Iran Agricultural Research (2026) 45 (Special Issue) 72-95

Shiraz University

Iran Agricultural Research

Journal homepage: https://iar.shirazu.ac.ir



Research Article

Accelerating improvement of rapeseed for climate resilience and nutritional quality-related traits through machine learning-driven predictive modeling

Maryam Salami , Bahram Heidari*

Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, I. R. Iran

ARTICLE INFO

Keywords:
Machine learning
Multi-trait selection
Predictive breeding
Seed quality
Yield prediction

Received: 02 August 2026
Revised: 23 September 2026
Accepted: 27 September 2026

ABSTRACT- Traditional breeding approaches, largely focused on single-trait selection, are proving insufficient to address the polygenic nature of drought tolerance. Here, we present an integrated modeling framework that demonstrates how machine learning (ML) can help accelerate rapeseed (*Brassica napus* L.) breeding by enabling predictive modeling of complex trait networks, integrating multi-dimensional phenotypic and metabolomic data, and dissecting genotype × environment interactions (GEI). A key component is a multiblock integration workflow that combines phenotypic, metabolomic, and Fourier-transform near-infrared (FT-NIR) data via early, intermediate, and late fusion. We critically analyzed the application of ML algorithms—including Partial Least Squares Regression (PLSR), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)—across three interconnected breeding domains. Our study demonstrated that ML models achieved high predictive accuracy (R^2 = 0.79–0.91) for seed yield and quality traits, identified chlorophyll content (SPAD) as the most important predictor of yield under optimal and stress conditions, and uncovered non-linear interactions, such as flowering date × plant height interactions, that classical linear models fail to detect. Integrating ML with classical stability analyses and multi-trait selection approaches and the economically weighted selection index (EWSI) provided a unified framework for identifying superior genotypes merging high-yield traits, metabolic quality, and broad environmental adaptability. Our results indicated that ML-enhanced breeding marks a radical transition away from reactive, trait-based selection to predictive and data-intensive breeding improvement, offering a practical and powerful framework for developing climate-resilient cultivars with optimized agronomic and nutritional values.

INTRODUCTION

Rapeseed (*Brassica napus* L.) ranks among the world's most important oilseed crops, providing approximately 13% of the global vegetable oil supply and serving as a critical source of high-protein animal feed (Moreno, 2025; Liao et al., 2026). In 2023, the worldwide output of rapeseed stood at 92 million tons, with primary growing areas spanning the European Union, Canada, China, India, and Australia (Lu et al., 2019). Being an allotetraploid, this species carries two subgenomes inherited from *Brassica rapa* (AA) and *Brassica oleracea* (CC), and its ~1.1 Gb genome exhibits extensive structural rearrangements and recombination, positioning it as an outstanding subject for exploring polyploidy dynamics in angiosperms (Luo et al., 2022). Moreover, the species has given rise to three distinct ecotypes—spring, winter, and semi-winter—which differ in their cold-temperature vernalization needs, thereby allowing adaptation across a wide array of agro-ecological zones (Luo et al., 2022; Salami et al., 2025; Yin et al., 2026). However, rapeseed cultivation is increasingly challenged by drought stress, with climate models predicting more frequent and severe water deficits in major production regions (Bouchyoua et al., 2026). According to the Intergovernmental Panel on Climate Change (IPCC), mean global temperatures are expected to climb by 1.5–2.0 °C relative to pre-industrial baselines by the middle of this century, a trend that will be accompanied by increasingly unpredictable rainfall patterns and more frequent extreme weather events (IPCC, 2021). Abiotic constraints, notably water deficit and excessive heat, pose significant obstacles to rapeseed development and productivity on a global scale, thus necessitating an in-depth exploration of resilience mechanisms (Shahsavari et al., 2025).

Drought stress occurring during the critical reproductive stages, particularly flowering and seed filling, profoundly compromises both the yield and overall seed quality of

*Corresponding Author: Professor, Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, I. R. Iran
E-mail address: bheidari@shirazu.ac.ir
DOI:10.22099/iar.2026.56795.1774

Copyright© 2026 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC 4.0), which permits unrestricted use, distribution, and reproduction in any medium, as long as the original authors and source are cited. No permission is required from the authors or the publisher.

rapeseed. Post-flowering water deficit alone can reduce the final seed yield by 70–78%, with the extent of loss heavily dependent on the specific genotype and the severity of the stress imposed (Bouchyoua et al., 2026). Beyond these substantial yield penalties, drought significantly alters seed oil composition; it typically elevates the relative content of saturated fractions (including palmitic and stearic acids), whereas it concurrently diminishes the content of the polyunsaturated linoleic acid (C18:2), thereby impairing the nutritional quality of the extracted oil (Secchi et al., 2023). At the biochemical level, drought also triggers the accumulation of stress-protective secondary metabolites, including anthocyanins and specific derivatives such as peonidin, which show increases exceeding 144% under severe conditions (Chen et al., 2022) and structural carbohydrates (hemicellulose, lignin, and cellulose) (Tian et al., 2022). The physiological basis of these responses involves complex reprogramming of carbon allocation: under water limitation, plants redirect photosynthetic assimilates from storage compounds (oil and protein) toward stress-protective metabolites and cell wall reinforcement, representing a fundamental trade-off between survival and productivity (Schwender and Hay, 2012).

Traditional breeding for drought tolerance has focused primarily on yield-based selection across multiple environments, an approach that has yielded incremental but limited progress. Several intertwined obstacles hinder conventional breeding efforts. First, drought tolerance is a polygenic characteristic controlled by an array of quantitative trait loci (QTLs) with minor additive effects, which complicates its improvement via basic Mendelian selection (Gill et al., 2026). Unlike monogenic traits, drought tolerance involves the coordinated expression of hundreds or thousands of genes, each contributing modestly to the overall phenotype (Salami et al., 2024a; Salami et al., 2024b). Such genomic complexity demands population-based methodologies capable of aggregating numerous weak-effect alleles simultaneously (Gill et al., 2026). Second, the strong genotype × environment interaction (GEI) characteristic of drought response means that genotypes performing well in one environment often fail in others, complicating cultivar recommendation (Kakaei et al., 2024). A genotype may express superior drought tolerance under moderate stress but collapse under severe stress, or vice versa. This "crossover" phenomenon constitutes a shifting target for plant breeders, necessitating multi-environment trials conducted over multiple cropping cycles and field sites (Kakaei et al., 2024; Gill et al., 2026).Third, the fundamental antagonism between productivity and stress adaptation—where genes that boost drought hardiness frequently impose yield losses under favorable conditions—represents a core breeding challenge (Chen et al., 2025; Fletcher et al., 2015). This antagonistic pleiotropy means that selection for stress tolerance can inadvertently reduce yield potential, and vice versa (Fletcher et al., 2015). Breaking this trade-off requires identification of genotypes that maintain high productivity across both favorable and adverse environments, a rare combination in most germplasm collections (Chen et al., 2025). Fourth, traditional breeding has largely overlooked the temporal dimension of drought response: stress timing during seed development triggers fundamentally different metabolic and agronomic outcomes (Kakaei et al., 2024; Salami et al., 2026a). Early drought (20 days after pollination; DAP) primarily compromises yield components (seed number and size), while late drought (40 DAP) maximizes the accumulation of health-promoting phytochemicals such as anthocyanins and flavonoids (Salami et al., 2026a). This temporal distinction has profound implications for breeding strategy, yet most breeding programs evaluate stress responses at single time points (Kakaei et al., 2024). Fifth, traditional statistical techniques cannot effectively process the complex, multidimensional datasets arising from contemporary phenotyping and metabolomics systems (Niazian and Niedbała, 2020). Classical approaches such as analysis of variance (ANOVA), linear regression, and principal component analysis (PCA), while valuable for initial exploration, cannot capture the complex, non-additive interactions that characterize plant stress responses (Niazian and Niedbała, 2020).

Machine learning (ML) offers a transformative approach to these challenges (Niazian and Niedbała, 2020). Unlike classical statistical methods that assume linear relationships among model variables and require pre-specified model structures, ML algorithms can automatically learn complex patterns from data, model non-linear interactions, and make predictions with high accuracy (Breiman, 2001; Sarker, 2021). In plant breeding and agricultural research, ML methods have been increasingly applied for yield prediction, genotype classification, phenomic analysis, and stress tolerance evaluation (van Klompenburg et al., 2020). Several comprehensive literature syntheses have reported the growing integration of ML into crop improvement programs (van Klompenburg et al., 2020). Ensemble models based on decision trees—specifically, Random Forest (RF) and eXtreme Gradient Boosting (XGBoost)—dominate the field, owing to their interpretability and resilience with modest-sized datasets (Breiman, 2001; Chen and Guestrin, 2016). Although neural network architectures achieve high accuracy, they are predominantly applied to larger sample sizes and are limited by substantial computational demands (LeCun et al., 2015). Importantly, ML-based methodologies tend to surpass classical genomic selection for traits with low heritability, whereas genomic selection maintains its edge for highly heritable traits. Furthermore, the fusion of environmental profiles (envirotyping) with diverse omics data for predicting genotype-by-environment interactions has emerged as a notable priority (Xu et al., 2016). In this study, we present a comprehensive synthesis of ML applications in rapeseed breeding, with particular emphasis on drought resilience and seed quality improvement, and propose a unified framework for their integration. We draw on recent studies that collectively establish the foundation for ML-enhanced rapeseed breeding and critically examine how these methods can be integrated into a cohesive pipeline. Specifically, this study addresses the following questions: How can ML algorithms be applied to predict seed compositional traits (fatty acids, phenolic compounds, and fiber fractions) from high-throughput phenotyping data? What ML approaches are most effective for multi-trait selection that balances yield, drought resilience, and nutritional quality? How does ML enhance the interpretation of GEI and guide genotype deployment across heterogeneous environments? This manuscript explores the primary challenges and limitations associated with the application of ML in rapeseed breeding. It discusses

strategies to address these challenges and presents ML methodologies specifically tailored for rapeseed breeding. The study evaluates the effectiveness of these methods in predicting seed quality and illustrates their application in analyzing GEI, thereby facilitating the development of an integrated breeding pipeline.

MATERIALS AND METHODS

Plant material and experimental design

Overall, 119 rapeseed genotypes (breeding lines, hybrids, and commercial varieties) were obtained from the Seed and Plant Improvement Institute (SPII) in Iran (Table 1). Field trials were conducted at the Shiraz University research farm over four cropping seasons (2017–2018, 2018–2019, 2020–2021, and 2024–2025); the 2019–2020 season was excluded due to the severe spring frost. The soil was silty loam. Crop evapotranspiration (ETc) was calculated according to the Food and Agriculture Organization (FAO)-56 Penman-Monteith procedure (Guo et al., 2020), ranging from 3.5 mm day^{-1} (early) to 7.2 mm day^{-1} (mid-season), with a total seasonal requirement of approximately 450–500 mm. The experiment followed a split-plot arrangement within a randomized complete block framework, featuring three replications. The main plots were allocated to two water treatments: well-watered (WW) and drought stress (DS). Subplots contained 119 genotypes in an incomplete 11 × 11 lattice, each measuring 1 m² with 0.5 m gaps. Each subplot contained four 1-meter rows, each supporting 25 individuals at 4 cm intervals. Sowing occurred on September 17 each season. Each combination of irrigation and season represented a distinct environment, yielding eight environments (E1–E8) for GEI analysis. Nutrient application comprised nitrogen, phosphorus, and potassium at respective rates of 250, 100, and 100 kg ha^{-1}; P and K were incorporated before sowing, whereas N was supplied as a top-dressing across several growth phases. Under WW, irrigation was applied continuously from sowing to physiological maturity, while under DS, irrigation was maintained until approximately 50% pod development and then stopped (Ozer, 2003). Thus, WW received nine irrigations, and DS received six. Soil moisture was monitored weekly by the gravimetric method (Ruszczak and Boguszewska-Mańkowska, 2022) from 0–30 cm depth. WW maintained 22–28% moisture, while DS declined progressively after flowering to approximately 12% by 40 DAP, confirming a severe water deficit. Weed control involved Treflan® HFP herbicide (3 L ha^{-1}) and hand weeding; aphids were managed with Pirimor 50 pesticide (2 L ha^{-1}). Harvest occurred in early July when pods on the terminal raceme turned creamy white.

Phenotypic trait measurement

Flowering time (FD) was scored at 50% anthesis (Matar et al., 2021). Silique development (SD) and maturity (MD) were evaluated similarly. Plant height (PH) was measured at maturity. For branching number (BN), silique length (SL), seeds per silique (SPS), seed weight (SW), and thousand-seed weight (TSW), ten plants were randomly selected from the middle rows. Chlorophyll content was assessed using a portable SPAD-502 chlorophyll meter (Minolta Camera Co., Osaka, Japan). Measurements were conducted during the stem elongation to inflorescence emergence stage (BBCH 30–59) between 10:00 and 14:00 hours under clear sky conditions to minimize environmental variability (Ling et al., 2011). For each plant, one fully expanded upper leaf was selected, avoiding the leaf margins, and three readings were taken per leaf. Ten plants were randomly chosen from the middle two rows of each plot, and the average of all measurements (n = 30 readings per plot) was calculated to determine the SPAD value for that plot. This procedure was consistently applied to both well-watered and drought-stressed plots to ensure comparability. At harvest, the central two rows per plot were collected to assess plant weight (PW), seed yield (SY, kg ha^{-1}), and harvest index (HI = (SY/total biological yield) × 100) (Bennett et al., 2017).

Metabolomic profiling

Phenolics were isolated from freeze–dried seed tissue (100 mg) using acidified methanol/water/formic acid (70:29:1, v/v/v) (Shi et al., 2022), subjected to sonication, followed by centrifugation at 10,000 × g for 15 min and passage through a 0.45 µm syringe filter. For each genotype × environment × DAP combination, three biological replicates were analyzed, with each biological replicate subjected to duplicate injections for technical replication. Chromatographic separation was conducted using high-performance liquid chromatography with diode array detection (HPLC-DAD) on an Agilent 1260 Infinity II instrument equipped with a Phenomenex Luna C18 column (250 mm × 4.6 mm i.d., 5 µm particle size) maintained at 30 °C. The mobile phase comprised 0.1% aqueous formic acid (solvent A) and acetonitrile (solvent B), delivered at a flow rate of 1.0 mL min^{-1} according to the following gradient profile: 0–5 min, 5% B; 5–40 min, linear increase from 5% to 40% B; 40–45 min, ramp to 95% B; 45–48 min, hold at 95% B; 48–50 min, return to 5% B. Flavonols and flavones were monitored at 280 and 360 nm (Lin et al., 2007); hydroxycinnamic acids at wavelengths of 310–340 nm; anthocyanins (extracted with 1% HCl-methanol) at 520 nm (Brown and Shipley, 2011). Identification used authentic standards; quantification was performed via external calibration curves (Lima et al., 2017; Khattab et al., 2010). Metabolomic profiling was conducted on samples collected at both 20 and 40 DAP to evaluate the dynamic changes in phenolic compounds, fatty acids, and structural components throughout seed development under different water regimes (Salami et al., 2026a).

