026) 'Harnessing super pan-genomics for accelerated crop improvement', *New Crops*, 100109.
- Zuffo, A.M., Steiner, F., Aguilera, J.G., Teodoro, P.E., Teixeira, M.C.M. and Busch, A. (2021) 'Machine learning approaches for prediction of drought tolerance in crops', *Agronomy*, 11(10), pp. 1–15.