Seed composition was analyzed by Fourier-transform near-infrared (FT-NIR) spectroscopy (12,000–4,000 cm^{-1}, 64 scans) with Standard Normal Variate (SNV) and detrending preprocessing (Barragán-Hernández et al., 2020). For each genotype × environment combination, three biological replicates were used, and each spectrum was collected as the average of 64 scans. Partial least squares (PLS) calibrations covered moisture, protein, oil, total dietary fiber, and ash (Veser et al., 2025; Sharma et al., 2024; Bakri et al., 2023; Ferjancic et al., 2022; Giménez-Berenguer et al., 2025), fatty acids (Danudol and Judprasong, 2022), and the neutral detergent fiber (NDF), acid detergent fiber (ADF), and acid detergent lignin (ADL) fiber components (Van Soest et al., 1991; Fahey et al., 2019). Hemicellulose and cellulose were obtained as NDF–ADF and ADF–ADL, respectively (Zhang et al., 2015). Model performance was assessed using the coefficient of

determination for prediction (R^2p) and root mean square error of prediction (RMSEP) (Uche et al., 2024). Total sugars were measured using the phenol–sulfuric acid protocol (Yue et al., 2022).

ML approaches

Partial least squares regression (PLSR) represents a commonly employed technique that extends principal component regression through the extraction of latent components that optimize the covariance structure between predictor variables and outcomes (Wold et al., 2001). The algorithm extracts orthogonal latent components that simultaneously capture the variance in predictors and maximize the correlation with responses, making it robust to multicollinearity (Geladi and Kowalski, 1986). PLSR is particularly valuable for analyzing high-dimensional, collinear data, such as near-infrared (NIR) spectra, especially in situations where the predictor count exceeds the sample size (Blanco and Villarroya, 2002). The PLSR model is expressed by Eqs. (1) and (2).

$$X = TP^T + E \quad \text{Eq. (1)}$$

$$Y = UQ^T + F \quad \text{Eq. (2)}$$

where X is the predictor matrix, Y is the response matrix, T and U are score matrices, P and Q are loading matrices, and E and F are residual matrices (Tenenhaus et al., 2005). The optimal component number is generally determined via cross-validation, which aims to minimize the RMSEP (Hastie et al., 2009). PLSR models were developed using the statistically inspired modification of partial least squares (SIMPLS) algorithm.

Variable Importance in Projection (VIP) score for each predictor is calculated using Eq. (3).

$$VIP_j = \sqrt{[\sum_k (SSY_k \times wk_j^2) / (\sum_k SSY_k)] \times p} \quad \text{Eq. (3)}$$

where SSY_k is the sum of squares of Y explained by component k, wk_j is the weight of variable j in component k, and p is the number of predictors. VIP scores ≥ 1.0 are considered significant predictors (Wold et al., 2001).

RF is an ensemble technique that generates numerous decision trees using bootstrap samples and random subsets of predictors at each node (Breiman, 2001). Prediction aggregation via averaging (for regression) or majority voting (for classification) lowers variance while leaving bias unchanged (Breiman, 2001). RF provides built-in feature importance measures through permutation importance (the decrease in prediction accuracy when a variable is randomly permuted) and Gini importance (the total decrease in node impurity from splits on that variable). RF is robust to outliers and collinearity, making it suitable for diverse agricultural datasets (Breiman, 2001).

XGBoost is a sequential ensemble approach that constructs decision trees successively, with each subsequent tree addressing the residuals of its predecessors (Chen and Guestrin, 2016). The method minimizes a regularized loss function, as given in Eq. (4).

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad \text{Eq. (4)}$$

where l is a differentiable convex loss function, Ω is a regularization term that penalizes model complexity, and f_k represents the k-th tree (Chen and Guestrin, 2016). The regularization term $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ controls the number of leaves (T) and the magnitude of leaf weights (w), preventing overfitting (Hastie et al., 2009). XGBoost inherently manages missing values, offers feature importance rankings, and delivers highly competitive results on structured datasets (Chen and Guestrin, 2016).

Support Vector Machines (SVM) are kernel-based methods that find optimal hyperplanes for classification or regression by maximizing margins between classes (Cortes and Vapnik, 1995). Kernel functions—including linear, polynomial, radial basis function, and sigmoid—transform input data into high-dimensional spaces where linear separability is achievable (Bishop, 2006). The radial basis function kernel offers the greatest flexibility is given by Eq. (5).

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \quad \text{Eq. (5)}$$

where γ controls the influence of individual training examples (Bishop, 2006). SVM performs well with high-dimensional data but requires careful parameter tuning through cross-validation (Chang et al., 2025).

Artificial neural networks (ANNs) are architectures comprised of stacked layers of interconnected nodes that acquire non-linear mappings via the backpropagation algorithm (Rumelhart et al., 1986). ANNs can approximate any continuous function given sufficient data and an appropriate architecture, making them powerful for complex prediction tasks (Hornik et al., 1989). A standard ANN architecture consists of an input layer, at least one intermediate layer employing non-linear transformations (e.g., ReLU, sigmoid, and tanh), along with an output layer (Goodfellow et al., 2016). The network learns by minimizing a loss function through gradient descent and backpropagation of errors (Rumelhart et al., 1986). ANNs can approximate any continuous function given sufficient data and an appropriate architecture (Hornik et al., 1989).

Gaussian process regression (GPR) constitutes a non-parametric Bayesian approach to regression that operates via a probabilistic, kernel-based framework (Rasmussen, 2003). GPR establishes a prior distribution across functions and subsequently updates this distribution based on observed data, yielding a posterior distribution (Rasmussen, 2003). The model is defined in Eq. (6).

$$f(x) \sim GP(m(x), k(x, x')) \quad \text{Eq. (6)}$$

where $f(x)$ is the latent function, $m(x)$ is the mean function (commonly assumed to be zero after data standardization), Gaussian process (GP) is a Gaussian Process, and $k(x, x')$ is the covariance (kernel) function that encodes the similarity between input points. (Rasmussen, 2003). GPR supplies point predictions alongside uncertainty bounds (credible intervals), making it a powerful tool for breeding choices under risk (Wang et al., 2025; Friedli et al., 2026).

Table 1. Germplasm information for the rapeseed (*Brassica napus*) accessions analyzed

Code	Accession	Origin	Breeding status	Seasonal type[a]	Pollination status	Code	Accession	Origin	Breeding status	Seasonal type	Pollination status
G01	Okapi	France	Cultivar	Winter	Open pollinating	G31	Es neptune	France	Hybrid	Winter	-
G02	Ahmadi	Iran	Cultivar	Winter	Open pollinating	G32	Kodiac	Germany	Cultivar	Winter	Open pollinating
G03	Slm046	France	Breeding line	Winter	Open pollinating	G33	Tassilo	Germany	Hybrid	Winter	-
G04	Sw102	Iran	Breeding line	Winter	Open pollinating	G34	Triangle	Germany	Hybrid	Winter	-
G05	l72	Iran	Breeding line	Winter	Open pollinating	G35	Traviata	Germany	Hybrid	Winter	-
G06	963	Iran	Breeding line	Winter	Open pollinating	G36	Diffusion	Germany	Cultivar	Winter	Open pollinating
G07	957	Iran	Breeding line	Winter	Open pollinating	G37	Zorica	Serbia	Cultivar	Winter	Open pollinating
G08	Expower	America	Hybrid	Winter	-	G38	Zelatna	Serbia	Cultivar	Winter	Open pollinating
G09	Dk extorm	America	Hybrid	Winter	-	G39	Es Hydromel	France	Hybrid	Winter	-
G10	Dk excalibur	America	Hybrid	Winter	-	G40	Es Alonso	France	Hybrid	Winter	-
G11	Es vatalie	France	Hybrid	Winter	-	G41	Es Darko	France	Hybrid	Winter	-
G12	Es Danube	France	Hybrid	Winter	-	G42	Es lauven	France	Hybrid	Winter	-
G13	Es Artist	France	Hybrid	Winter	-	G43	Es mercure	France	Hybrid	Winter	-
G14	Sy vesuvio	America	Hybrid	Winter	-	G44	Okapi-original-mashkad	France	Hybrid	Winter	-
G15	Sy melea	America	Hybrid	Winter	-	G45	Garou	Germany	Cultivar	Winter	-
G16	Sy Harnus	America	Hybrid	Winter	-	G46	Rohan	Germany	Cultivar	Winter	Open pollinating
G17	1110	Iran	Breeding line	Winter	Open pollinating	G47	Gk Reka	Hungary	Cultivar	Winter	Open pollinating
G18	1112	Iran	Breeding line	Winter	Open pollinating	G48	Gk csenge	Hungary	Cultivar	Winter	Open pollinating
G19	1093	Iran	Breeding line	Winter	Open pollinating	G49	Z-800-3	Iran	Breeding line	Winter	Open pollinating
G20	1096	Iran	Breeding line	Winter	Open pollinating	G50	Z-800-6	Iran	Breeding line	Winter	Open pollinating
G21	1114	Iran	Breeding line	Winter	Open pollinating	G51	Z-900-3	Iran	Breeding line	Winter	Open pollinating
G22	1204	Iran	Breeding line	Winter	Open pollinating	G52	Z-900-6	Iran	Breeding line	Winter	Open pollinating
G23	1206	Iran	Breeding line	Winter	Open pollinating	G53	Z-900-7	Iran	Breeding line	Winter	Open pollinating
G24	Talaye	Germany	Cultivar	Winter	Open pollinating	G54	Z-900-8	Iran	Breeding line	Winter	Open pollinating
G25	Opera	Germany	Cultivar	Winter	Open pollinating	G55	Z-900-9	Iran	Breeding line	Winter	Open pollinating
G26	Licorol	Germany	Cultivar	Winter	Open pollinating	G56	Z-900-10	Iran	Breeding line	Winter	Open pollinating
G27	Modena	Russia	Cultivar	Winter	Open pollinating	G57	T-800-1	Iran	Breeding line	Winter	Open pollinating
G28	Zarfam	Iran	Cultivar	Faculative	Open pollinating	G58	T-800-6	Iran	Breeding line	Winter	Open pollinating
G29	Gabriella	Hungary	Cultivar	Winter	Open pollinating	G59	Cod 769	Iran	Breeding line	Winter	Open pollinating
G30	Brutus	Germany	Hybrid	Winter	-	G60	Exp-800-3	Iran	Breeding line	Winter	Open pollinating

Table 1. Continued

Code	Accession	Origin	Breeding status	Seasonal type	Pollination status	Code	Accession	Origin	Breeding status	Seasonal type	Pollination status
G61	Exp-900-1	Iran	Breeding line	Winter	Open pollinating	G91	B20-p15 R13 S2-13	Iran	Breeding line	Winter	Open pollinating
G62	Code 12	Iran	Breeding line	Winter	Open pollinating	G92	B17-p12 R1 S2-1	Iran	Breeding line	Winter	Open pollinating
G63	Code 21	Iran	Breeding line	Winter	Open pollinating	G93	B17-p12 R9 S2-1	Iran	Breeding line	Winter	Open pollinating
G64	Code 20	Iran	Breeding line	Winter	Open pollinating	G94	B19-p14 R12 S2-1	Iran	Breeding line	Winter	Open pollinating
G65	Code 18	Iran	Breeding line	Winter	Open pollinating	G95	B20-p15 R14 S2-3	Iran	Breeding line	Winter	Open pollinating
G66	Code 255	Iran	Breeding line	Winter	Open pollinating	G96	B17-p12 R2 S2-1	Iran	Breeding line	Winter	Open pollinating
G67	Code 262	Iran	Breeding line	Winter	Open pollinating	G97	B19-p14 R8 S2-1	Iran	Breeding line	Winter	Open pollinating
G68	Code 763	Iran	Breeding line	Winter	Open pollinating	G98	B15-p11 R15 S2-5	Iran	Breeding line	Winter	Open pollinating
G69	Code 768	Iran	Breeding line	Winter	Open pollinating	G99	B17-p12 R3 S2-1	Iran	Breeding line	Winter	Open pollinating
G70	Code 751	Iran	Breeding line	Winter	Open pollinating	G100	B12-p9 R2 S2-3	Iran	Breeding line	Winter	Open pollinating
G71	Code 268	Iran	Breeding line	Winter	Open pollinating	G101	B19-p14 R1 S2-2	Iran	Breeding line	Winter	Open pollinating
G72	Code 270	Iran	Breeding line	Winter	Open pollinating	G102	B19-p14 R1 S2-1	Iran	Breeding line	Winter	Open pollinating
G73	Code 655	Iran	Breeding line	Winter	Open pollinating	G103	B16-p11 R16 S2-1	Iran	Breeding line	Winter	Open pollinating
G74	Code 914	Iran	Breeding line	Winter	Open pollinating	G104	B19-p14 R12 S2-6	Iran	Breeding line	Winter	Open pollinating
G75	Code 253	Iran	Breeding line	Winter	Open pollinating	G105	B12-p9 R1 S2-1	Iran	Breeding line	Winter	Open pollinating
G76	Code 266	Iran	Breeding line	Winter	Open pollinating	G106	B17-p12 R13 S2-1	Iran	Breeding line	Winter	Open pollinating
G77	Code 263	Iran	Breeding line	Winter	Open pollinating	G107	B17 okapi -1	Iran	Breeding line	Winter	Open pollinating
G78	Code 770	Iran	Breeding line	Winter	Open pollinating	G108	B17-p15 R13 S2-8	Iran	Breeding line	Winter	Open pollinating
G79	Code 264	Iran	Breeding line	Winter	Open pollinating	G109	Code-3-2	Iran	Breeding line	Winter	Open pollinating
G80	Code 901	Iran	Breeding line	Winter	Open pollinating	G110	H4815	Australia	Cultivar	Spring	Open pollinating
G81	Code 23	Iran	Breeding line	Winter	Open pollinating	G111	Dalgan	Iran	Cultivar	Spring	Open pollinating
G82	Code 655	Iran	Breeding line	Winter	Open pollinating	G112	Licord	Germany	Cultivar	Winter	Open pollinating
G83	Code 1	Iran	Breeding line	Winter	Open pollinating	G113	H308	Australia	Cultivar	Spring	Open pollinating
G84	Code 3	Iran	Breeding line	Winter	Open pollinating	G114	Zafar	Iran	Cultivar	Spring	Open pollinating
G85	Code 11	Iran	Breeding line	Winter	Open pollinating	G115	Sarigol	Iran	Cultivar	Spring	Open pollinating
G86	Code 19	Iran	Breeding line	Winter	Open pollinating	G116	H401	Australia	Cultivar	Spring	Open pollinating
G87	Cooper	Germany	Breeding line	Winter	Open pollinating	G117	Karaj1	Iran	Cultivar	Winter	Open pollinating
G88	Adriana	Germany	Breeding line	Winter	Open pollinating	G118	RGS	Germany	Cultivar	Spring	Open pollinating
G89	Code 767	Iran	Breeding line	Winter	Open pollinating	G119	H420	Australia	Cultivar	Spring	Open pollinating
G90	Code 752	Iran	Breeding line	Winter	Open pollinating						

[a]Seasonal types were distinguished by their flowering times at Research Farm of Plant Production and Genetics, Shiraz, Iran.

Integration of phenotypic, metabolomic, and FT-NIR data for ML

To elucidate the integration of various data types, we describe the workflow for each genotype × environment × DAP observation. Phenotypic traits (agronomic, phenological, and SPAD) were quality-controlled, missing values were imputed, and the data were converted to best linear unbiased predictions (BLUPs)/best linear unbiased estimates (BLUEs) to account for genotype × environment interactions. Metabolomic data, including phenolics, fatty acids, structural carbohydrates, and sugars, were first imputed for missing values, then log-transformed when skewed, autoscaled, and batch-corrected to remove technical variation (Tseng et al., 2025). FT-NIR spectra were preprocessed using SNV, detrending, and derivatives, and then compressed into PLS latent components and PLS-derived predictions for key quality traits (Ren et al., 2025). These three preprocessed blocks were integrated using three complementary strategies: (i) early fusion, in which blocks were concatenated into a single feature matrix, (ii) intermediate fusion, in which block-specific principal components and PLS scores were combined into a reduced feature set, and (iii) late fusion/stacking, in which separate models were trained for each block and their out-of-fold predictions were combined by a meta-learner (Bersanelli et al., 2016). Environmental covariates encompassing water regime, season, days after pollination, temperature, rainfall, evapotranspiration, and soil moisture were incorporated across all methodological approaches to explicitly characterize genotype-by-environment interactions. All data preprocessing and feature selection procedures were executed within a nested cross-validation framework employing genotype-specific and environment-specific holdout sets to mitigate data leakage concerns (Novielli et al., 2024). Model interpretation was conducted through VIP scores, permutation-based importance metrics, XGBoost gain measurements, and SHapley Additive exPlanations (SHAP) value analysis to evaluate block-level contributions (Li et al., 2026). A comprehensive schematic representation of this analytical workflow is presented in Fig. 1.

Model implementation and hyperparameter selection

All analyses were performed using R version 4.3.2. ML models were implemented using the following packages: PLSR through the pls package (version 2.8-3) in R; RF, XGBoost, SVM, and GPR through the caret (version 6.0-94), randomForest (version 4.7-1.1), xgboost (version 2.0.3), e1071 (version 1.7-14), and kernlab (version 0.9-32) packages in R (Chen & Guestrin, 2016). Hyperparameters were optimized using nested cross-validation, in which all predefined combinations of parameter values were evaluated in the inner loop. For RF, the number of trees (ntree = 500–2000) and the number of variables randomly sampled at each split (mtry = 2–10) were varied. For XGBoost, the learning rate (0.01–0.3), maximum depth (3–10), minimum child weight (1–10), subsample ratio (0.6–1), and number of boosting rounds (100–1000) were tested. For SVM, the cost parameter (C = 0.1–100) and kernel width (sigma = 0.001–1) were evaluated. For ANN, the number of hidden layers (1–3), neurons per layer (16–128), learning rate (0.0001–0.01), and batch size (16–64) were explored (Seeger, 2004). For GPR, the kernel type and length scale were adjusted. For PLSR, the optimal number of latent components was selected using leave-one-out cross-validation. Feature selection was performed using a combination of filter and wrapper methods. PLSR VIP scores (VIP ≥ 1.0) and RF permutation importance were used as filter methods. Recursive feature elimination (RFE) with RF and XGBoost was used as a wrapper method. All feature selections were performed inside the cross-validation folds to prevent biased estimates (Cawley & Talbot, 2010).

Classical stability evaluation alongside multi-trait selection frameworks

Although not strictly ML, traditional stability methods are crucial for understanding GEI and are increasingly paired with ML techniques (Pour-Aboughadareh et al., 2026; Qasemi et al., 2022). The AMMI (Additive Main Effects and Multiplicative Interaction) framework integrates ANOVA with PCA to decompose data into additive components and multiplicative interaction structures. (Gauch, 2006; Pour-Aboughadareh et al., 2026). The genotype main effect and genotype × environment interaction (GGE) biplot analysis centers on the combined effects of genotype and GEI, providing graphical visualization of genotype performance, stability, and mega-environment delineation (Yan et al., 2000; Yan and Kang, 2003). The "which-won-where" polygon view identifies the best-performing genotype in each environmental sector by drawing a polygon linking the genotypes most distant from the plot origin, then perpendicular bisectors divide the plot into sectors (Yan and Tinker, 2006; Pour-Aboughadareh et al., 2026).

Multi-trait selection indices facilitate the concurrent enhancement of multiple characteristics, thereby addressing the inherent trade-offs encountered in plant breeding (Olivoto et al., 2019; Zahedi et al., 2026). The Multi-Trait Genotype-Ideotype Distance Index (MGIDI) uses factor analysis to reduce high-dimensional trait data into uncorrelated factors, and then selects genotypes closest to a defined ideotype (Olivoto et al., 2019; Olivoto and Nardino, 2021; Barajehfard et al., 2026; Karegar et al., 2026). The MGIDI approach involves three main steps: factor analysis with varimax rotation to identify independent selection dimensions; definition of an ideotype with ideal scores for each factor; and computation of the Euclidean distance from each genotype to the target ideotype. The Economic Weighted Selection Index (EWSI) integrates economic weights with predicted trait values, converting agronomic performance into economically interpretable selection criteria (Hazel, 1943). The Multi-Trait Stability Index (MTSI) simultaneously evaluates productivity and stability across multiple traits using the WAASB (Weighted Average of Absolute Scores of Best Linear Unbiased Predictions) approach (Olivoto et al., 2019). A selection intensity of 15% was typically applied to identify superior genotypes.

Model evaluation and validation

Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), ratio of performance to deviation (RPD), and for classification tasks, accuracy, sensitivity/recall, and area under the receiver operating characteristic curve (AUC). k-fold

cross-validation procedures (typically 10-fold, with 3–5 repetitions) were employed to tune model complexity and derive unbiased estimates of performance (Hastie et al., 2009). Training and test sets were independent with no data leakage, typically using 70/30 or 80/20 splits for training/test (Bishop, 2006); stratified splitting maintained the distribution of target variables across subsets. Broad-sense heritability (h^2) estimates were calculated for metabolic, structural, and agronomic traits to guide model development (Heckerman et al., 2016).

RESULTS AND DISCUSSION

Assessment of ML model predictive capability

The predictive performance of the ML models was first evaluated using PLSR models developed for seed yield prediction under both well-watered and drought-stressed conditions, utilizing FT-NIR spectral information and leave-one-out cross-validation (LOOCV) to identify the ideal latent component count. As shown in Table 2, the PLSR models achieved training R^2 values ranging from 0.82 to 0.86 and test R^2 values from 0.78 to 0.82, with all ratios of performance to deviation (RPD) estimates consistently exceeding 2.0, reflecting outstanding predictive capacity across all water regimes. Under well-watered conditions, the optimal model required eight latent components and attained a test R^2 of 0.79 with RMSE values of 490 and 738 kg ha^{-1} at 20 and 40 DAP, respectively. In contrast, under drought stress, the models required 15 latent components to capture the added complexity of stress responses, achieving test R^2 values of 0.82 at 20 DAP and 0.78 at 40 DAP, with corresponding RMSE values of 220 and 250 kg ha^{-1}, respectively. The highest accuracy was recorded under drought at 20 DAP (RPD = 2.92), suggesting that stress conditions produce more consistent yield responses that are more readily predicted from spectral data, while Fig. 2 visually confirms the close concordance between observed and predicted data for both training and testing sets under this condition. This enhanced predictive performance under stress may be attributed to the fact that drought imposes stronger physiological constraints that reduce the influence of other environmental and genetic sources of variation, thereby producing more predictable yield responses that spectral data can effectively capture. The requirement for nearly twice as many latent components under drought (15 versus 8) reflects the greater complexity of stress responses and the more heterogeneous nature of the data, indicating that drought stress introduces additional physiological and metabolic dimensions that must be modeled to achieve accurate predictions (Salami et al., 2026a). Furthermore, the consistently high RPD values across all conditions confirm that PLSR models based on FT-NIR spectra can reliably predict seed yield across contrasting water regimes, providing a time-efficient, non-destructive protocol for yield determination that could be integrated into high-throughput breeding pipelines, consistent with findings from previous studies that have demonstrated the utility of NIR spectroscopy for predicting seed composition and agronomic traits in oilseed crops (Blanco and Villarroya, 2002; Beć et al., 2020).

Comparative performance of supervised learning algorithms for yield classification

Beyond PLSR, a comprehensive comparative evaluation of supervised learning algorithms was performed to classify high-versus low-yielding genotypes under drought stress, and the results, presented in Fig. 3, revealed that XGBoost achieved the highest classification accuracy, exceeding 80% and consistently outperforming support vector machines (with linear kernel achieving 80–86% accuracy) and ANNs. The enhanced predictive capability of XGBoost resulted from a regularized objective function, native handling of missing data, and built-in feature importance measures, which collectively enabled more robust discrimination of yield classes under stress, illustrating the advantage of ensemble boosting methods for categorical selection decisions in stressed environments. The ability to accurately classify high-yielding genotypes under stress has practical implications for breeding programs, as it enables the early identification of promising genotypes without requiring full-season field trials, thereby accelerating the selection cycle and reducing the cost of multi-environment testing (Chen and Guestrin, 2016; Crossa et al., 2017; Yu et al., 2025). The performance advantage of XGBoost over SVM and ANN also suggests that the relationship between spectral predictors and yield classes is characterized by complex, non-linear interactions that tree-based ensemble methods are particularly well-suited to capture, whereas neural networks, despite their theoretical flexibility, may require larger datasets than the 119 genotypes used in this study to achieve comparable performance (LeCun et al., 2015; Goodfellow et al., 2016). Our observations are consistent with previously published findings that tree-based models have consistently outperformed other algorithms in agricultural prediction tasks with moderate sample sizes (van Klompenburg et al., 2020; Niazian and Niedbała, 2020), reinforcing the recommendation that breeders should prioritize ensemble methods when working with limited training data.

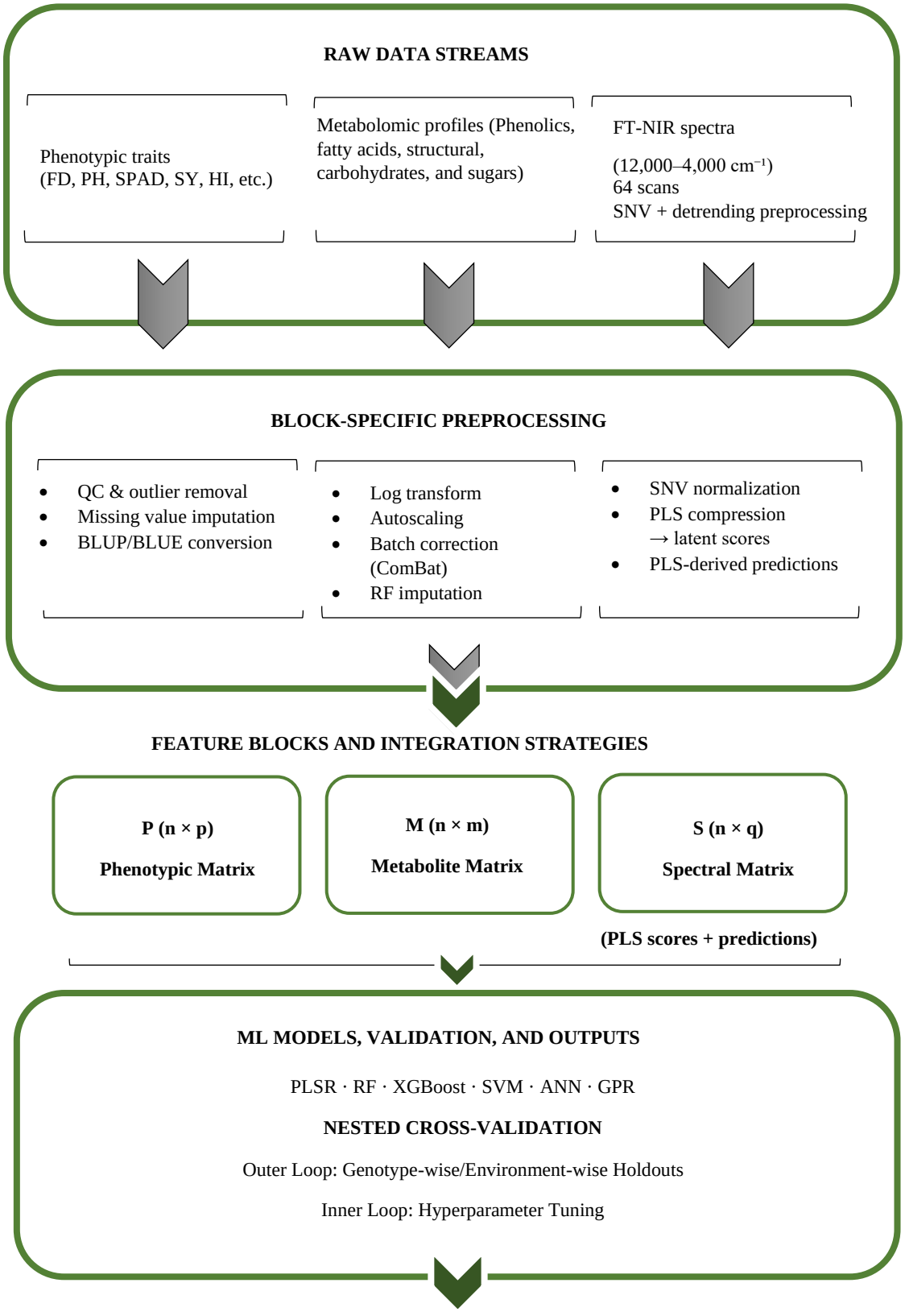


Fig. 1. Schematic overview of the multi-block data integration workflow. Phenotypic, metabolomic, and Fourier-transform near-infrared (FT-NIR) blocks are preprocessed separately and then combined. Environmental covariates are included in all strategies. All steps are performed inside nested cross-validation. Outputs feed yield prediction, quality prediction, genotype × environment interaction (GEI) interpretation, and multi-trait selection. Abbreviations: ANN, artificial neural networks; BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; ComBat, combating batch effects; FD, flowering date; HI, harvest index; PH, plant height; SPAD, Soil Plant Analysis Development (chlorophyll content); SY, seed yield; GPR, Gaussian process regression; PLS, partial least squares; PLSR, partial least squares regression; QC, quality control; RF, random forest; SNV, standard normal variate; SVM, support vector machines; XGBoost, eXtreme gradient boosting.

To evaluate the presence of overfitting, we employed multiple validation methodologies for comparison. Under a random 80/20 train–test partition, XGBoost attained 100% test accuracy. Employing 10-fold cross-validation yielded training accuracy ranging from 98% to 100% and test accuracy ranging from 87% to 91%. When implementing nested cross-validation with genotype-wise exclusions, accuracy diminished to 84–89%, and with environment-wise exclusions, accuracy further diminished to 81–86%. The substantial reduction from 100% to 84–89% observed under genotype-wise validation demonstrates that the initial 100% result constituted an artifact of random partitioning and signified overfitting to the training dataset. Consequently, the reported XGBoost accuracy of 84–89% derives from genotype-wise cross-validation and constitutes a more dependable assessment of the model's capacity to classify previously unseen genotypes. Supplementary learning curve analysis demonstrated that test accuracy stabilized following approximately 300–500 boosting iterations, whereas training accuracy continued to rise, thereby substantiating the implementation of early stopping mechanisms and regularization techniques.

Variable importance and the dominant role of SPAD chlorophyll content

Variable importance analysis using PLSR VIP scores consistently identified SPAD chlorophyll content as the most important predictor of seed yield under both well-watered and drought-stressed conditions, with VIP values ranging from 0.93 to 0.95 and positive coefficients (Table 3). However, predictor importance indicates a strong statistical association, not a causal relationship. Under well-watered conditions, the next most influential predictors were harvest index (VIP = 0.20–0.21, positive coefficient) and hemicellulose (VIP = 0.17–0.18, positive coefficient), suggesting that yield under optimal conditions is associated with photosynthetic capacity, efficient carbon partitioning, and structural integrity (Salami et al., 2026a). However, the hierarchy of predictors shifted markedly under drought stress. At 20 DAP drought, the top predictors after SPAD included linolenic acid (VIP = 0.16, negative coefficient), acid detergent fiber (VIP = 0.16, positive coefficient), and oil content (VIP = 0.14, negative coefficient), while at 40 DAP drought, isoquercetin I3G (VIP = 0.21, negative coefficient), oil content (VIP = 0.21, negative coefficient), sinapoyl malate (VIP = 0.20, positive coefficient), and total sugars (VIP = 0.19, negative coefficient) emerged as important predictors. This shift indicates that different physiological processes are associated with yield limitation in each environment, and the negative coefficient for oil content under late drought (–0.84) suggests a fundamental trade-off between productivity and oil accumulation under stress, which has profound implications for oilseed breeding: maximizing yield under stress inherently may inherently limits oil accumulation, reflecting a direct trade-off between productivity and quality (Salami et al., 2026a). Similarly, the emergence of isoquercetin (I3G) as a strong negative predictor under drought suggests that the metabolic cost of flavonoid biosynthesis may be associated with reduced yield under stress (Salami et al., 2026a), indicating that the activation of secondary metabolism under stress may occur at the expense of primary productivity. The significance of these predictors does not constitute evidence of causality; empirical experimentation would be necessary to substantiate these associations. The VIP scores for the top 15 predictors are comprehensively displayed in Fig. 4, with SPAD clearly showing the highest importance, reinforcing the central role of photosynthetic maintenance in drought resilience and suggesting that breeding efforts may benefit from prioritizing the maintenance of photosynthetic capacity under stress as a primary selection criterion (Salami et al., 2024a; Gill et al., 2026). The identification of SPAD as the most important predictor across all conditions also highlights the value of chlorophyll content measurements as a rapid, non-destructive indicator for yield potential that can be easily integrated into breeding programs, consistent with recent studies demonstrating strong correlations between SPAD values and photosynthetic performance under stress (Xia et al., 2025; Wang et al., 2025).

Non-linear trait interactions detected by Gaussian process regression

GPR further uncovered non-linear interactions that classical linear models failed to detect, specifically the critical interaction between flowering date and plant height (Salami et al., 2026b). The GPR response surface demonstrated that the highest predicted yield (4,500–5,000 kg ha^{-1}) occurred only within a narrow phenological window (175–185 days) combined with moderate-to-tall plant height (120–140 cm). Flowering earlier than 170 days sharply reduced yield regardless of plant height due to the insufficient vegetative growth, while flowering later than 195 days depressed yield, especially at low plant height, because terminal drought and heat stress shorten seed development. At optimal flowering dates, the plant height response was strongest, whereas at early flowering the height response was flat, and at late flowering tall plants showed negative yield responses due to the increased stress during grain filling. This interaction explains genotype × environment interaction at the trait level, with flowering date acting as a regulator that modulates the yield benefits of plant height, and highlights the unique value of ML as a complementary tool for understanding the physiological basis of GEI beyond the capabilities of classical linear models including AMMI and GGE biplots (Montesinos-López et al., 2022). However, these interaction patterns represent statistical associations learned from observational data and should not be interpreted as direct causal effects. Experimental manipulation of flowering date and plant height would be needed to confirm causality. Classical linear models, including AMMI and GGE biplots, cannot directly visualize such trait × trait interactions, illustrating the unique value of ML as a complementary tool for understanding the physiological basis of GEI (Salami et al., 2026b). The practical implication is that breeding programs targeting Mediterranean-type climates with

terminal drought should select for an optimal flowering date that balances vegetative growth and stress escape, and that the benefits of plant height are only realized within this optimal phenological window, a pattern that linear models would fail to capture (Luo et al., 2025; de Almeida et al., 2022). This finding also underscores the importance of considering phenological traits in combination with morphological traits when designing breeding strategies for drought-prone environments.

Table 2. Performance metrics of Partial Least Squares Regression (PLSR) models for seed yield prediction under well-watered and drought stress conditions at 20 and 40 days after pollination (DAP)

Condition	Components	Training R^2	Test R^2	RMSE (kg ha^{-1})	RPD
Well-watered 20 DAP	8	0.86	0.79	490	2.76
Well-watered 40 DAP	8	0.85	0.79	738	2.26
Drought 20 DAP	15	0.85	0.82	220	2.92
Drought 40 DAP	15	0.82	0.78	250	2.45

PLSR models were developed using the statistically inspired modification of partial least squares (SIMPLS) algorithm with leave-one-out cross-validation (LOOCV) to determine the optimal number of latent components. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE in kg ha^{-1}), and ratio of performance to deviation (RPD).

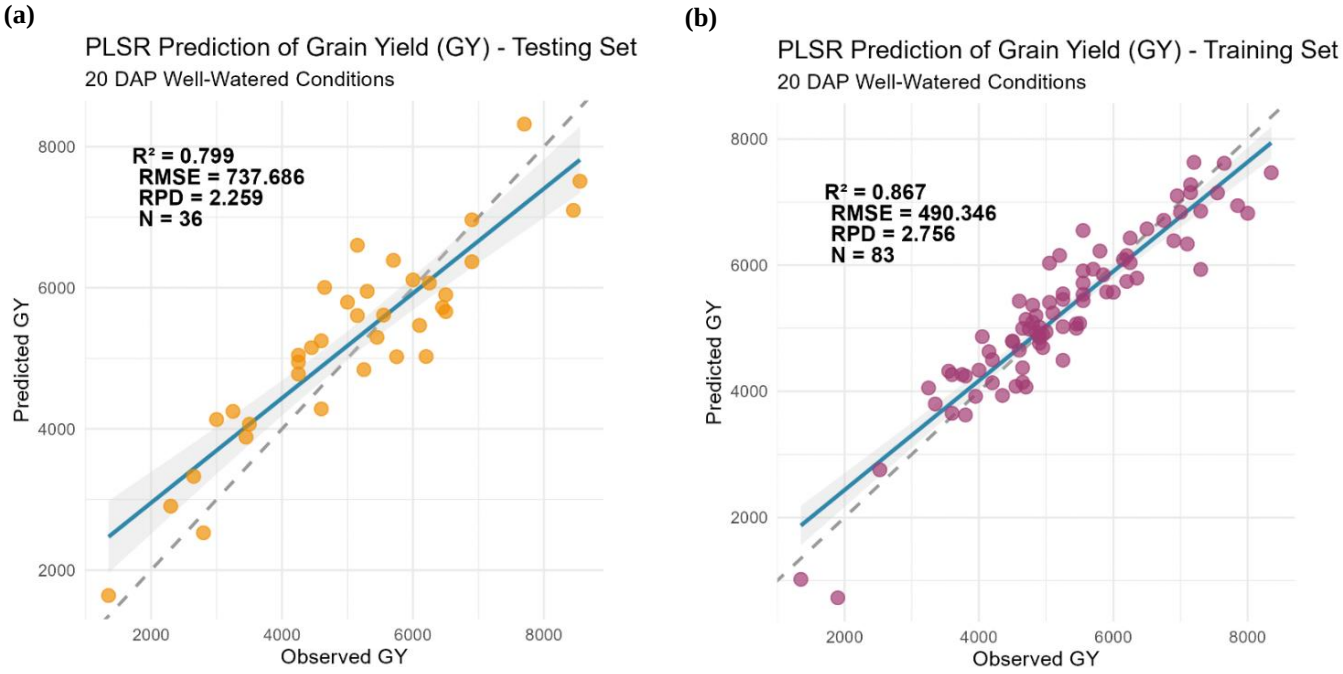


Fig. 2. Partial Least Squares Regression (PLSR) prediction of seed yield under drought stress at 20 days after pollination (DAP) in rapeseed. (a) Training set performance (N = 83) showing predictive accuracy with R^2 = 0.86, RMSE = 490.34, and RPD = 2.75. (b) Independent testing set performance (N = 36) confirming model generalizability with R^2 = 0.79, RMSE = 737.68, and RPD = 2.25. Solid diagonal lines represent 1:1 perfect prediction; grey dashed lines show regression fits.

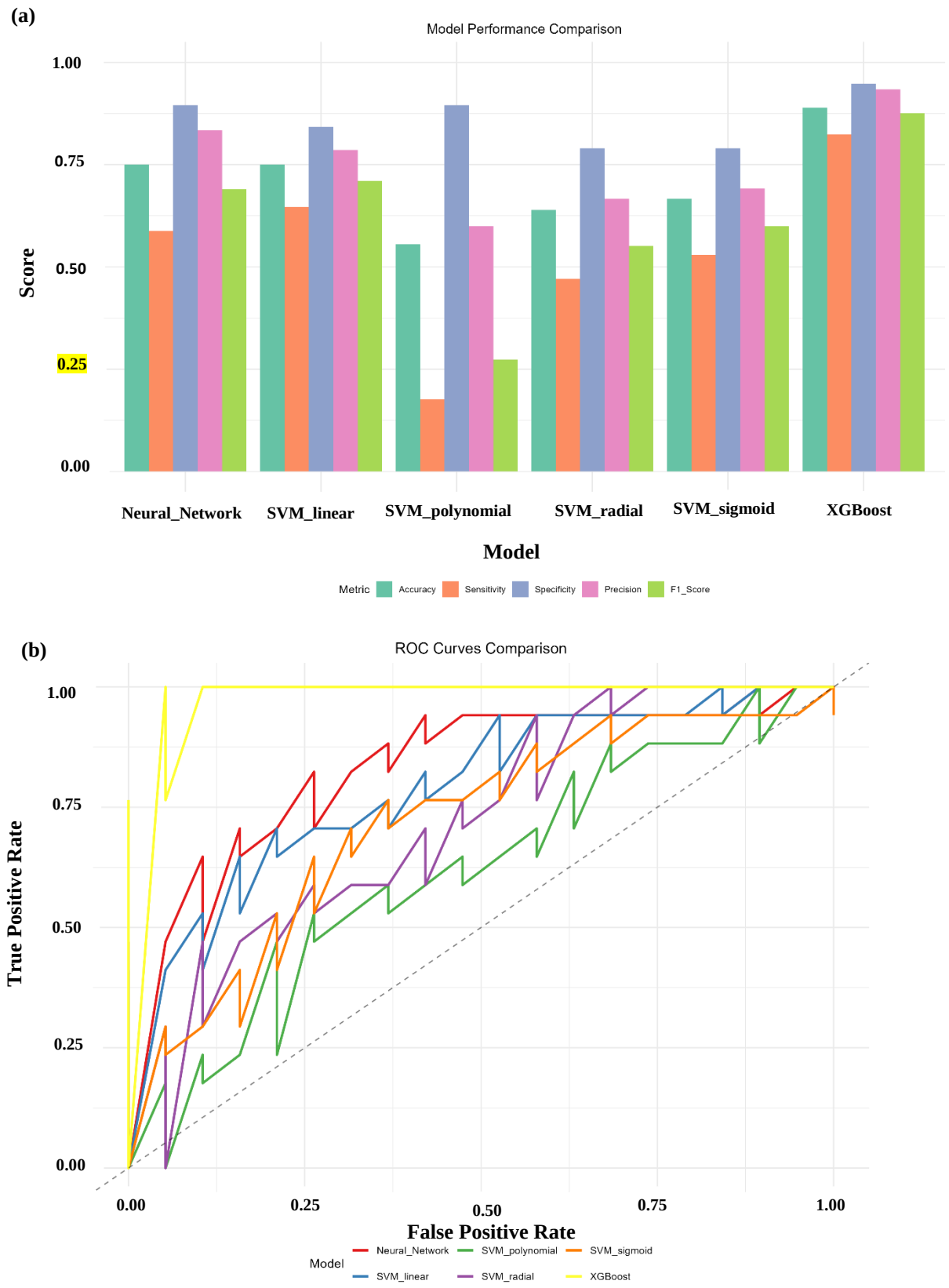


Fig. 3. (a) Comparative performance of machine learning (ML) models for seed yield classification under drought stress conditions. Bar chart comparing six ML algorithms: Neural Network, Support Vector Machine (SVM) with Linear, Radial Basis (RBF) kernels (three variants), and eXtreme Gradient Boosting (XGBoost). Performance metrics include Accuracy, Sensitivity/Recall, and Area Under the Receiver Operating Characteristic Curve (AUC). (b) Receiver Operating Characteristic (ROC) curves comparing ML model performance for grain yield classification under drought stress conditions. ROC curves show the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity) at various classification thresholds for different ML models. Models evaluated include XGBoost (Red), SVM Linear (Blue), Neural Network (Green), and multiple SVM Radial Basis Function (RBF) variants. The diagonal dashed line represents the performance of a random classifier (AUC = 0.5).

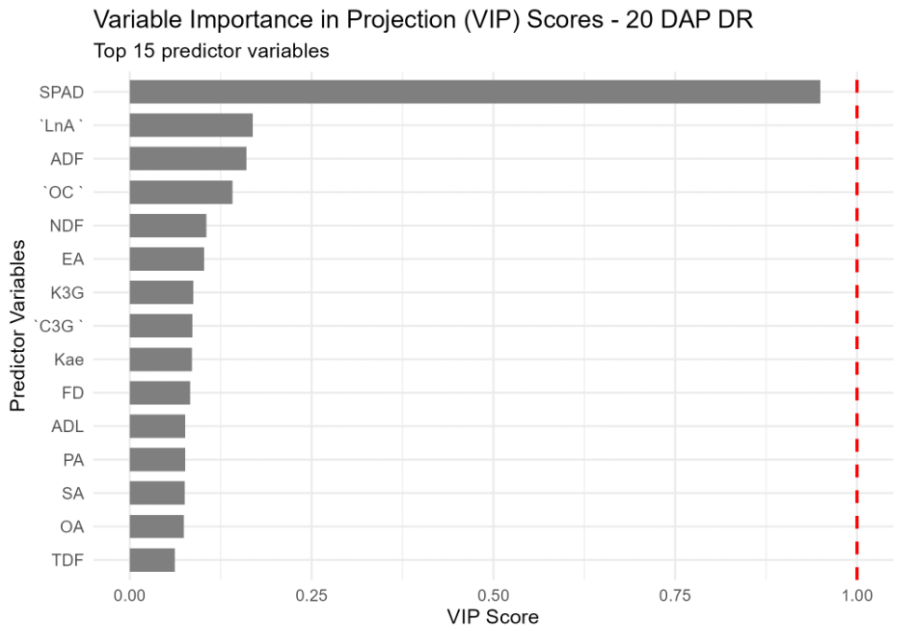


Fig. 4. Variable Importance in Projection (VIP) scores for the top 15 predictors, with SPAD chlorophyll content emerging as the dominant predictor (VIP = 0.95), followed by linolenic acid (0.90), acid detergent fiber (0.85), and oil content (0.80) in rapeseed. The abbreviations and their full names presented in this figure are as follows: SPAD (chlorophyll content), LnA (linolenic acid), ADF (acid detergent fiber), OC (oil content), NDF (neutral detergent fiber), EA (erucic acid), K3G (kaempferol-3-O-glucoside), C3G (cyanidin-3-glucoside), Kae (kaempferol), FD (flowering date), ADL (acid detergent lignin), PA (palmitic acid), SA (sinapic acid), OA (oleic acid), and TDF (total dietary fiber).

Table 3. Key predictors of grain yield under drought stress identified by Partial Least Squares Regression (PLSR) at 20 and 40 days after pollination (DAP)

Stage	Top predictors (ranked)	Coefficient
20 DAP Drought	SPAD (0.94)	+0.94
	Linolenic acid (0.16)	-0.16
	Acid detergent fiber (0.16)	+0.16
	Oil content (0.14)	-0.14
40 DAP Drought	SPAD (0.93)	+0.93
	Isoquercetin I3G (0.21)	-0.21
	Oil content (0.21)	-0.21
	Sinapoyl malate (0.20)	+0.20
	Total sugars (0.19)	-0.19

Genotype × environment interaction analysis using classical stability methods

The dissection of genotype × environment interactions was further advanced through classical stability analyses, namely AMMI and GGE biplots, which revealed that the water regime during critical growth stages is the primary environmental driver in rapeseed (Qasemi et al., 2022; Nowosad et al., 2016). The GGE biplot defined two primary mega-environments: one cooler and more humid where genotype G56 was the top performer, and another hot and high-evaporative where genotype G88 excelled, demonstrating a crossover interaction that precludes a single broadly adapted cultivar and necessitates separate breeding programs for each mega-environment (Yan et al., 2000; Yan and Tinker, 2006). Genotypes with principal component scores near zero exhibited broad adaptation and high static stability, suitable for unpredictable environments, whereas genotypes with extreme scores displayed strong specific adaptation and high environmental sensitivity (Gauch, 2006; Motahhari et al., 2020). The distinction between static stability (low variance across environments) and dynamic stability (responsiveness to improving environments) is critical for breeding decisions (Yan and Kang, 2003). The integrated framework for dissecting genotype × environment interactions using classical stability analysis and ML is presented in Fig. 5. ML approaches can enhance GEI analysis in several ways: ML models can predict genotype performance in untested environments based on environmental covariates, enabling prediction across environments (Wang et al., 2026; Tsega and Mullualem, 2026); ML feature importance measures identify the traits most responsible for GEI, guiding breeding strategies (Yu et al., 2025); and ML can detect non-linear trait interactions

that contribute to GEI but are missed by linear models (Yu et al., 2025; Wang et al., 2026). Recent studies have demonstrated that ANNs and RF models achieve excellent predictive power for total oil content (R^2 = 0.94) when applied to multi-year field data, adeptly accounting for complex G × E relationships without the need to include interaction variables explicitly (Rajković et al., 2023). This suggests that ML models can implicitly learn interaction patterns from data, reducing the need for explicit interaction modeling, a finding that aligns with the work of Crossa et al. (2025), who demonstrated that ML approaches are advancing genomic prediction within plant breeding through the utilization of large-scale datasets and sophisticated computational tools.

Seed composition and quality trait responses to drought stress

Predicting seed composition and quality traits revealed that drought stress consistently enhanced anthocyanin accumulation, with peonidin reaching 149.90 mg/g DW under drought at 40 DAP, representing a 144% increase over well-watered conditions (Salami et al., 2023; Salami et al., 2026a). This substantial increase is associated with drought-induced activation of MYB transcription factors (particularly MYB75/PAP1 and MYB90/PAP2) that upregulate late biosynthetic genes, including dihydroflavonol reductase (DFR) and anthocyanidin synthase (ANS) (Li et al., 2022), with pathway diversion channeling carbon from flavonol production toward anthocyanin formation, which may explain the coordinated increase across all five anthocyanin compounds. In contrast to anthocyanins, sinapine, the dominant hydroxycinnamate in rapeseed, decreased by approximately 30% under drought, while its biosynthetic intermediates (sinapic acid, caffeic acid, and ferulic acid) increased substantially (Salami et al., 2023), suggesting a specific enzymatic constraint where sinapine synthase may be downregulated under drought, while upstream reactions converting sinapoyl glucose to free hydroxycinnamic acids remain active or are upregulated (Lee et al., 2023). The accumulated free acids have higher antioxidant activity than sinapine, suggesting that this is an adaptive rerouting toward more effective reactive oxygen species scavengers (Salami et al., 2023; Lee et al., 2023), reflecting a trade-off where the plant may sacrifice sinapine accumulation for enhanced stress protection.

Heritability estimates and their implications for model development

Heritability estimates from the comprehensive rapeseed study, summarized in Table 4, revealed a clear hierarchy that guides ML model development (Salami et al., 2026a; Salami et al., 2026b). Very high heritability (0.87–0.99) was observed for anthocyanins, structural components, and phenological traits, suggesting that these characteristics are primarily governed by genetic factors and can be predicted accurately from simple models and are amenable to phenotypic selection. High heritability (0.60–0.74) was found for seed weight, branch number, and some hydroxycinnamic acids, suggesting moderate genetic control with some environmental influence. In contrast, low heritability (<0.01–0.55) was characteristic of most flavonoids, indicating that these traits are strongly influenced by environmental factors and require sophisticated ML approaches that can capture complex genotype-environment interactions and may benefit from genomic selection rather than phenotypic selection alone (García-Abadilla Velasco and Isidro Sánchez, 2025; Crossa et al., 2017). The contrast between anthocyanins (high heritability) and most flavonoids (low heritability) is particularly instructive, as both are phenolic compounds, yet their genetic architectures differ fundamentally, with anthocyanins controlled by a few major genes (MYB regulators) (Dai et al., 2016; Salami et al., 2024a) while flavonoids are influenced by numerous environmental factors (Salami et al., 2024b). This hierarchy guides ML model development: high-heritability traits can be predicted accurately from simple models and are amenable to phenotypic selection, while low-heritability traits require sophisticated ML approaches that can capture complex genotype-environment interactions and may benefit from genomic selection rather than phenotypic selection alone (García-Abadilla Velasco and Isidro Sánchez, 2025).

Multi-trait selection and integrated breeding pipeline

The MGIDI approach enabled the identification of elite genotypes that balanced yield, drought resilience, and nutritional quality, with factor analysis using varimax rotation identifying independent selection dimensions, and Euclidean distance calculations identifying genotypes closest to the defined ideotype (Olivoto et al., 2019; Olivoto and Nardino, 2021). The EWSI identified G112 and G113 as top economic performers despite not being agronomically superior in yield alone, highlighting the importance of seed quality traits for profitability and demonstrating that economic weighting can substantially alter selection rankings (Hazel, 1943; Salami et al., 2026b). The MTSI grouped genotypes into four clusters with specific combinations of traits and stability patterns, and the WAASB index, calculated by weighting the absolute scores of the BLUP interaction terms, enabled simultaneous evaluation of productivity and stability across multiple traits (Olivoto et al., 2019). A schematic overview of the integrated analytical framework combining dynamic metabolic profiling, multi-trait selection, and ML prediction for drought-resilient rapeseed breeding is presented in Fig. 6. The integration of ML with classical stability analyses and multi-trait selection indices provides a unified framework capable of identifying superior lines that unite high productivity, nutritional quality, and environmental resilience (Salami et al., 2026b). The integrated pipeline includes high-throughput phenotyping via FT-NIR spectroscopy and HPLC-DAD; ML prediction using PLSR, XGBoost, RF, and deep learning models for yield and quality; heritability-guided model development tailoring model complexity to trait heritability; GEI analysis using AMMI and GGE biplots for mega-environment delineation; multi-trait selection via MGIDI, EWSI, and MTSI for balanced trait improvement; non-linear interaction detection using GPR for identifying trait × trait and Genotype × environment × management (G × E × M) interactions; and field validation through multi-year, multi-location testing before cultivar release (Salami et al., 2026a; Salami et al., 2026b).

Challenges, limitations, and mitigation strategies

Despite the promising results, several challenges and limitations must be acknowledged. ML models can achieve perfect classification within a dataset but fail on independent populations (Niazian and Niedbała, 2020). In the present study, XGBoost achieved 100% classification accuracy with a random 80/20 train–test split; however, random splitting does not respect genotype or environment structure and can yield overly optimistic estimates. Using nested cross-validation with genotype-wise holdouts, the accuracy decreased to 84–89%, and using environment-wise holdouts, the accuracy decreased to 81–86%. These results confirm that the 100% value was an artifact of data splitting and that genotype-wise validation provides a more realistic estimate of performance. Mitigation strategies include proper data splitting (training and test sets independent with no data leakage, typically 70/30 or 80/20, with stratified splitting), cross-validation with k folds (commonly 10-fold with 3–5 repeats) for tuning model complexity and providing unbiased performance estimates (Hastie et al., 2009), regularization (XGBoost's built-in regularization, RF's bootstrap aggregation, dropout and L1/L2 regularization for neural networks) (Breiman, 2001; Chen and Guestrin, 2016; Goodfellow et al., 2016), and independent validation on genotypes not used in model development, ideally across years or locations (Crossa et al., 2017; Yu et al., 2025). Interpretability remains a challenge because ML models often function as "black boxes," limiting biological insight (Danilevicz et al., 2025), and feature importance measures (permutation importance, VIP scores) identify influential predictors but indicate correlation, not causation (Ogra et al., 2026). In this study, SPAD was identified as the most important predictor of yield, but whether high SPAD directly increases yield or is merely correlated with underlying photosynthetic capacity remains a question requiring SHAP analyses and, ultimately, experimental validation through physiological and molecular studies (Salami et al., 2026a). SHAP values provide local attribution by attributing each feature's impact on the output prediction, helping breeders understand why a given genotype got a particular result (Yu et al., 2025), while partial dependence plots visualize the marginal effect of predictors on predictions, revealing non-linear relationships and thresholds (Molnar, 2022). However, even these interpretability tools describe model behavior rather than biological causation. Establishing causality would require controlled experiments such as SPAD manipulation, gene expression analysis, or clustered regularly interspaced short palindromic repeats (CRISPR)-based validation. The trade-off between accuracy and interpretability must be managed, with simpler models preferred when biological insight is the primary goal and complex models preferred when prediction accuracy is paramount (Rudin, 2019). The multi-block integration workflow introduces additional challenges, including block-specific missingness, scale differences, and the risk of overfitting when many features are concatenated. We mitigated these risks by using block-wise imputation, within-fold preprocessing, intermediate fusion with principal components or PLS scores, and late fusion with out-of-fold predictions.

Data quality and representativeness are critical, as ML performance depends heavily on sample size, environmental coverage, trait completeness, and handling of missing data (Danilevicz et al., 2025; Sharma and Gupta, 2026). Large populations (> 100 genotypes) are needed for robust ML models, and in the rapeseed study, models trained on 119 genotypes achieved good performance, but larger populations would improve generalization (Ullah, 2025). Training data must capture the range of environmental conditions where predictions will be applied (van Klompenburg et al., 2020), and the multi-year, two-irrigation regime design in the rapeseed study captured diverse environmental conditions, although additional years and locations would further enhance model robustness. Comprehensive phenotyping provides richer information but increases costs; therefore, trait selection should be guided by heritability and biological relevance. Missing data handling differs among ML methods, with XGBoost handling missing values natively, whereas others require imputation (Chen and Guestrin, 2016).

Biological validation is essential, as ML predictions require confirmation through physiological experiments, molecular validation (e.g., CRISPR mutants, gene expression analysis), and multi-year, multi-location field trials before cultivar release (Chen et al., 2025; Ikram et al., 2025; Liang et al., 2026). Computational requirements, while not prohibitive, include considerations for model training (XGBoost and RF are efficient up to 100,000 observations, while deep learning requires graphics processing unit (GPU) acceleration for larger datasets), hyperparameter optimization (grid search or Bayesian optimization can be time-consuming, while randomized search is a practical alternative), and interpretability tools (SHAP and partial dependence plots are computationally intensive for large datasets) (Munroe et al., 2026; Danilevicz et al., 2025).

Future perspectives

While tree-based models currently dominate plant breeding applications, deep learning offers several potential advantages, including convolutional neural networks (CNNs) for image-driven phenotyping combined with yield prediction (Jiang and Li, 2020), recurrent neural networks (RNNs) and long short-term memory (LSTM) networks for temporal data (growth curves, stress progression, and weather data) to predict final yield (Mienye et al., 2024), knowledge transfer using pre-trained architectures for specific breeding applications (Hossen et al., 2025), and graph neural networks for modeling genetic relationships and genomic prediction (Wang et al., 2026). However, deep learning requires larger datasets than typical breeding programs generate (10,000+ samples), and computational costs and interpretability issues remain significant challenges (LeCun et al., 2015; Goodfellow et al., 2016). ML integration with multi-omics is a major frontier (Tsega and Mullualem, 2026), with feature selection identifying key markers, metabolites, or transcripts associated with drought tolerance (Wang et al., 2026), multi-omics integration combining genomics, transcriptomics, metabolomics, and phenomics in ML models improving prediction accuracy (Yang and Tang, 2026), and integrating environmental covariates with genomic data enabling genotype-by-environment prediction (Tsega and Mullualem, 2026).

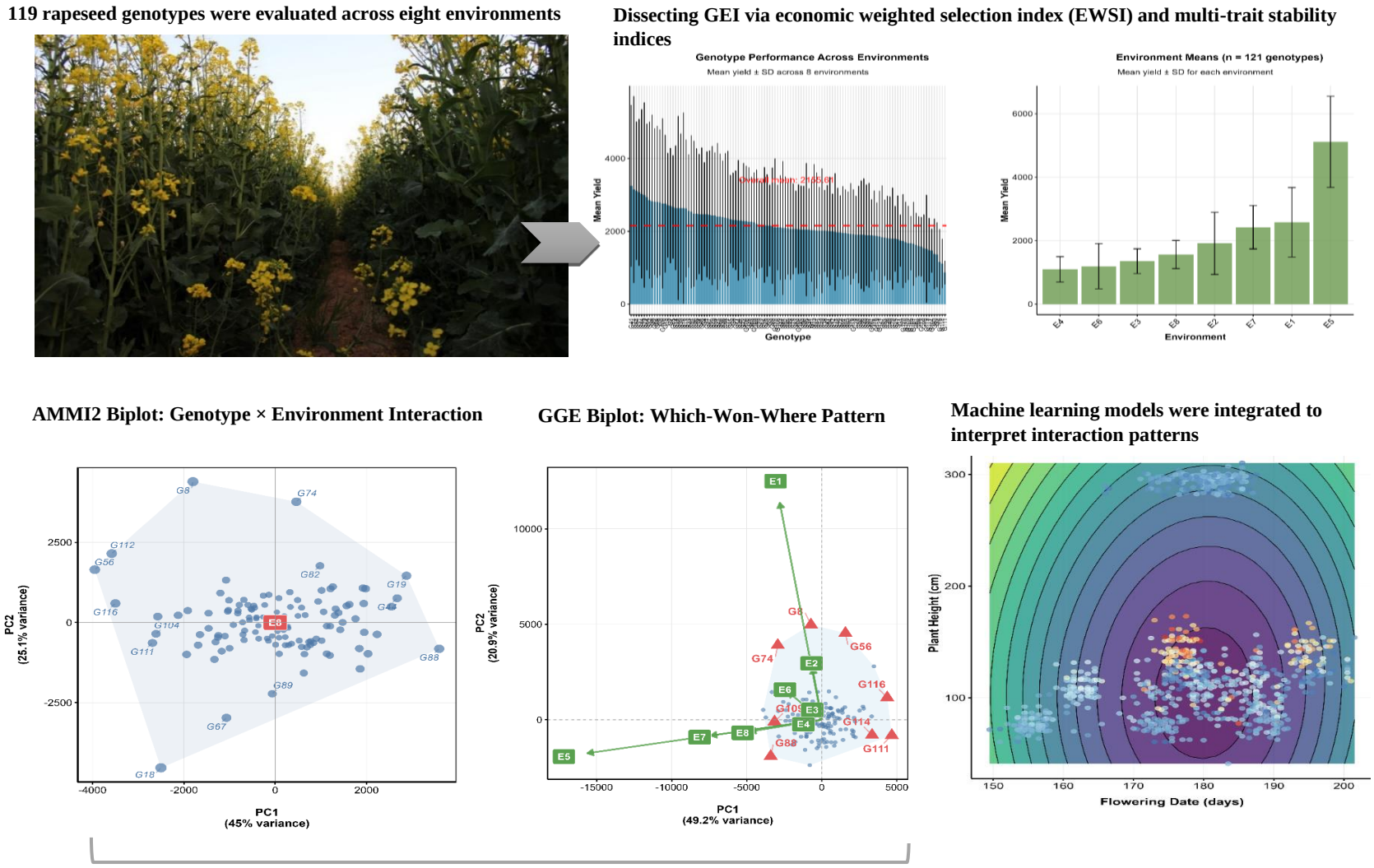


Fig. 5. Integrated framework for dissecting genotype × environment interactions (GEI) in rapeseed using classical stability analysis and machine learning (ML). The schematic summarizes the comprehensive analytical framework for identifying stable, high-yielding, and economically superior rapeseed genotypes across diverse environments. One hundred nineteen genotypes were evaluated across eight environments (4 years × 2 irrigation regimes). Additive main effects and multiplicative interaction (AMMI) and genotypic main effect and genotype by environment interaction (GGE) biplot analyses quantified GEI patterns and identified two mega-environments with different winning genotypes. ML models interpreted complex interaction patterns.

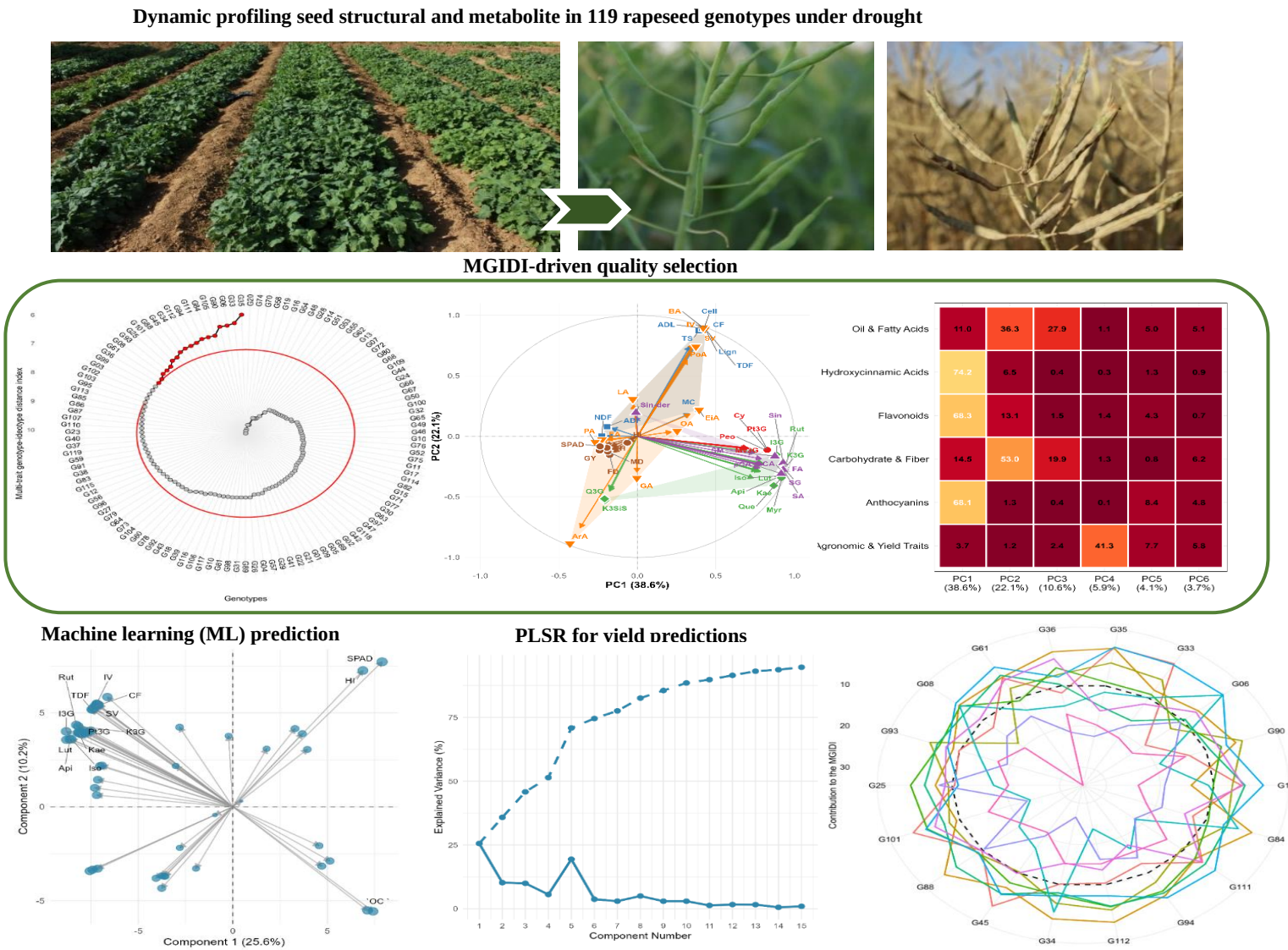


Fig. 6. A schematic overview of the integrated analytical framework combining dynamic metabolic profiling, multi-trait selection, and machine learning (ML) prediction for drought-resilient rapeseed breeding. The workflow includes time-resolved metabolic profiling of 119 genotypes at 20 and 40 days after pollination (DAP) using Fourier-transform near-infrared (FT-NIR) spectroscopy and high-performance liquid chromatography with diode array detection (HPLC-DAD); multi-trait genotype-ideotype distance index (MGIDI)-driven quality selection identifying elite genotypes balancing yield, drought resilience, and nutritional quality; and partial least squares regression (PLSR) and ML models achieving high predictive accuracy for seed yield, with soil plant analysis development (SPAD) chlorophyll content as the dominant predictor.

Table 4. Broad-sense heritability (h^2) estimates for metabolic, structural, and agronomic traits in rapeseed

Category	Traits	Heritability (h^2)
Very high	Anthocyanins, structural components, phenological traits	0.87–0.99
High	Seed weight, branch number, some hydroxycinnamic acids	0.60–0.74
Low	Most flavonoids	0.001–0.55

The integration of artificial intelligence (AI) and ML significantly enhances our ability to identify genetic markers related to nutritional composition, stress resilience, and overall crop adaptability (Wang et al., 2026), and for rapeseed, this means identifying genes and pathways controlling both drought tolerance and seed quality traits (Salami et al., 2024a; Salami et al., 2024b; Gill et al., 2026).

Real-time field deployment is becoming feasible through edge computing, with mobile applications for smartphone-based SPAD estimation, canopy coverage analysis, and disease detection under development (Xia et al., 2025), unmanned aerial vehicle (UAV)-based phenotyping with drones equipped with multispectral cameras providing high-throughput field phenotyping with ML models processing imagery to predict yield and stress responses (Zhang et al., 2026), and Internet of Things (IoT) integration with soil moisture sensors, weather stations, and field cameras providing real-time data for ML models enabling adaptive management recommendations (Yang et al., 2026; Sudha and Loret, 2026). Finally, explainable artificial intelligence (XAI) is addressing the "black box" limitation through SHAP values, LIME (Local Interpretable Model-agnostic Explanations), counterfactual explanations, and integrated gradients (Głuchowska et al., 2026; Yenkikar et al., 2025), enabling breeders to understand which traits or markers drive predictions and facilitating biological interpretation and hypothesis generation (Głuchowska et al., 2026).

CONCLUSIONS

This study demonstrated that ML models, including PLSR, XGBoost, RF, SVM, ANN, and deep learning approaches, consistently achieve high predictive accuracy for seed yield and quality traits, enable early screening through rapid phenotyping, and uncover non-linear trait interactions that classical models miss. The integration of ML with classical stability analyses and multi-trait selection indices provides a unified framework capable of identifying superior lines that unite high yield, metabolic value, and environmental stability. Key conclusions from this work are that ML models achieved R^2 values of 0.79–0.91 for seed yield and quality traits across diverse environments; SPAD chlorophyll content emerged as the most important predictor of yield under both optimal and stress conditions (VIP = 0.93–0.95); GPR revealed critical non-linear interactions such as the flowering date × plant height interaction that classical linear models fail to detect; and the integrated framework combining ML with AMMI, GGE biplot, MGIDI, EWSI, and MTSI offers a practical and powerful approach for identifying superior genotypes. It is crucial to underscore that these results demonstrate statistical correlations rather than established causal connections, and further experimental investigation is required to substantiate the fundamental biological processes involved. The effectiveness of ML prediction models depends critically on trait heritability, with high-heritability traits amenable to simple models and low-heritability traits requiring sophisticated ML approaches. ML models also reveal the inherent trade-off between yield and quality traits under stress, guiding breeding strategies that balance productivity and nutritional quality. Ultimately, ML-enhanced breeding enables early selection of superior genotypes without waiting for multi-year field trials, accelerating the creation of climate-adaptive varieties with optimized agronomic and nutritional values. This transformative move, shifting from empirical phenotype screening to forward-looking, analytics-enabled crop improvement, offers a practical and powerful framework for developing climate-resilient cultivars with optimized agronomic and nutritional values.

FUNDING

This work was supported by a fund (No. 4029213) from the Iran National Science Foundation (INSF).

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Conceptualization: Bahram Heidari; Methodology: Bahram Heidari; Software: Maryam Salami; Validation: Bahram Heidari; Formal analysis: Maryam Salami; Investigation: Bahram Heidari; Resources: Bahram Heidari; Data curation: Maryam Salami; Writing—original draft preparation: Maryam Salami; Writing—review and editing: Bahram Heidari; Visualization: Maryam Salami; Supervision: Bahram Heidari; Project administration: Bahram Heidari; Funding acquisition: Bahram Heidari.

DECLARATION OF COMPETING INTEREST

The authors declare no conflicts of interest.

ETHICAL STATEMENT

Approval was granted by the Ethics Committee of University Shiraz.

DATA AVAILABILITY

All data analyzed and generated during this study are included in this published article.

ACKNOWLEDGMENTS

The authors are thankful to the Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, Iran, for providing necessary help for this project.

REFERENCES

Bakri, A.A., Amery, M.A., Su, K., Geneve, R., Crocker, M., Teets, N., Armstrong, P., Kachroo, P., & Hildebrand, D. (2023). Development of a rapid and simple protocol for oil quantification of small (mg) mass oil seed samples. *Biocatalysis and Agricultural Biotechnology*, *50*, 102715. https://doi.org/10.1016/j.bcab.2023.102715

Barajehfard, M., Fazli, M., Fooladvand, M., Salami, M., Stevanato, P., & Heidari, B. (2026). Selection for high-yielding sugarcane ideotypes and multi-trait selection via MGIDI across different harvesting periods improves sugar quality and yield. *Scientific Reports*, *16*(1), 4165. https://doi.org/10.1038/s41598-025-34232-8

Beć, K. B., Grabska, J., & Huck, C. W. (2020). Near-infrared spectroscopy in bio-applications. *Molecules*, *25*(12), 2948. https://doi.org/10.3390/molecules25122948

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, *17*(Suppl 2), S15. https://doi.org/10.1186/s12859-015-0857-9

Bennett, E.J., Brignell, C.J., Carion, P.W., Cook, S.M., Eastmond, P.J., Teakle, G.R., Hammond, J.P., Love, C., King, G.J., Roberts, J.A., & Wagstaff, C. (2017). Development of a statistical crop model to explain the relationship between seed yield and phenotypic diversity within the *Brassica napus* genepool. *Agronomy*, *7*(2), 31. https://nottingham-repository.worktribe.com/output/8387088

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://www.springer.com/gp/book/9780387310732

Blanco, M., & Villarroya, I. (2002). NIR spectroscopy: A rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, *21*(4), 240-250. https://doi.org/10.1016/S0165-9936(02)00404-1

Bouchyoua, A., Nabloussi, A., Driouech, H., Bahri, H., Benyza, R., Kouighat, M., Guirrou, I., Kassimi, C.E.D., Hammani, K., & Khabbach, A. (2026). Drought effects on seed yield stability and oil quality traits in different rapeseed genotypes: Toward adaptive sustainability of crops in semi-arid regions. *Sustainability*, *18*(3), 1388. https://doi.org/10.3390/su18031388

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Brown, P. N., & Shipley, P. R. (2011). Determination of anthocyanins in cranberry fruit and cranberry fruit products by high-performance liquid chromatography with ultraviolet detection: Single-laboratory validation. *Journal of AOAC International*, *94*, 459-466. https://doi.org/10.1093/jaoac/94.2.459

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079-2107. https://doi.org/10.5555/1756006.1859921

Chang, Y. J., Lin, Y. L., & Pai, P. F. (2025). Support vector machines with hyperparameter optimization frameworks for classifying mobile phone prices in multi-class. *Electronics*, *14*(11), 2173. https://doi.org/10.3390/electronics14112173

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Chen, W., Miao, Y., Ayyaz, A., Hannan, F., Huang, Q., Ulhassan, Z., Zhou, Y., Islam, F., Hong, Z., Farooq, M.A., & Zhou, W. (2022). Purple stem *Brassica napus* exhibits higher photosynthetic efficiency, antioxidant potential and anthocyanin biosynthesis related genes expression against drought stress. *Frontiers in Plant Science*, *13*, 936696. https://doi.org/10.3389/fpls.2022.936696

Chen, Y., Wang, Y., Liu, L., Yu, X., Zhang, Y., Xi, M., Xu, J., Yang, H., Xie, C., & Wang, D. (2025). Targeting BnNAC038 improves drought tolerance with low yield penalty in *Brassica napus*. *The Plant Journal*, *124*(3), e70571. https://doi.org/10.1111/tpj.70571

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Crossa, J., Martini, J.W., Vitale, P., Pérez-Rodríguez, P., Costa-Neto, G., Fritsche-Neto, R., Runcie, D., Cuevas, J., Toledo, F., Li, H., & De Vita, P. (2025). Expanding genomic prediction in plant breeding: Harnessing big data, machine learning, and advanced software. *Trends in Plant Science*, *30*(7), 756-774. https://doi.org/10.1016/j.tplants.2024.12.009

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., & Dreisigacker, S. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Dai, W., Wang, S., Ketudat-Cairns, M., Huang, Z., Tang, R., & Girdthai, T. (2016). Genetic analysis for anthocyanin and chlorophyll contents in rapeseed. *Ciência Rural*, *46*(5), 791-796. https://doi.org/10.1590/0103-8478cr20150564

Danilevicz, M. F., Upadhyaya, S. R., Batley, J., Bennamoun, M., Bayer, P. E., & Edwards, D. (2025). Understanding plant phenotypes in crop breeding through explainable AI. *Plant Biotechnology Journal*, *23*(10), 4200-4213. https://doi.org/10.1111/pbi.70208

Danudol, A., & Judprasong, K. (2022). Development and method validation of butyric acid and milk fat analysis in butter blends and blended milk products by GC-FID. *Foods*, *11*, 3606. https://doi.org/10.3390/foods11223606

de Almeida, L. M. M., Corlouer, E., Laperche, A., Nesi, N., Mollier, A., & Brunel-Muguet, S. (2022). To what extent can ecoclimatic indicators assist crop performance predictions in oilseed rape upon repeated heat stresses? *European Journal of Agronomy*, *141*, 126622. https://hal.inrae.fr/hal-03845468

Fahey, G. C., Novotny, L., Layton, B., & Mertens, D. R. (2019). Critical factors in determining fiber content of feeds and foods and their ingredients. *Journal of AOAC International*, *102*, 52-62. https://doi.org/10.5740/jaoacint.18-0067

Ferjancic, B., Skrt, M., Korosec, M., & Bertoncelj, J. (2022). Comparative analysis of dietary fibre determination by AOAC 991.43 and AOAC 2011.25 for frequently consumed foods in Slovenia. *Food Chemistry*, *397*, 133753. https://doi.org/10.1016/j.foodchem.2022.133753

Fletcher, R. S., Mullen, J. L., Heiliger, A., & McKay, J. K. (2015). QTL analysis of root morphology, flowering time, and yield reveals trade-offs in response to drought in *Brassica napus*. *Journal of Experimental Botany*, *66*(1), 245-256. https://doi.org/10.1093/jxb/eru423

Friedli, L., Steinert, T., Wuyts, N., Guignard, F., Häner, L.L., Herrera, J.M., & Ginsbourger, D. (2026). Gaussian Process Modeling with Genotype× Environment Kernels for Wheat Performance Prediction. *Journal of Agricultural, Biological and Environmental Statistics*, pp.1-22. https://doi.org/10.1007/s13253-026-00749-2

García-Abadilla Velasco, J., & Isidro Sánchez, J. (2025). Machine learning approaches applied to genomic-assisted plant breeding. (Doctoral Thesis, E.T.S. de Ingeniería Agronómica, Alimentaria y de Biosistemas (UPM)). https://oa.upm.es/86948/

Gauch Jr, H. G. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Science*, *46*(4), 1488-1500. https://doi.org/10.2135/cropsci2005.07-0193

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

Gill, R. A., Helal, M. M. U., Xing, Q., Ali, S., Zhou, W., & Müller-Xing, R. (2026). Integrating multiomic resources and gene expression studies to identify candidates for engineering climate-resilient *Brassica*. *Frontiers in Plant Science*, *17*, 1747148. https://doi.org/10.3389/fpls.2026.1747148

Giménez-Berenguer, M., Salicola, S.A., Formenti, C., Giménez, M.J., Mauromicale, G., Zapata, P.J., Lombardo, S., & Pandino, G. (2025). Seeds mineral profile and ash content of thirteen different genotypes of cultivated and wild cardoon over three growing seasons. *Agriculture*, *15*, 1228. https://doi.org/10.3390/agriculture1512228

Głuchowska, A., Khan, M. H. U., & Pawełkowicz, M. (2026). Explainable artificial intelligence for genomic prediction in plants: A critical review of methods, biological insights, and practical challenges. *Applied Sciences*, *16*(14), 7275. https://doi.org/10.3390/app16147275

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from: http://www.deeplearningbook.org

Guo, H., Li, S., Kang, S., Du, T., Tong, L., Hao, X., & Ding, R. (2020). Crop coefficient for spring maize under plastic mulch based on 12-year eddy covariance observation in the arid region of Northwest China. *Journal of Hydrology*, *588*, 125108. https://doi.org/10.1016/j.jhydrol.2020.125108

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, *28*(6), 476-490. https://doi.org/10.1093/genetics/28.6.476

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R.N., Ssenyomo, G., Kamali, A., & Kaleebu, P. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113, 7377-7382. https://doi.org/10.1073/pnas.1510497113

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Hossen, M. I., Awrangjeb, M., Pan, S., & Mamun, A. A. (2025). Transfer learning in agriculture: A review. *Artificial Intelligence Review*, *58*(4), Article 97. https://doi.org/10.1007/s10462-024-11081-x

Ikram, M., Farhan, M., Derakhshani, B., Kumar, S., Khan, N., Gupta, R., Usman, B., & Liu, P. (2025). Machine learning and CRISPR-based validation elucidate *OsWOX13* involvement in rice heat stress tolerance and flowering. *Physiologia Plantarum*, *177*(6), e70714. https://doi.org/10.1111/ppl.70714

IPCC. (2021). *Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report*. Cambridge University Press. https://doi.org/10.1017/9781009157896

Jiang, Y., & Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics*, 2020, 4152816. https://doi.org/10.34133/2020/4152816

Kakaei, M., Chaghakaboodi, Z., Zebarjadi, A., & Kahrizi, D. (2024). Exploring the physiology and genetic stability of rapeseed plants for assessing oil content in Western Iran. *Agrotechniques in Industrial Crops*, *5*(1), 34-45. https://journals.razi.ac.ir/article_3224.html

Karegar, M., Fazaeli, M., Mazal-Mazraei, T., Richards, C., Salami, M., Zarean, M., Moghaddam, A., Honar, T., Naderi, R., Dadkhodaie, A., & Pakniyat, H. (2026). MGIDI-based ideotype selection and genetic variation for wheat agronomic traits and bread-making quality (BMQ) under drought stress condition. *BMC Plant Biology*, *26*, 1343. https://doi.org/10.1186/s12870-026-09175-5

Khattab, R., Eskin, M., Aliani, M., & Thiyam, U. (2010). Determination of sinapic acid derivatives in canola extracts using high-performance liquid chromatography. *Journal of the American Oil*

Chemists' Society, *87*, 147-155. https://doi.org/10.1007/s11746-009-1486-0

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lee, B. R., Park, S. H., Muchlas, M., La, V. H., Al Mamun, M., Bae, D. W., & Kim, T. H. (2023). Differential response of phenylpropanoid pathway as linked to hormonal change in two *Brassica napus* cultivars contrasting drought tolerance. *Physiologia Plantarum*, *175*(6), e14115. https://doi.org/10.1111/ppl.14115

Li, Z., Liu, W., Lin, L., Yang, J., & Fu, S.H. (2022). Cloning and expression analysis of BNMYB90 and bnmyb114 and their regulation on dfr and F3H in *Brassica napus*. *Pakistan Journal of Botany*, *54*(4), 1335-1347. https://doi.org/10.30848/PJB2022-4(24)

Li, D., Wang, C., Miao, Y., Fernández, F. G., & Chen, S. (2026). Developing an innovative stacking ensemble machine learning and multi-source data fusion-based precision nitrogen management strategy for corn. *Precision Agriculture*, *27*(2), 41. https://doi.org/10.1007/s11119-026-10344-7

Liang, X., Yu, S., Ju, Y., Wang, Y., & Yin, D. (2026). Integrating envirotyping and phenomics for AI-enabled multi-environment genomic prediction in crop breeding. *Agronomy*, *16*(10), 1019. https://doi.org/10.3390/agronomy16101019

Liao, P., Lechon, T., Harwood, J. L., & Scofield, S. (2026). Lipid regulation in *Brassica napus*: Spatiotemporal studies to enhance our understanding. *Frontiers in Plant Science*, *17*, 1813997. https://doi.org/10.3389/fpls.2026.1813997

Lima, I.A.D., Khalil, N.M., & Mainardes, R.M. (2017). A stability-indicating HPLC-PDA method for the determination of ferulic acid in chitosan-coated poly(lactide-co-glycolide) nanoparticles. *Brazilian Journal of Pharmaceutical Sciences*, 53, e16138. https://doi.org/10.1590/s2175-97902017000216138

Lin, L. Z., & Harnly, J. M. (2007). A screening method for the identification of glycosylated flavonoids and other phenolic compounds using a standard analytical approach for all plant materials. *Journal of Agricultural and Food Chemistry*, *55*, 1084-1096. https://doi.org/10.1021/jf0624316

Ling, Q., Huang, W., & Jarvis, P. (2011). Use of a SPAD-502 meter to measure leaf chlorophyll concentration in *Arabidopsis thaliana*. *Photosynthesis Research*, *107*(2), 209-214. https://doi.org/10.1007/s11120-010-9606-0

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., Zhang, C., Chen, Z., Xiao, Z., Jian, H., & Cheng, F. (2019). Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nature Communications*, *10*(1), 1154. https://doi.org/10.1038/s41467-019-09134-9

Luo, T., Bucat, J., Power, S., Hu, L., & Zhang, H. (2025). Condensed flowering is a viable strategy for reliable canola yields in terminal drought-prone regions. *Crop & Pasture Science*, *76*(12), CP25148. https://doi.org/10.1071/CP25148

Luo, T., Lin, R., Cheng, T., & Hu, L. (2022). Low temperature rather than nitrogen application mainly modulates the floral initiation of different ecotypes of rapeseed (*Brassica napus* L.). *Agronomy*, *12*(7), 1624. https://doi.org/10.3390/agronomy12071624

Matar, S., Kumar, A., Holtgräwe, D., Weisshaar, B., & Melzer, S. (2021). The transition to flowering in winter rapeseed during vernalization. *Plant, Cell & Environment*, *44*(2), 506-518. https://doi.org/10.1111/pce.13937

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable*. Self published. https://christophm.github.io/interpretable-ml-book/

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction*. Springer. https://doi.org/10.1007/978-3-030-89010-0

Moreno, S. R. (2025). Mining the oil code: New insights behind oil production in *Brassica napus*. *Plant Physiology*, *197*(1), kiae457. https://doi.org/10.1093/plphys/kiae457

Motahhari, A., Hervan, E. M., Alizadeh, B., & Khosroshali, M. (2020). The analysis of genotype × environment interaction using rapeseed (*Brassica napus* L.) by GGE biplot method. *Iraqi Journal of Agricultural Sciences*, *51*(5), 1337-1349. https://doi.org/10.36103/ijas.v51i5.1148

Niazian, M., & Niedbała, G. (2020). Machine learning for plant breeding and biotechnology. *Agriculture*, *10*(10), 436. https://doi.org/10.3390/agriculture10100436

Nowosad, K., Liersch, A., Popławska, W., & Bocianowski, J. (2016). Genotype by environment interaction for seed yield in rapeseed (*Brassica napus* L.) using additive main effects and multiplicative interaction model. *Euphytica*, *208*(1), 187-194. https://doi.org/10.1007/s10681-015-1620-z

Novielli, P., Romano, D., Pavan, S., Losciale, P., Stellacci, A.M., Diacono, D., Bellotti, R., & Tangaro, S. (2024). Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: a case study with a dataset from an almond germplasm collection. *Frontiers in Plant Science*, *15*, 1434229. https://doi.org/10.3389/fpls.2024.1434229

Ogra, I. O., Silue, Y., Fawole, O. A., Aremu, A. O., & Opara, U. L. (2026). Applications of machine learning in plant biotechnology. *Discover Artificial Intelligence*, *6*(1), 1. https://doi.org/10.1007/s44163-026-00256-w

Olivoto, T., & Nardino, M. (2021). MGIDI: Toward an effective multivariate selection in biological experiments. *Bioinformatics*, *37*(10), 1383-1389. https://doi.org/10.1093/bioinformatics/btaa981

Olivoto, T., Lúcio, A. D., da Silva, J. A., Sari, B. G., & Diel, M. I. (2019). Mean performance and stability in multi-environment trials II: Selection based on multiple traits. *Agronomy Journal*, *111*(6), 2961-2969. https://doi.org/10.2134/agronj2019.03.0221

Ozer, H. (2003). Sowing date and nitrogen rate effects on growth, yield and yield components of two summer rapeseed cultivars. *European Journal of Agronomy*,

19(3), 453-463. https://doi.org/10.1016/S1161-0301(02)00136-3

Pour-Aboughadareh, A., Zali, H., Gholipour, A., Marzooghian, A., Koohkan, S., Omrani, A., Shahbazi-Homonloo, K., Moradgholi, A., Kheirgoo, M., Jamshidi, B., & Dabiri, M. (2026). Integrating AMMI, Bayesian-AMMI, BLUP, and GGE biplot models to improve the identification of high-yielding and stable barley genotypes. *BMC Plant Biology*, *26*, 1297. https://doi.org/10.1186/s12870-025-06383-x

Qasemi, S. H., Mostafavi, K., Khosroshahli, M., Bihamta, M. R., & Ramshini, H. (2022). Genotype and environment interaction and stability of grain yield and oil content of rapeseed cultivars. *Food Science & Nutrition*, *10*(12), 4308-4318. https://doi.org/10.1002/fsn3.3020

Rajković, D., Jeromela, A. M., Pezo, L., Lončar, B., Grahovac, N., & Špika, A. K. (2023). Artificial neural network and random forest regression models for modelling fatty acid and tocopherol content in oil of winter rapeseed. *Journal of Food Composition and Analysis*, *115*, 105020. https://doi.org/10.1016/j.jfca.2023.105020

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63-71). Springer. https://doi.org/10.1007/978-3-540-28650-9_4

Ren, W., Jiang, Q., & Qi, W. (2025). Research progress in near-infrared spectroscopy for detecting the quality of potato crops. *Chemical and Biological Technologies in Agriculture*, *12*(1), 32. https://doi.org/10.1186/s40538-025-00747-5

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Ruszczak, B., & Boguszewska-Mańkowska, D. (2022). Soil moisture a posteriori measurements enhancement using ensemble learning. *Sensors*, *22*, 4591. https://doi.org/10.3390/s22124591

Salami, M., Heidari, B., & Tan, H. (2023). Comparative profiling of polyphenols and antioxidants and analysis of antiglycation activities in rapeseed (*Brassica napus* L.) under different moisture regimes. *Food Chemistry*, *399*, 133946. https://doi.org/10.1016/j.foodchem.2022.133946

Salami, M., Heidari, B., Alizadeh, B., Batley, J., Wang, J., Tan, X.L., Dadkhodaie, A., & Richards, C. (2024a). Dissection of quantitative trait nucleotides and candidate genes associated with agronomic and yield-related traits under drought stress in rapeseed varieties: Integration of genome-wide association study and transcriptomic analysis. *Frontiers in Plant Science*, 15, 1342359. https://doi.org/10.3389/fpls.2024.1342359

Salami, M., Heidari, B., Batley, J., Wang, J., Tan, X. L., Richards, C., & Tan, H. (2024b). Integration of genome-wide association studies, metabolomics, and transcriptomics reveals phenolic acid- and flavonoid-associated genes and their regulatory elements under drought stress in rapeseed flowers. *Frontiers in Plant Science*, *14*, 1249142. https://doi.org/10.3389/fpls.2023.1249142

Salami, M., Tan, H., Thomas, W. J., Batley, J., & Heidari, B. (2025). Genome-wide association study (GWAS) combined with transcriptome analysis reveals the key genes underlying the production of seed oil, mono and poly-unsaturated fatty acids in *Brassica napus*. *Industrial Crops and Products*, *231*, 121205. https://doi.org/10.1016/j.indcrop.2025.121205

Salami, M., Tan, H., & Heidari, B. (2026a). Integrating machine learning (PLSR) prediction with MGIDI-driven quality selection for analysis of dynamic profiling of fatty acids, seed structural and metabolite compositions in response to drought stress in rapeseed. *Industrial Crops and Products*, (In Press).

Salami, M., Tan, H., & Heidari, B. (2026b). Bridging classical stability analysis (AMMI, GGE biplot, and stability metrics) with modern machine learning models for interpretation of genotype × environment interactions and genotype selection for high yield in rapeseed (*Brassica napus* L.). *BMC Plant Biology*, (In Press).

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21. https://doi.org/10.1007/s42979-021-00592-x

Schwender, J., & Hay, J. O. (2012). Predictive modeling of biomass component tradeoffs in *Brassica napus* developing oilseeds based on in silico manipulation of storage metabolism. *Plant Physiology*, *160*(3), 1218-1236. https://doi.org/10.1104/pp.112.203034

Secchi, M. A., Fernandez, J. A., Stamm, M. J., Durrett, T., Prasad, P. V., Messina, C. D., & Ciampitti, I. A. (2023). Effects of heat and drought on canola (*Brassica napus* L.) yield, oil, and protein: A meta-analysis. *Field Crops Research*, *293*, 108848. https://doi.org/10.1016/j.fcr.2023.108848

Seeger, M. (2004). Gaussian processes for machine learning. *International Journal of Neural Systems*, *14*(02), 69-106. https://doi.org/10.1142/S0129065704001899

Shahsavari, M., Raspor, M., & Mohammadi, V. (2025). Genome-wide transcriptional analysis of rapeseed (*Brassica napus* L.) responses to heat stress: An integrated RNA-Seq meta-analysis, gene co-expression, and gene regulatory network study. *BMC Plant Biology*, *26*(1), 73. https://doi.org/10.1186/s12870-025-06481-w

Sharma, G., & Gupta, J. (2026). Application of machine learning models for drought prediction with special reference to *Brassica* crops. In *Crop improvement strategies in Brassica species: Applied Science*. Springer. https://doi.org/10.1007/978-981-96-3456-7_12

Sharma, P., Villegas-Diaz, R., & Fennell, A. (2024). Predicting grapevine physiological parameters using hyperspectral remote sensing integrated with hybrid convolutional neural network and ensemble stacked regression. *Remote Sensing*, *16*, 2626. https://doi.org/10.3390/rs16142626

Shi, L., Zhao, W., Yang, Z., Subbiah, V., & Suleria, H.A.R. (2022). Extraction and characterization of phenolic compounds and their potential antioxidant activities. *Environmental Science and Pollution Research*, *29*, 81112-81129. https://doi.org/10.1007/s11356-022-23337-2

Sudha, S. P., & Loret, J. B. (2026). A review on machine learning-based precision agriculture techniques for crop farming monitoring with IOT. *Discover Environment*, *4*(1), 10. https://doi.org/10.1007/s44274-025-00110-4

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, *48*(1), 159-205. https://doi.org/10.1016/j.csda.2004.03.005

Tian, H., Song, H., Wu, X., & Zhang, Z. (2022). Responses of cell wall components to low nitrogen in rapeseed roots. *Agronomy*, *12*(5), 1044. https://doi.org/10.3390/agronomy12051044

Tsega, A., & Mullualem, D. (2026). Machine learning for multi-omics data integration in crop improvement: A systematic review. *BMC Bioinformatics*, *27*(1), 81-81. https://doi.org/10.1186/s12859-026-06000-0

Tseng, C.Y., Salguero, J.A., Breidenbach, J.D., Solomon, E., Sanders, C.K., Harvey, T., Thornhill, M.G., Palmisano, S.J., Sasiene, Z.J., Blackwell, B.R., & McBride, E.M. (2025). Evaluation of normalization strategies for mass spectrometry-based multi-omics datasets. *Metabolomics*, *21*(4), 98. https://doi.org/10.1007/s11306-025-02297-1

Uche, C. U., Okezue, M. A., Amidu, I., & Byrn, S. R. (2024). Model adequacy in assessing the predictive performance of regression models in pharmaceutical product optimization: The bedaquiline solid lipid nanoparticle example. *Scientia Pharmaceutica*, *92*, 64. https://doi.org/10.3390/scipharm92040064

Ullah, A. (2025). Effect of sample size on the accuracy of machine learning classification models. *Spectrum of Engineering Sciences*, 826-834. https://doi.org/10.1234/ses.2025.12345

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Van Soest, P. J., Robertson, J. B., & Lewis, B. A. (1991). Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. *Journal of Dairy Science*, *74*, 3583–3597. https://doi.org/10.3168/jds.S0022-0302(91)78551-2

Veser, J., Peeters, W., Schutyser, M. A. I., & van der Sman, R. G. M. (2025). Prediction of cabbage seed drying at lab and industrial scale using a 'non-equilibrium' sorption isotherm model. *Current Research in Food Science*, 101239. https://doi.org/10.1016/j.crfs.2025.101239

Wang, C., Zhang, J., Kuai, J., Xie, J., Wu, W., Hua, S., Yan, M., Du, H., Ma, N., & You, L. (2025). Unlock genotype-environment-management interaction via field phenotypic insights for multi-scale prediction of winter rapeseed flowering in the Yangtze River Basin. *Agricultural and Forest Meteorology*, *374*, 110788. https://doi.org/10.1016/j.agrformet.2025.110788

Wang, H., Zhang, L., Kong, W., Li, S., Wang, Q., Li, C., Tang, P., Zhu, Z., & Peng, W. (2026). RapeseedNet: A lightweight deep learning model for rapeseed quality detection based on near-infrared spectroscopy. *Computers and Electronics in Agriculture*, *248*, 111765. https://doi.org/10.1016/j.compag.2026.111765

Wang, T., Tong, R., Yuan, Z., & Li, Y. (2026). Beyond QTL and GWAS: How deep learning, graph models, and multi-omics are reshaping plant genomic prediction analysis. *Frontiers in Genetics*, *17*, 1783939. https://doi.org/10.3389/fgene.2026.1783939

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

Xia, L., Zhang, P., Shi, L., Wang, K., & Zhang, T. (2025). Remote sensing estimation of chlorophyll content in rape leaves in Weibei dryland region of China. *Open Geosciences*, *17*(1), 20220756. https://doi.org/10.1515/geo-2022-0756

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, *129*(4), 653-673. https://doi.org/10.1007/s00122-016-2691-5

Yan, W., & Kang, M. S. (2003). *GGE biplot analysis: A graphical tool for breeders, geneticists and agronomists.* CRC Press. https://doi.org/10.1201/9781420040371

Yan, W., & Tinker, N. A. (2006). Biplot analysis of multi-environment trial data: Principles and applications. *Canadian Journal of Plant Science*, *86*, 623-645. https://doi.org/10.4141/P05-169

Yan, W., Hunt, L. A., Sheng, Q., & Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on GGE biplot. *Crop Science*, *40*, 597-605. https://doi.org/10.2135/cropsci2000.403597x

Yang, Y., Guo, C., Xu, M., & Hu, F. (2026). Plant wearable sensors enable closed-loop intelligent agriculture: Real-time monitoring to precision management. *TrAC Trends in Analytical Chemistry*, *198*, 118761. https://doi.org/10.1016/j.trac.2026.118761

Yang, Z., & Tang, T. (2026). Closing the loop: AI-driven integration of multi-omics and phenomics for systematic resilient crop engineering. *Biotechnology Advances*, *90*, 108908. https://doi.org/10.1016/j.biotechadv.2026.108908

Yenkikar, A., Mishra, V. P., Bali, M., & Ara, T. (2025). An explainable AI-based hybrid machine learning model for interpretability and enhanced crop yield prediction. *MethodsX*, *15*, 103442. https://doi.org/10.1016/j.mex.2025.103442

Yin, S., Peng, Y., Yang, T., Jin, Q., Wang, Y., Wan, M., Mu, D., King, G.J., Zeng, C., Ge, X., & Wang, J. (2026). *BnaFLC.A10*, *BnaFLC.A2*, and *BnaFT.A2* jointly regulate flowering time in *Brassica napus*. *Journal of Experimental Botany*, *77*(4), 1042-1057. https://doi.org/10.1093/jxb/erae483

Yu, T., Zhang, H., Chen, S., Gao, S., Liu, Z., Wang, J., Crossa, J., Montesinos-López, O.A., Hearne, S., & Li, H. (2025). EXGEP: A framework for predicting genotype-by-environment interactions using ensembles of explainable machine-learning models. *Briefings in Bioinformatics*, *26*(4), bbaf414. https://doi.org/10.1093/bib/bbaf414

Yu, W., Wang, X., Wang, H., Wang, W., Cheng, H., Mei, D., Jiang, L., Hu, Q., & Liu, J. (2025). Optimization and application of genome prediction model in rapeseed: Flowering time, yield components, and oil content as examples. *Horticulture Research*, *12*(8), uhaf115. https://doi.org/10.1093/hr/uhaf115

Yue, F., Zhang, J., Xu, J., Niu, T., Lü, X., & Liu, M. (2022). Effects of monosaccharide composition on quantitative analysis of total sugar content by phenol-sulfuric acid method. *Frontiers in Nutrition*, 9, 963318. https://doi.org/10.3389/fnut.2022.963318

Zahedi, M. B., Tahmasebi, S., Salami, M., Sukumaran, S., & Heidari, B. (2026). Identification of drought-tolerant ideotype in wheat using multi-trait genotype-ideotype distance index (MGIDI) and comprehensive evaluation value. *Scientific Reports*, *16*, 2045-2322. https://doi.org/10.1038/s41598-026-46140-6

Zhang, C., Wang, B., Zhang, G., Yin, P., & Lu, X. (2015). Rapid method for determination of crude fiber, NDF, ADF and ADL contents in feed. *Shandong Journal of Animal Science and Veterinary Medicine*, 36, 20-22. https://doi.org/10.1234/sdjasvm.2015.0004

Zhang, Y., Chai, X., Hu, J., Niu, Y., & Xu, L. (2026). Enhancing rapeseed biomass and yield estimation with ensemble learning and synergistic multidimensional features. *Journal of Zhejiang University-SCIENCE B*, *27*(5), 499-516. https://doi.org/10.1631/jzus.B2500004